Python for NLP: Sentiment Analysis with Scikit-Learn

https://stackabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/ (https://stackabuse.com/python-for-nlp-sentiment-analysis-with-scikit-learn/)

Problem Definition:

Given tweets about six US airlines, the task is to predict whether a tweet contains positive, negative or neautral sentiment about the airline.

In [1]:
```
import numpy as np
import pandas as pd
import re
import nltk
import matplotlib.pyplot as plt
%matplotlib inline
```

In [2]:
```
data_source_url = "https://raw.githubusercontent.com/kolaveridi/kaggle-Twitter-US-Airline-Sentiment-/master/Tweets.csv"
airline_tweets = pd.read_csv(data_source_url)
```

In [3]:
```
# Peaking at the data
airline_tweets.head()
```

Out[3]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_locatio
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn said.	NaN	2015-02-24 11:35:52 -0800	Nal
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials t...	NaN	2015-02-24 11:15:59 -0800	Nal
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn	NaN	0	@VirginAmerica I didn't today... Must mean I n...	NaN	2015-02-24 11:15:48 -0800	Lets Pla
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica it's really aggressive to blast...	NaN	2015-02-24 11:15:36 -0800	Nal
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica and it's a really big bad thing...	NaN	2015-02-24 11:14:45 -0800	Nal

Data Analysis:

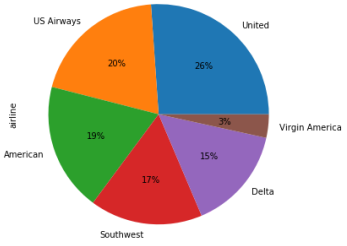
In [4]:
```
plot_size = plt.rcParams["figure.figsize"]
print(plot_size[0])
print(plot_size[1])

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

```
6.0
4.0
```

Percentage of public tweets for each airline

In [5]:
```
airline_tweets.airline.value_counts().plot(kind='pie', autopct='%1.0f%%')
```

Out[5]: <matplotlib.axes._subplots.AxesSubplot at 0x22ca6d8aa08>

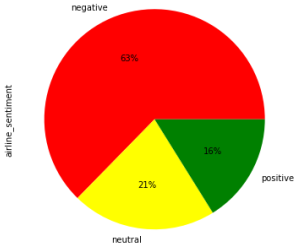


It is noted that United has the most tweets with 26%.

Distribution of sentiments across all tweets

In [6]:
```
airline_tweets.airline_sentiment.value_counts().plot(kind='pie', autopct='%1.0f%%', colors=["red", "yellow", "green"])
```

Out[6]: <matplotlib.axes._subplots.AxesSubplot at 0x22ca71f6888>

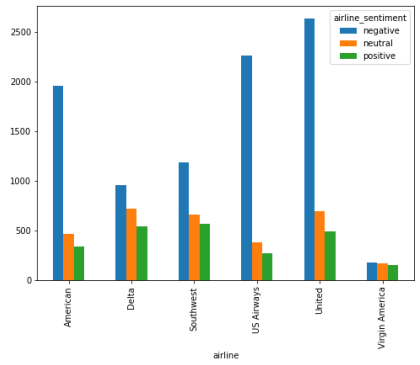


It is noted that an overwhelming majority of the tweets are negative with 63%

Distribution of sentiments for each individual airline

In [7]:
```
airline_sentiment = airline_tweets.groupby(['airline', 'airline_sentiment']).airline_sentiment.count().unstack()
airline_sentiment.plot(kind='bar')
```

Out[7]: <matplotlib.axes._subplots.AxesSubplot at 0x22ca7266048>

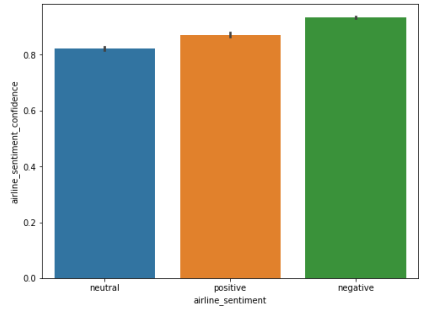


It is noted that for most airlines, the majority of the tweets are negative, followed by neutral and positive tweets. It is also noted that Virgin American is the only airline where the ratio of the sentiments is somewhat similar.

Confidence level for the tweets belonging to three sentiment categories

In [8]:
```
import seaborn as sns
sns.barplot(x='airline_sentiment', y='airline_sentiment_confidence', data=airline_tweets)
```

Out[8]: <matplotlib.axes._subplots.AxesSubplot at 0x22ca96fbc48>



**It is noted that the confidence level for negative tweets is higher compared to positive and neutral tweets.

Data Cleaning