SENTIMENT ANALYSIS FOR MARKETING

BATCH MEMBER

211121104054 : THULASI NAVANEETHAN . N

Phase 3 Submission Document

Phase 3 : Development Part 1

Topic : Start building the sentiment analysis model by loading and preprocessing the dataset .

Introduction To Sentiment Analysis

Sentiment analysis refers to analyzing an opinion or feelings about something using data like text or images, regarding almost anything. Sentiment analysis helps companies in their decision-making process. For instance, if public sentiment towards a product is not so good, a company may try to modify the product or stop the production altogether in order to avoid any losses.

There are many sources of public sentiment e.g. public interviews, opinion polls, surveys, etc. However, with more and more people joining social media platforms, websites like Facebook and Twitter can be parsed for public sentiment.

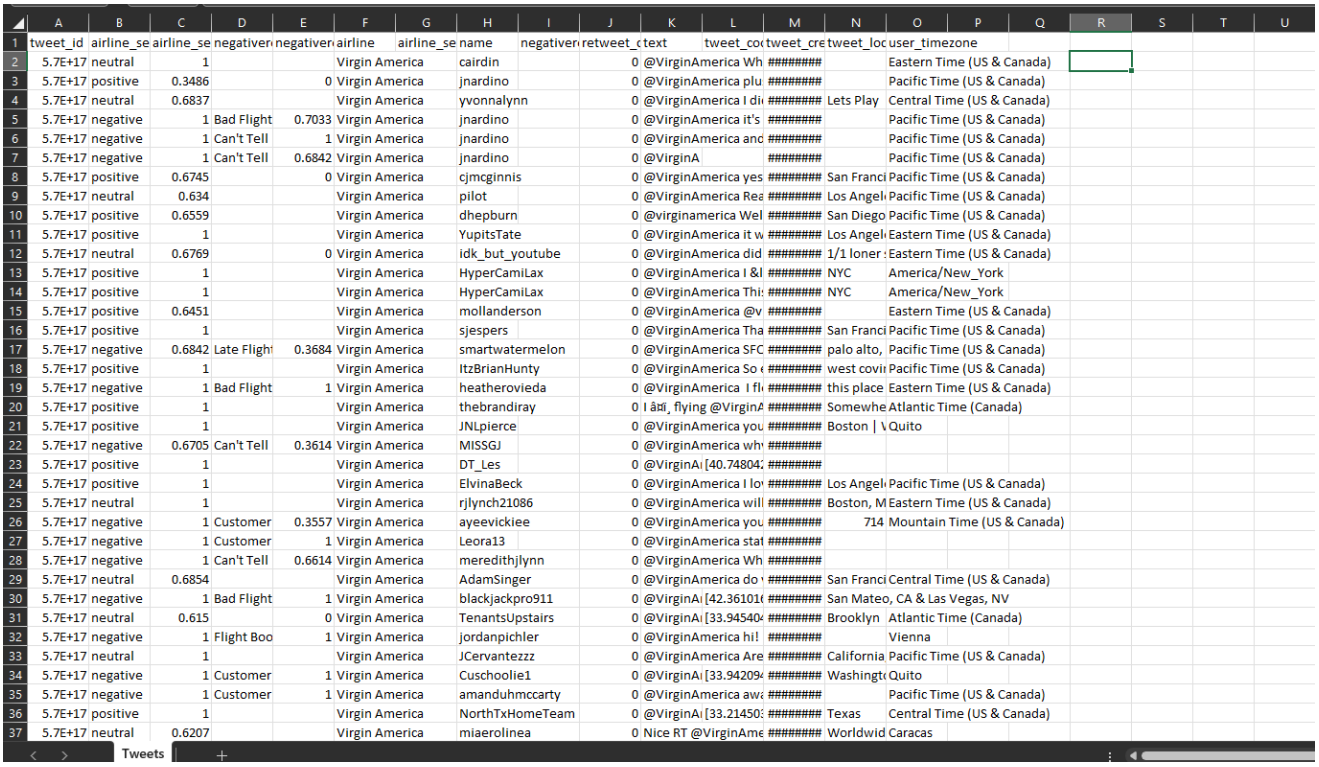
Problem Definition

Given tweets about six US airlines, the task is to predict whether a tweet contains positive, negative, or neutral sentiment about the airline. This is a typical supervised learning task where given a text string, we have to categorize the text string into predefined categories.

Solution

To solve this problem, we will follow the typical machine learning pipeline. We will first import the required libraries and the dataset. We will then do exploratory data analysis to see if we can find any trends in the dataset. Next, we will perform text preprocessing to convert textual data to numeric data that can be used by a machine learning algorithm. Finally, we will use machine learning algorithms to train and test our sentiment analysis models.

Given dataset:



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	tweet_id	airline_se	airline_se	negativer	negativer	airline	airline_se	name	negativer	retweet_c	text	tweet_co	tweet_cre	tweet_loc	user_timezone		
2	5.7E+17	neutral	1			Virgin America		cairdin		0	@VirginAmerica Wh		########		Eastern Time (US & Canada)		
3	5.7E+17	positive	0.3486		0	Virgin America		jnardino		0	@VirginAmerica plu		########		Pacific Time (US & Canada)		
4	5.7E+17	neutral	0.6837			Virgin America		yvonnalynn		0	@VirginAmerica I di		########	Lets Play	Central Time (US & Canada)		
5	5.7E+17	negative	1	Bad Flight	0.7033	Virgin America		jnardino		0	@VirginAmerica it's		########		Pacific Time (US & Canada)		
6	5.7E+17	negative	1	Can't Tell	1	Virgin America		jnardino		0	@VirginAmerica and		########		Pacific Time (US & Canada)		
7	5.7E+17	negative	1	Can't Tell	0.6842	Virgin America		jnardino		0	@VirginA		########		Pacific Time (US & Canada)		
8	5.7E+17	positive	0.6745		0	Virgin America		cjmcginnis		0	@VirginAmerica yes		########	San Franci	Pacific Time (US & Canada)		
9	5.7E+17	neutral	0.634			Virgin America		pilot		0	@VirginAmerica Rea		########	Los Angel	Pacific Time (US & Canada)		
10	5.7E+17	positive	0.6559			Virgin America		dhepburn		0	@virginamerica Wel		########	San Diego	Pacific Time (US & Canada)		
11	5.7E+17	positive	1			Virgin America		YupitsTate		0	@VirginAmerica it w		########	Los Angel	Eastern Time (US & Canada)		
12	5.7E+17	neutral	0.6769		0	Virgin America		idk_but_youtube		0	@VirginAmerica did		########	1/1 loner :	Eastern Time (US & Canada)		
13	5.7E+17	positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica I &l		########	NYC	America/New_York		
14	5.7E+17	positive	1			Virgin America		HyperCamiLax		0	@VirginAmerica Thi		########	NYC	America/New_York		
15	5.7E+17	positive	0.6451			Virgin America		mollanderson		0	@VirginAmerica @v		########		Eastern Time (US & Canada)		
16	5.7E+17	positive	1			Virgin America		sjespers		0	@VirginAmerica Tha		########	San Franci	Pacific Time (US & Canada)		
17	5.7E+17	negative	0.6842	Late Flight	0.3684	Virgin America		smartwatermelon		0	@VirginAmerica SFC		########	palo alto,	Pacific Time (US & Canada)		
18	5.7E+17	positive	1			Virgin America		ItzBrianHunty		0	@VirginAmerica So		########	west covi	Pacific Time (US & Canada)		
19	5.7E+17	negative	1	Bad Flight	1	Virgin America		heatherovieda		0	@VirginAmerica I fl		########	this place	Eastern Time (US & Canada)		
20	5.7E+17	positive	1			Virgin America		thebrandiray		0	I â¤ï¸ flying @VirginA		########	Somewhe	Atlantic Time (Canada)		
21	5.7E+17	positive	1			Virgin America		JNLpierce		0	@VirginAmerica you		########	Boston | \	Quito		
22	5.7E+17	negative	0.6705	Can't Tell	0.3614	Virgin America		MISSGJ		0	@VirginAmerica wh		########				
23	5.7E+17	positive	1			Virgin America		DT_Les		0	@VirginAi [40.74804		########				
24	5.7E+17	positive	1			Virgin America		ElvinaBeck		0	@VirginAmerica I lo		########	Los Angel	Pacific Time (US & Canada)		
25	5.7E+17	neutral	1			Virgin America		rjlynch21086		0	@VirginAmerica will		########	Boston, M	Eastern Time (US & Canada)		
26	5.7E+17	negative	1	Customer	0.3557	Virgin America		ayeevickiee		0	@VirginAmerica you		########	714	Mountain Time (US & Canada)		
27	5.7E+17	negative	1	Customer	1	Virgin America		Leora13		0	@VirginAmerica stat		########				
28	5.7E+17	negative	1	Can't Tell	0.6614	Virgin America		meredithjlynn		0	@VirginAmerica Wh		########				
29	5.7E+17	neutral	0.6854			Virgin America		AdamSinger		0	@VirginAmerica do		########	San Franci	Central Time (US & Canada)		
30	5.7E+17	negative	1	Bad Flight	1	Virgin America		blackjackpro911		0	@VirginAi [42.36101		########	San Mateo,	CA & Las Vegas, NV		
31	5.7E+17	neutral	0.615		0	Virgin America		TenantsUpstairs		0	@VirginAi [33.94540		########	Brooklyn	Atlantic Time (Canada)		
32	5.7E+17	negative	1	Flight Boo	1	Virgin America		jordanpichler		0	@VirginAmerica hi!		########		Vienna		
33	5.7E+17	neutral	1			Virgin America		JCervantezzz		0	@VirginAmerica Are		########	California	Pacific Time (US & Canada)		
34	5.7E+17	negative	1	Customer	1	Virgin America		Cuschoolie1		0	@VirginAi [33.94209		########	Washingt	Quito		
35	5.7E+17	negative	1	Customer	1	Virgin America		amanduhmccarty		0	@VirginAmerica awa		########		Pacific Time (US & Canada)		
36	5.7E+17	positive	1			Virgin America		NorthTxHomeTeam		0	@VirginAi [33.21450		########	Texas	Central Time (US & Canada)		
37	5.7E+17	neutral	0.6207			Virgin America		miaerolinea		0	Nice RT @VirginAme		########	Worldwid	Caracas		

Tweets

Importance Of loading and processing dataset :

Loading and processing a dataset is a crucial step in data analysis, machine learning, and many other data-related tasks. The importance of this step lies in its role in ensuring th e quality, integrity, and suitability of the data for subsequent analysis and modeling. Here a re some key reasons why loading and processing a dataset is important:

1. Data Quality Assurance: Loading and processing the dataset allows you to identify and ad dress data quality issues, such as missing values, outliers, inconsistencies, and errors. This i s critical for ensuring the accuracy and reliability of any analysis or modeling that follows.

2. Data Cleaning : Often, real-world datasets are messy and contain missing or irrelevant inf ormation. Pre-processing involves cleaning the data by removing or imputing missing value s and eliminating redundant or noisy features, making the dataset more suitable for analysi s.

3. Data Transformation : You may need to transform data into a more appropriate format o r scale. For example, normalizing features can be important for machine learning algorithm s that are sensitive to the scale of input features.

4. Feature Engineering : Feature engineering involves creating new features or modifying e xisting ones to better represent the underlying patterns in the data. Proper feature enginee ring can significantly improve the performance of machine learning models.

5. Data Exploration : Loading the dataset allows you to explore its characteristics, distribut ions, and relationships between variables. This exploratory data analysis (EDA) helps you g ain insights and inform subsequent analysis decisions.

6. Data Preprocessing for Machine Learning : In the context of machine learning, loading an d preprocessing data are essential steps. You need to split the data into training and testing sets, perform one-hot encoding or label encoding for categorical variables, and handle class imbalances, if any.

7. Data Security and Privacy : Ensuring that sensitive information is properly handled and p rotected is crucial. Loading and processing data offer opportunities to anonymize or mask personally identifiable information to comply with data privacy regulations.

8. Data Reduction : Large datasets can be computationally expensive to work with. Preproc essing can involve dimensionality reduction techniques like PCA to reduce the number of fe atures while preserving important information.

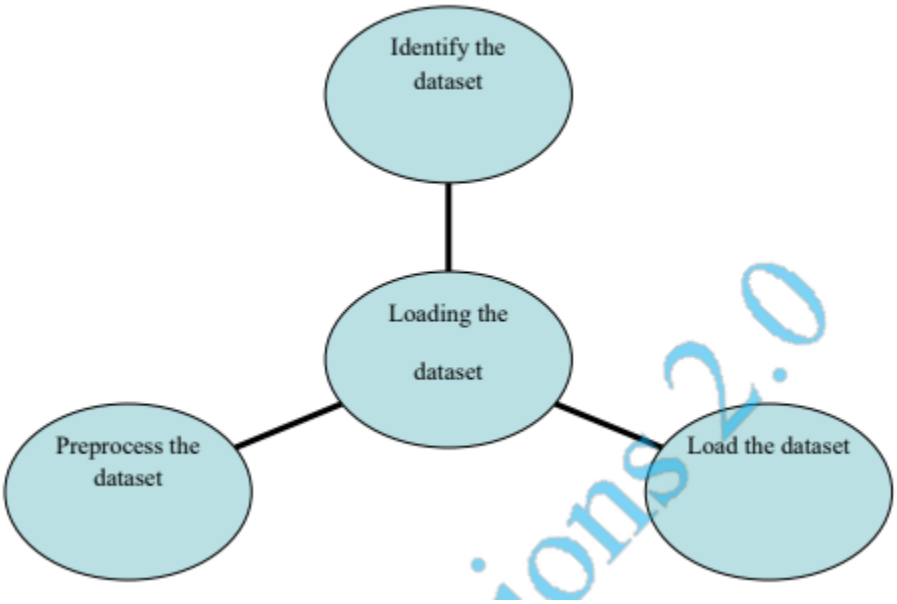
9. Data Standardization : In some cases, it's important to standardize data so that it can be e asily compared or combined with other datasets. This involves ensuring consistent units, ti me zones, or data formats.

10. Data Visualization : Data visualization often comes after data loading and processing. It helps you communicate your findings and insights effectively and aids in understanding the data's characteristics.

11. Model Interpretability : Proper preprocessing can make your models more interpretabl e by ensuring that input features are meaningful and appropriately scaled.

12. Time and Resource Efficiency : Preprocessing can also involve optimizing data storage f ormats or compression techniques to save storage space and reduce the time required for d ata loading and analysis.

In summary, loading and processing a dataset are essential steps in any data-driven project. Properly prepared data ensures that subsequent analysis, modeling, and decision-making processes are accurate, efficient, and reliable. It also helps in uncovering valuable in sights from the data and addressing issues of data quality, privacy, and compliance.



Program :

```
import numpy as np
import pandas as pd
import re
import seaborn as sns
import nltk
import matplotlib.pyplot as plt
%matplotlib inline

airline_tweets = pd.read_csv(r'/content/Tweets.csv')
airline_tweets.head()
```

```
             tweet_id airline_sentiment  airline_sentiment_confidence  \
0  570306133677760513           neutral                        1.0000
1  570301130888122368          positive                        0.3486
2  570301083672813571           neutral                        0.6837
3  570301031407624196          negative                        1.0000
4  570300817074462722          negative                        1.0000

  negativereason  negativereason_confidence         airline  \
0            NaN                        NaN  Virgin America
1            NaN                     0.0000  Virgin America
2            NaN                        NaN  Virgin America
3     Bad Flight                     0.7033  Virgin America
4     Can't Tell                     1.0000  Virgin America

  airline_sentiment_gold        name negativereason_gold  retweet_count  \
0                    NaN     cairdin                 NaN              0
1                    NaN    jnardino                 NaN              0
2                    NaN  yvonnalynn                 NaN              0
3                    NaN    jnardino                 NaN              0
4                    NaN    jnardino                 NaN              0

                                          text tweet_coord  \
0                @VirginAmerica What @dhepburn said.         NaN
1  @VirginAmerica plus you've added commercials t...         NaN
2  @VirginAmerica I didn't today... Must mean I n...         NaN
3  @VirginAmerica it's really aggressive to blast...         NaN
4  @VirginAmerica and it's a really big bad thing...         NaN

              tweet_created tweet_location               user_timezone
0  2015-02-24 11:35:52 -0800            NaN  Eastern Time (US & Canada)
1  2015-02-24 11:15:59 -0800            NaN  Pacific Time (US & Canada)
2  2015-02-24 11:15:48 -0800      Lets Play  Central Time (US & Canada)
3  2015-02-24 11:15:36 -0800            NaN  Pacific Time (US & Canada)
4  2015-02-24 11:14:45 -0800            NaN  Pacific Time (US & Canada)
```

Let's explore the dataset a bit to see if we can find any trends. But before that, we will change the default plot size to have a better view of the plots.

```
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
```

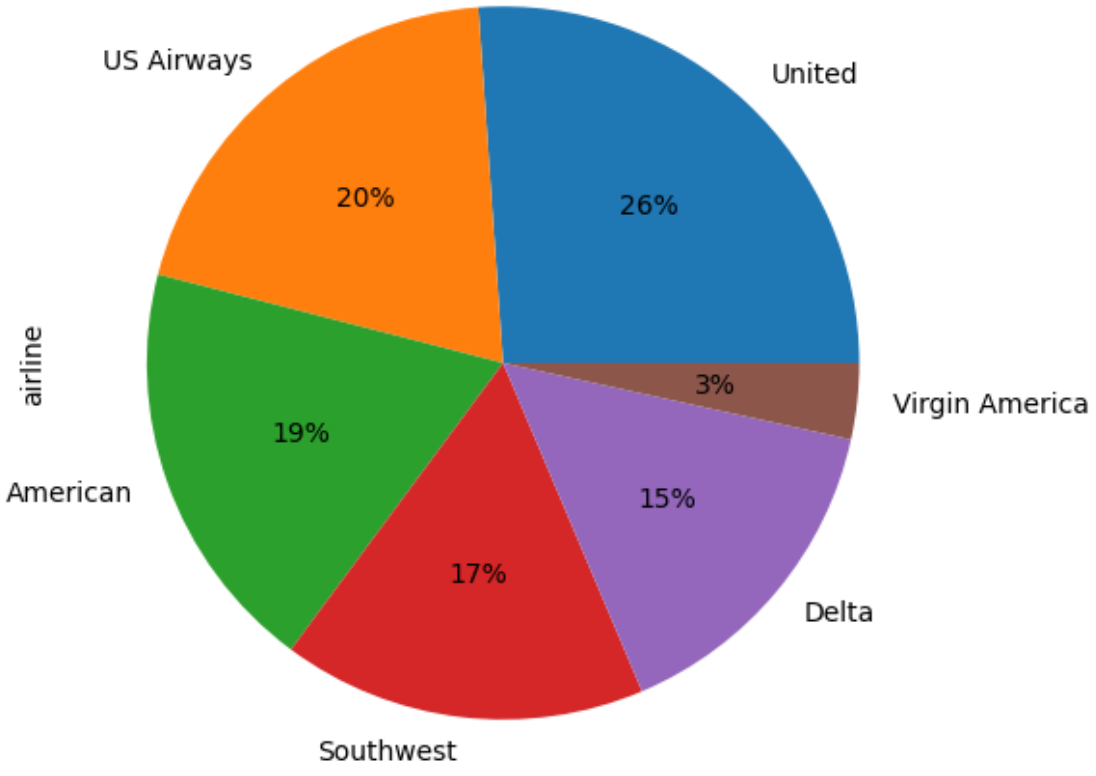
```
8.0
6.0
```

Exploration Of Data

Let's first see the number of tweets for each airline. We will plot a pie chart for that:

```
airline_tweets.airline.value_counts().plot(kind='pie', autopct='%1.0f%%')
```

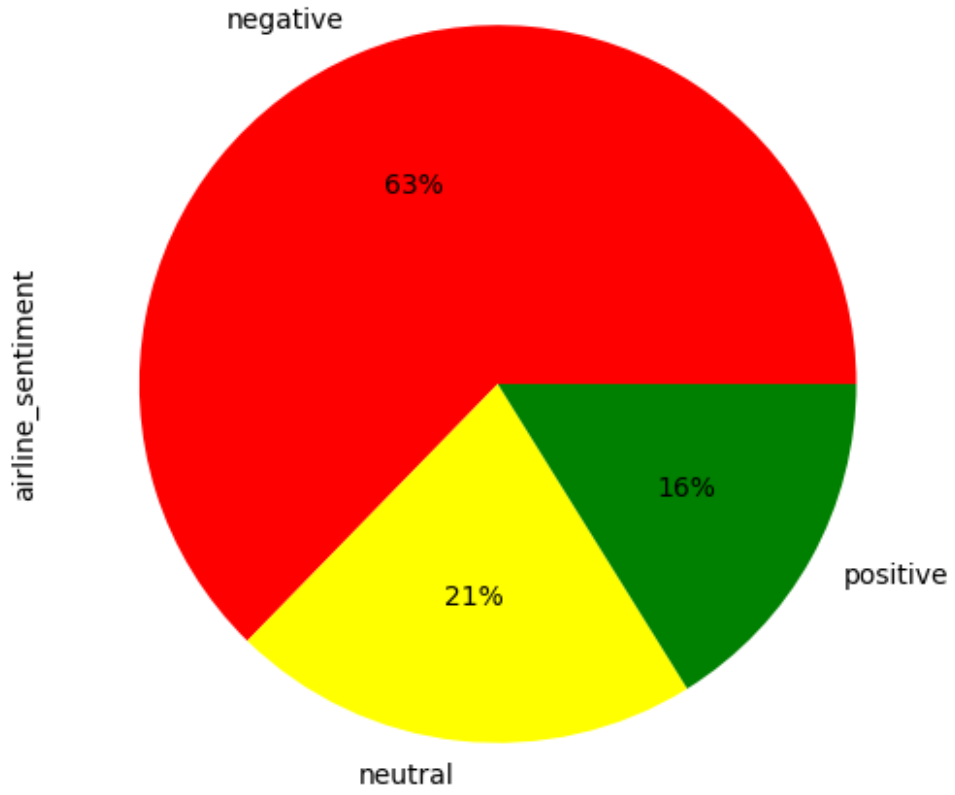
```
<Axes: ylabel='airline'>
```



Let's now see the distribution of sentiments across all the tweets.

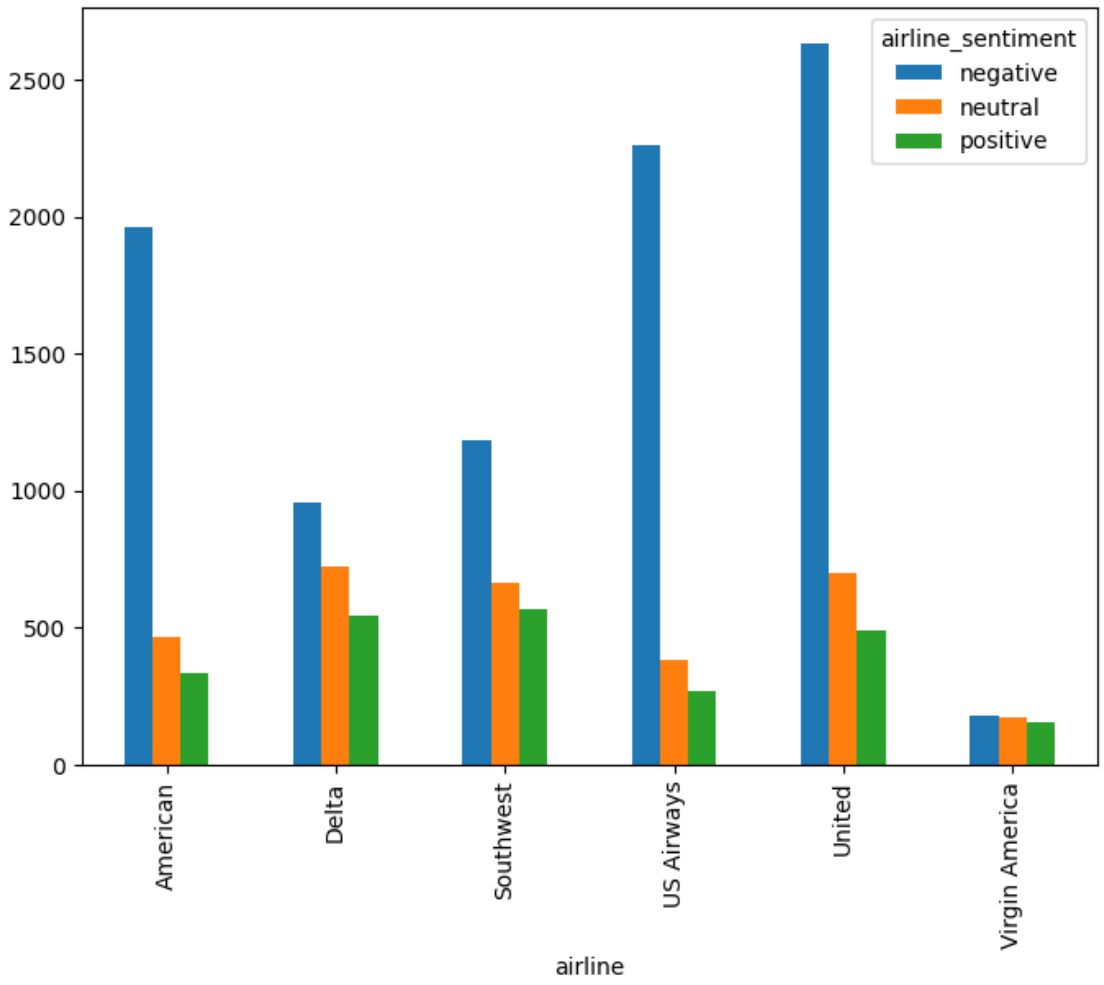
```
airline_tweets.airline_sentiment.value_counts().plot(kind='pie', autopct='%1.
0f%%', colors=["red", "yellow", "green"])
```

```
<Axes: ylabel='airline_sentiment'>
```




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

```
<Axes: xlabel='airline'>
```

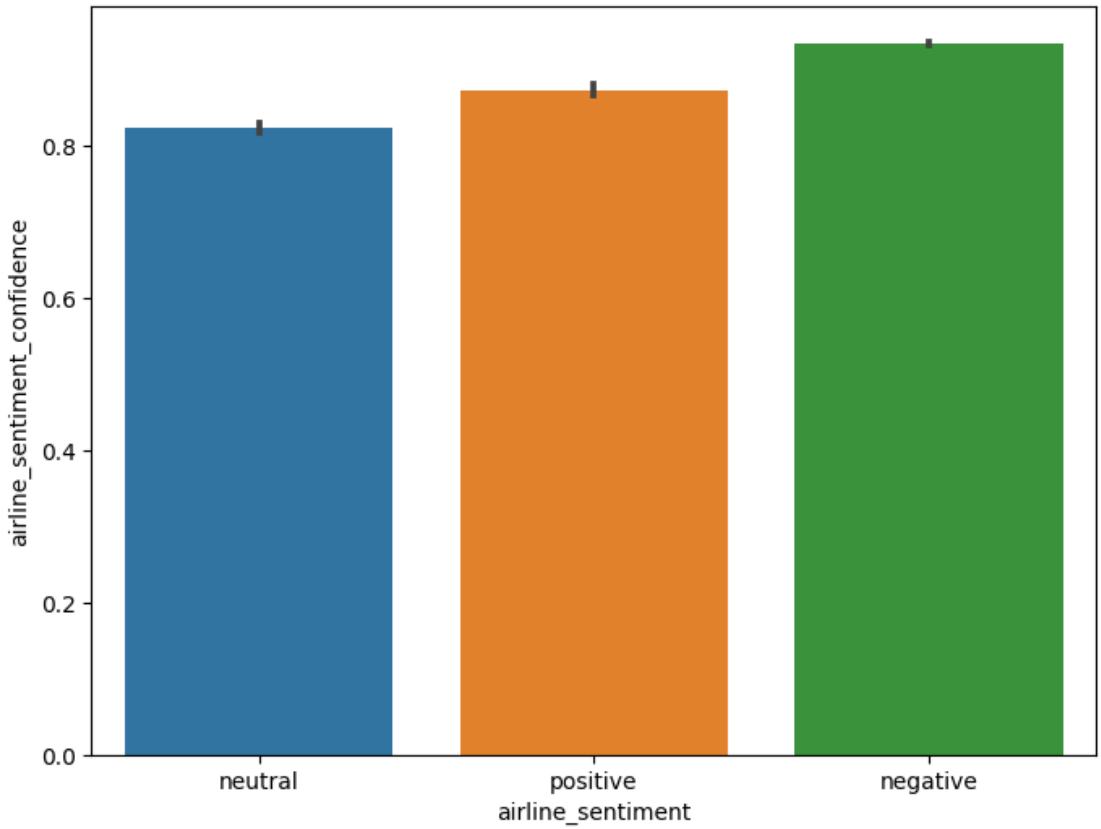


It is evident from the output that for almost all the airlines, the majority of the tweets are negative, followed by neutral and positive tweets. Virgin America is probably the only airline where the ratio of the three sentiments is somewhat similar.

Finally, let's use the Seaborn library to view the average confidence level for the tweets belonging to three sentiment categories.

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

```
<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.

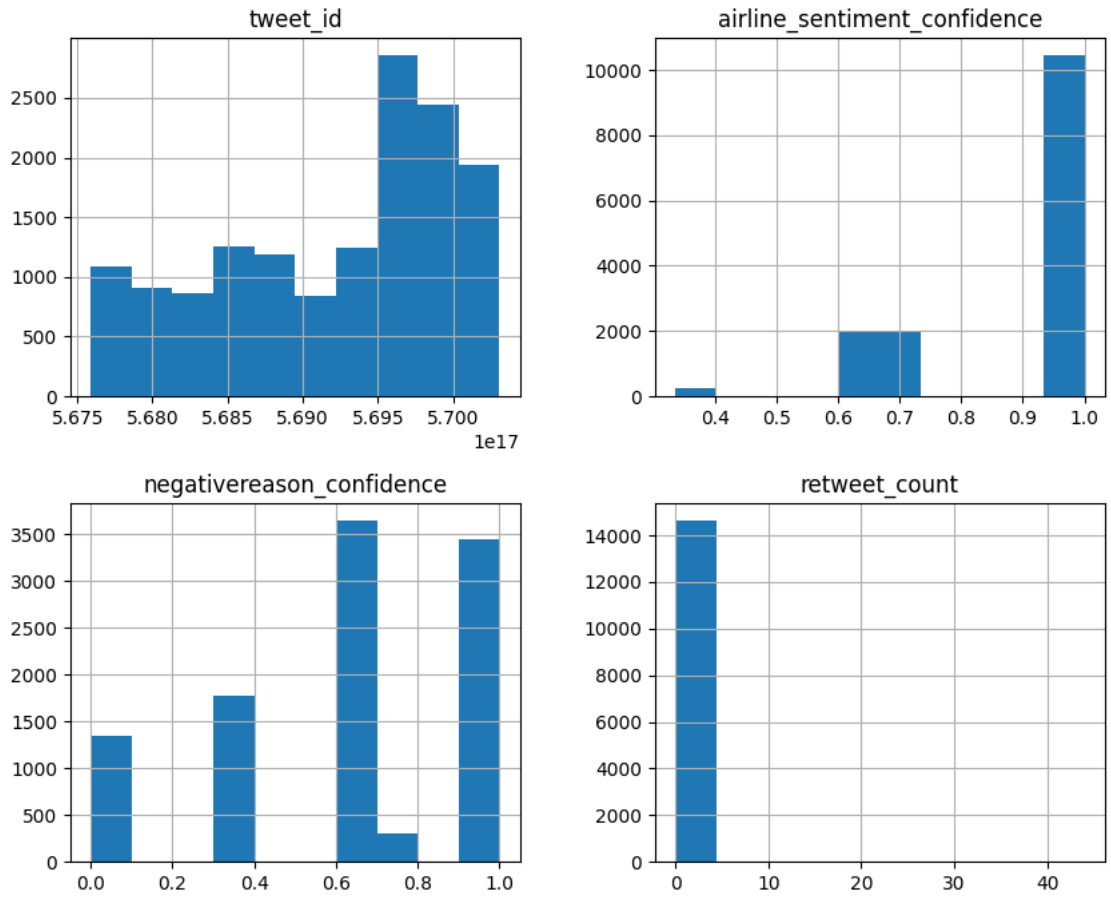
Data Cleaning

Tweets contain many slang words and punctuation marks. We need to clean our tweets before they can be used for training the machine learning model. However, before cleaning the tweets, let's divide our dataset into feature and label sets.

Our feature set will consist of tweets only. If we look at our dataset, the 11th column contains the tweet text. Note that the index of the column will be 10 since pandas columns follow zero-based indexing scheme where the first column is called 0th column. Our label set will consist of the sentiment of the tweet that we have to predict. The sentiment of the tweet is in the second column (index 1). To create a feature and a label set, we can use the iloc method off the pandas data frame.

```
airline_tweets.hist(figsize=(10,8))
```

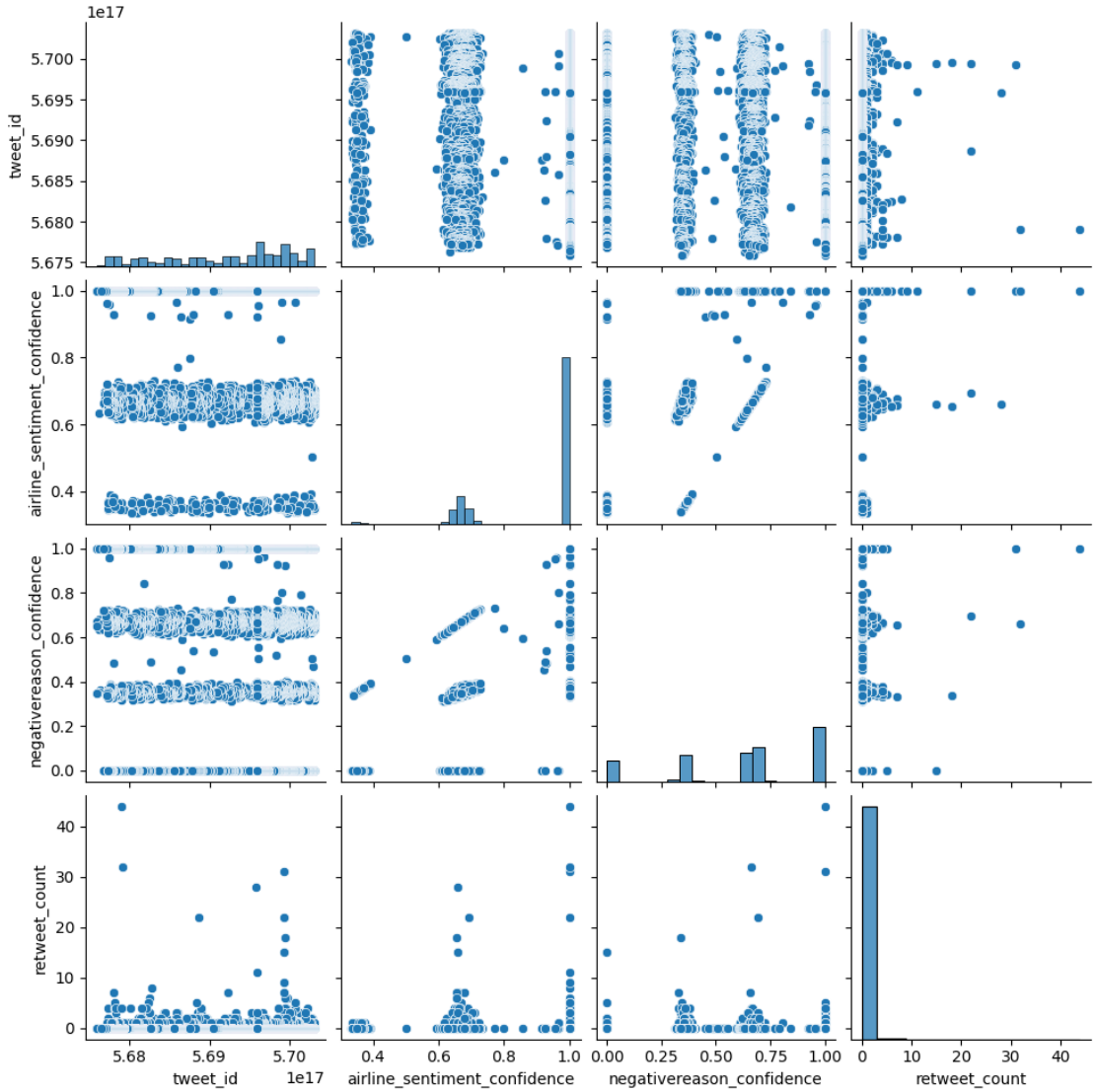
```
array([[<Axes: title={'center': 'tweet_id'}>,
        <Axes: title={'center': 'airline_sentiment_confidence'}>],
       [<Axes: title={'center': 'negativereason_confidence'}>,
        <Axes: title={'center': 'retweet_count'}>]], dtype=object)
```




```
plt.figure(figsize=(12,8))
sns.pairplot(airline_tweets)
```

```
<seaborn.axisgrid.PairGrid at 0x787bf77a6a10>
```

```
<Figure size 1200x800 with 0 Axes>
```




```
airline_tweets.corr(numeric_only=True)
```

```
                              tweet_id  airline_sentiment_confidence  \
tweet_id                      1.000000                      0.024840
airline_sentiment_confidence  0.024840                      1.000000
negativereason_confidence     0.021533                      0.685879
retweet_count                -0.008852                      0.012581

                              negativereason_confidence  retweet_count
tweet_id                                       0.021533      -0.008852
airline_sentiment_confidence                   0.685879       0.012581
negativereason_confidence                      1.000000       0.021574
retweet_count                                  0.021574       1.000000
```

```
plt.figure(figsize=(10,5))
sns.heatmap(airline_tweets.corr(numeric_only = True), annot=True)
```

<Axes: >

Conclusion :

In conclusion, sentiment analysis models are powerful tools for extracting valuable insights from text data, enabling businesses, researchers, and individuals to understand and harness sentiment in a wide range of applications.