TechnoHacks : Data science

Task 2 : Social media sentiment analysics

Use a dataset of tweets or Facebook posts and perform sentiment analysis to determine the overall sentiment of the posts.

Author : Sujata Gaikwad

In [1]:
```
## import librires
import pandas as pd
import numpy as np
```

In [2]:
```
data=pd.read_csv(r"C:\Users\HP\Downloads\Tweets.csv.zip")
data
```

Out[2]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	nega
0	570306133677760513	neutral	1.0000	NaN	
1	570301130888122368	positive	0.3486	NaN	
2	570301083672813571	neutral	0.6837	NaN	
3	570301031407624196	negative	1.0000	Bad Flight	
4	570300817074462722	negative	1.0000	Can't Tell	
...	...	...	...	...	
14635	569587686496825344	positive	0.3487	NaN	
14636	569587371693355008	negative	1.0000	Customer Service Issue	
14637	569587242672398336	neutral	1.0000	NaN	
14638	569587188687634433	negative	1.0000	Customer Service Issue	
14639	569587140490866689	neutral	0.6771	NaN	

14640 rows × 15 columns

In [3]:
```
data.info
```

```
Out[3]: <bound method DataFrame.info of                  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
...                   ...               ...                           ...
14635  569587686496825344          positive                        0.3487
14636  569587371693355008          negative                        1.0000
14637  569587242672398336           neutral                        1.0000
14638  569587188687634433          negative                        1.0000
14639  569587140490866689           neutral                        0.6771

               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
...                       ...                        ...             ...
14635                     NaN                     0.0000        American
14636  Customer Service Issue                     1.0000        American
14637                     NaN                        NaN        American
14638  Customer Service Issue                     0.6659        American
14639                     NaN                     0.0000        American

      airline_sentiment_gold             name negativereason_gold  \
0                        NaN          cairdin                 NaN
1                        NaN         jnardino                 NaN
2                        NaN       yvonnalynn                 NaN
3                        NaN         jnardino                 NaN
4                        NaN         jnardino                 NaN
...                      ...              ...                 ...
14635                    NaN  KristenReenders                 NaN
14636                    NaN         itsropes                 NaN
14637                    NaN         sanyabun                 NaN
14638                    NaN       SraJackson                 NaN
14639                    NaN        daviddtwu                 NaN

       retweet_count                                               text  \
0                  0                @VirginAmerica What @dhepburn said.
1                  0  @VirginAmerica plus you've added commercials t...
2                  0  @VirginAmerica I didn't today... Must mean I n...
3                  0  @VirginAmerica it's really aggressive to blast...
4                  0  @VirginAmerica and it's a really big bad thing...
...              ...                                                ...
14635              0  @AmericanAir thank you we got on a different f...
14636              0  @AmericanAir leaving over 20 minutes Late Flig...
14637              0  @AmericanAir Please bring American Airlines to...
14638              0  @AmericanAir you have my money, you change my ...
14639              0  @AmericanAir we have 8 ppl so we need 2 know h...

      tweet_coord              tweet_created tweet_location  \
0             NaN  2015-02-24 11:35:52 -0800            NaN
1             NaN  2015-02-24 11:15:59 -0800            NaN
2             NaN  2015-02-24 11:15:48 -0800      Lets Play
```

```
3              NaN  2015-02-24 11:15:36 -0800            NaN
4              NaN  2015-02-24 11:14:45 -0800            NaN
...            ...                        ...            ...
14635          NaN  2015-02-22 12:01:01 -0800            NaN
14636          NaN  2015-02-22 11:59:46 -0800          Texas
14637          NaN  2015-02-22 11:59:15 -0800  Nigeria,lagos
14638          NaN  2015-02-22 11:59:02 -0800     New Jersey
14639          NaN  2015-02-22 11:58:51 -0800     dallas, TX

                    user_timezone
0      Eastern Time (US & Canada)
1      Pacific Time (US & Canada)
2      Central Time (US & Canada)
3      Pacific Time (US & Canada)
4      Pacific Time (US & Canada)
...                           ...
14635                         NaN
14636                         NaN
14637                         NaN
14638  Eastern Time (US & Canada)
14639                         NaN

[14640 rows x 15 columns]>
```

In [4]:
```
data.describe()
```

Out[4]:

	tweet_id	airline_sentiment_confidence	negativereason_confidence	retweet_count
count	1.464000e+04	14640.000000	10522.000000	14640.000000
mean	5.692184e+17	0.900169	0.638298	0.082650
std	7.791112e+14	0.162830	0.330440	0.745778
min	5.675883e+17	0.335000	0.000000	0.000000
25%	5.685592e+17	0.692300	0.360600	0.000000
50%	5.694779e+17	1.000000	0.670600	0.000000
75%	5.698905e+17	1.000000	1.000000	0.000000
max	5.703106e+17	1.000000	1.000000	44.000000

In [5]:
```
data.shape
```

Out[5]: (14640, 15)

In [6]:
```
data.size
```

Out[6]: 219600

In [7]:
```
data.corr()
```

Out[7]:

	tweet_id	airline_sentiment_confidence	negativereason_confidence
tweet_id	1.000000	0.024840	0.021533
airline_sentiment_confidence	0.024840	1.000000	0.685879
negativereason_confidence	0.021533	0.685879	1.000000
retweet_count	-0.008852	0.012581	0.021574

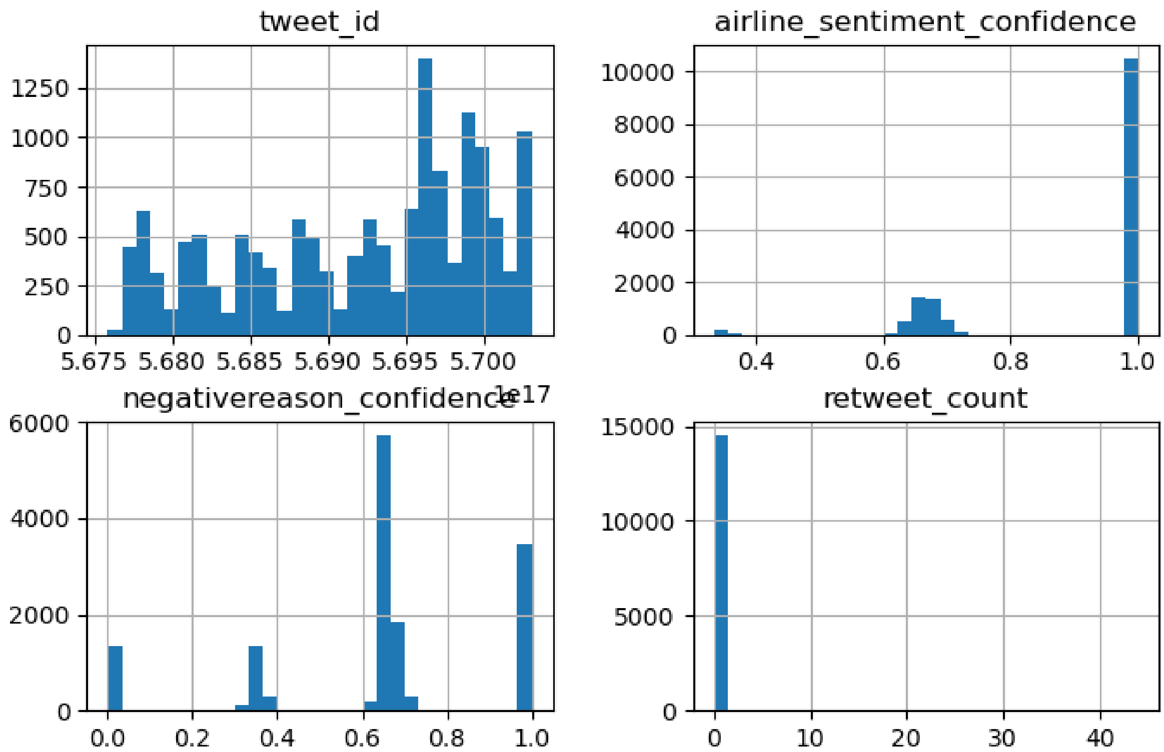
In [8]:
```
data.isnull().sum()
```

Out[8]:
```
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
```

In [9]:
```
# Deal with missing values
def deal_missing_values(x_full):
    x_full=x_full.drop("airline_sentiment_gold",axis=1)
    x_full=x_full.drop("negativereason_gold",axis=1)
    x_full=x_full.drop("tweet_coord",axis=1)
    #replace null values with mean
    x_full["negativereason_confidence"]= x_full["negativereason_confidence"].f
    return x_full
data=deal_missing_values(data)
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14640 entries, 0 to 14639
Data columns (total 12 columns):
 #   Column                        Non-Null Count  Dtype  
---  ------                        --------------  -----  
 0   tweet_id                      14640 non-null  int64  
 1   airline_sentiment             14640 non-null  object 
 2   airline_sentiment_confidence  14640 non-null  float64
 3   negativereason                9178 non-null   object 
 4   negativereason_confidence     14640 non-null  float64
 5   airline                       14640 non-null  object 
 6   name                          14640 non-null  object 
 7   retweet_count                 14640 non-null  int64  
 8   text                          14640 non-null  object 
 9   tweet_created                 14640 non-null  object 
 10  tweet_location                9907 non-null   object 
 11  user_timezone                 9820 non-null   object 
dtypes: float64(2), int64(2), object(8)
memory usage: 1.3+ MB
```

In [10]:
```python
## visualizing data to get better inslights
import matplotlib.pyplot as plt
data.hist(bins=30,figsize=(8,5))
plt.show()
```



In [11]:
```python
data["airline"].unique()
```

Out[11]:
```
array(['Virgin America', 'United', 'Southwest', 'Delta', 'US Airways',
       'American'], dtype=object)
```

In [12]:
```python
data["negativereason"].unique()
```

Out[12]:
```
array([nan, 'Bad Flight', "Can't Tell", 'Late Flight',
       'Customer Service Issue', 'Flight Booking Problems',
       'Lost Luggage', 'Flight Attendant Complaints', 'Cancelled Flight',
       'Damaged Luggage', 'longlines'], dtype=object)
```

In [13]:
```
data.head()
```

Out[13]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negative
0	570306133677760513	neutral	1.0000	NaN	
1	570301130888122368	positive	0.3486	NaN	
2	570301083672813571	neutral	0.6837	NaN	
3	570301031407624196	negative	1.0000	Bad Flight	
4	570300817074462722	negative	1.0000	Can't Tell	

In [14]:
```
data.tail()
```

Out[14]:

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	nega
14635	569587686496825344	positive	0.3487	NaN	
14636	569587371693355008	negative	1.0000	Customer Service Issue	
14637	569587242672398336	neutral	1.0000	NaN	
14638	569587188687634433	negative	1.0000	Customer Service Issue	
14639	569587140490866689	neutral	0.6771	NaN	

In [15]:
```
x=data.drop("airline_sentiment",axis=1)
y=data["airline_sentiment"]
```

In [16]:
```
## Split the feature and labels also training and test set
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=
```

In [17]:
```
# one hot encode"airline" attribute
from sklearn.compose import make_column_transformer
from sklearn.preprocessing import OneHotEncoder,MinMaxScaler
ct=make_column_transformer((MinMaxScaler(),["tweet_id"]))
# get all values between 0 and 1
(OneHotEncoder(handle_unknown="ignore"),["airline","retweet_count"])
ct.fit(x_train)
x_train_normal=ct.transform(x_train)
x_test_normal=ct.transform(x_test)
```

by using LogisticRegression

In [18]:
```
# now our data is ready to feed into the model
from sklearn.linear_model import LogisticRegression
l1=LogisticRegression(max_iter=1000)
l1.fit(x_train_normal,y_train)
```

Out[18]:
```
LogisticRegression(max_iter=1000)
```

In [19]:
```
y_pred=l1.predict(x_test_normal)
print(y_pred)
```

```
['negative' 'negative' 'negative' ... 'negative' 'negative' 'negative']
```

In [20]:
```
## Find the accuracy of the model
from sklearn.metrics import accuracy_score
accl=accuracy_score(y_pred,y_test)*100
print(accl)
```

```
60.82650273224044
```

By using SVM Model

In [21]:
```
from sklearn.svm import SVC
s1=SVC()
s1.fit(x_train_normal,y_train)
```

Out[21]:
```
SVC()
```

In [22]:
```
y_preds=s1.predict(x_test_normal)
print(y_preds)
```

```
['negative' 'negative' 'negative' ... 'negative' 'negative' 'negative']
```

In [23]:
```
accs=accuracy_score(y_preds,y_test)*100
print(accs)
```

```
60.82650273224044
```

By using Desicion Tree

In [24]:
```
from sklearn.tree import DecisionTreeClassifier
d1=DecisionTreeClassifier()
d1.fit(x_train_normal,y_train)
```

Out[24]:
```
DecisionTreeClassifier()
```

In [25]:
```
y_predD=d1.predict(x_test_normal)
print(y_predD)
```

```
['positive' 'negative' 'neutral' ... 'negative' 'positive' 'negative']
```

In [26]:
```
accD=accuracy_score(y_predD,y_test)*100
print(accD)
```

```
48.83879781420765
```

Thank you

In [ ]:
```

```