```
import kagglehub

# Download latest version
path = kagglehub.dataset_download("jp797498e/twitter-entity-sentiment-analysis")

print("Path to dataset files:", path)
```

```
Warning: Looks like you're using an outdated `kagglehub` version, please consider updating (latest vers
Downloading from https://www.kaggle.com/api/v1/datasets/download/jp797498e/twitter-entity-sentiment-ana
100%|██████████| 1.99M/1.99M [00:00<00:00, 35.2MB/s]Extracting model files...

Path to dataset files: /root/.cache/kagglehub/datasets/jp797498e/twitter-entity-sentiment-analysis/vers
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
import os


col=['Id' , 'Entity' , 'Sentiment' , 'Content']
df_train=pd.read_csv("/content/twitter_training.csv" , names=col)


df_test=pd.read_csv("/content/twitter_validation.csv" , names=col)


df_train
```

	Id	Entity	Sentiment	Content
0	2401	Borderlands	Positive	im getting on borderlands and i will murder yo...
1	2401	Borderlands	Positive	I am coming to the borders and I will kill you...
2	2401	Borderlands	Positive	im getting on borderlands and i will kill you ...
3	2401	Borderlands	Positive	im coming on borderlands and i will murder you...
4	2401	Borderlands	Positive	im getting on borderlands 2 and i will murder ...
...	...	...	...	...
74677	9200	Nvidia	Positive	Just realized that the Windows partition of my...
74678	9200	Nvidia	Positive	Just realized that my Mac window partition is ...
74679	9200	Nvidia	Positive	Just realized the windows partition of my Mac ...
74680	9200	Nvidia	Positive	Just realized between the windows partition of...
74681	9200	Nvidia	Positive	Just like the windows partition of my Mac is l...

74682 rows × 4 columns

Next steps: Generate code with df_train | View recommended plots | New interactive sheet

```
##data summary
```

```
df_train.shape
```

```
(74682, 4)
```

```
df_train.columns
```

```
Index(['Id', 'Entity', 'Sentiment', 'Content'], dtype='object')
```

```
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 74682 entries, 0 to 74681
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype 
---  ------     --------------  ----- 
 0   Id         74682 non-null  int64 
 1   Entity     74682 non-null  object
 2   Sentiment  74682 non-null  object
 3   Content    73996 non-null  object
dtypes: int64(1), object(3)
memory usage: 2.3+ MB
```

```
df_train.dtypes
```

	0
Id	int64
Entity	object
Sentiment	object
Content	object

```
##datacleaning
```

```
df_train.isnull().sum()
```

	0
Id	0
Entity	0
Sentiment	0
Content	686

```
df_train.dropna(subset=['Content'] , inplace=True)
```

```
df_train.shape
```

```
(73996, 4)
```

```
df_train.Sentiment.unique()
```

```
array(['Positive', 'Neutral', 'Negative', 'Irrelevant'], dtype=object)
```

```
df_train.Sentiment=df_train.Sentiment.replace('Irrelevant' , 'Neutral')
df_test.Sentiment=df_test.Sentiment.replace('Irrelevant' , 'Neutral')
```

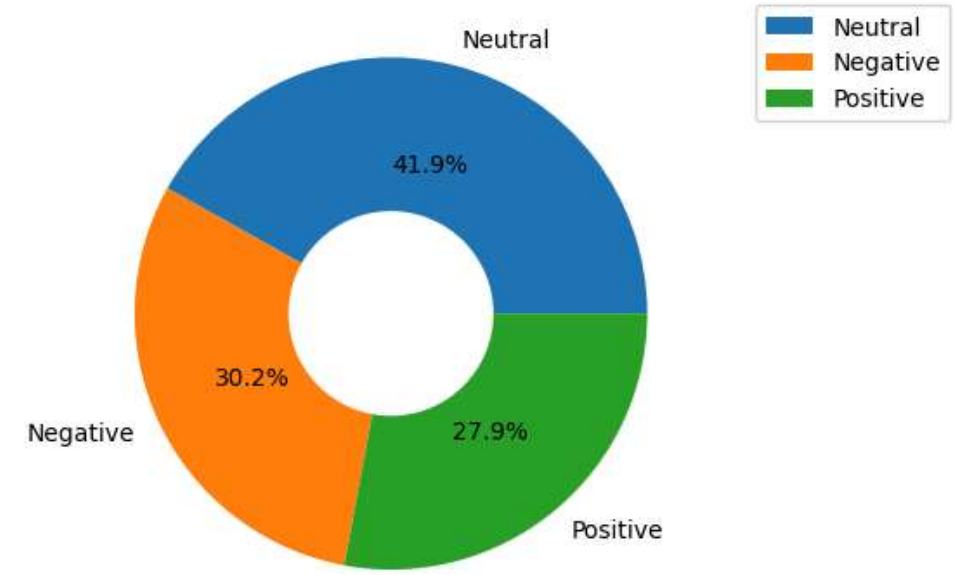
```
df_train.Sentiment.unique()
```

```
array(['Positive', 'Neutral', 'Negative'], dtype=object)
```

```
sentiment_count=df_train.Sentiment.value_counts()
sentiment_count
```

Sentiment	count
Neutral	30983
Negative	22358
Positive	20655

```
y=['Neutral' , 'Negative' , 'Positive']
plt.pie(sentiment_count , labels=y, autopct='%0.1f%%' )
circle=plt.Circle((0,0),0.4, facecolor='white')
plt.gca().add_patch(circle)
plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left')
plt.show()
```




```
df_train.Entity.unique()
```

```
array(['Borderlands', 'CallOfDutyBlackopsColdWar', 'Amazon', 'Overwatch',
       'Xbox(Xseries)', 'NBA2K', 'Dota2', 'PlayStation5(PS5)',
       'WorldOfCraft', 'CS-GO', 'Google', 'AssassinsCreed', 'ApexLegends',
       'LeagueOfLegends', 'Fortnite', 'Microsoft', 'Hearthstone',
       'Battlefield', 'PlayerUnknownsBattlegrounds(PUBG)', 'Verizon',
       'HomeDepot', 'FIFA', 'RedDeadRedemption(RDR)', 'CallOfDuty',
       'TomClancysRainbowSix', 'Facebook', 'GrandTheftAuto(GTA)',
       'MaddenNFL', 'johnson&johnson', 'Cyberpunk2077',
       'TomClancysGhostRecon', 'Nvidia'], dtype=object)
```

```
Entity_count=df_train.Entity.value_counts()
Entity_count
```

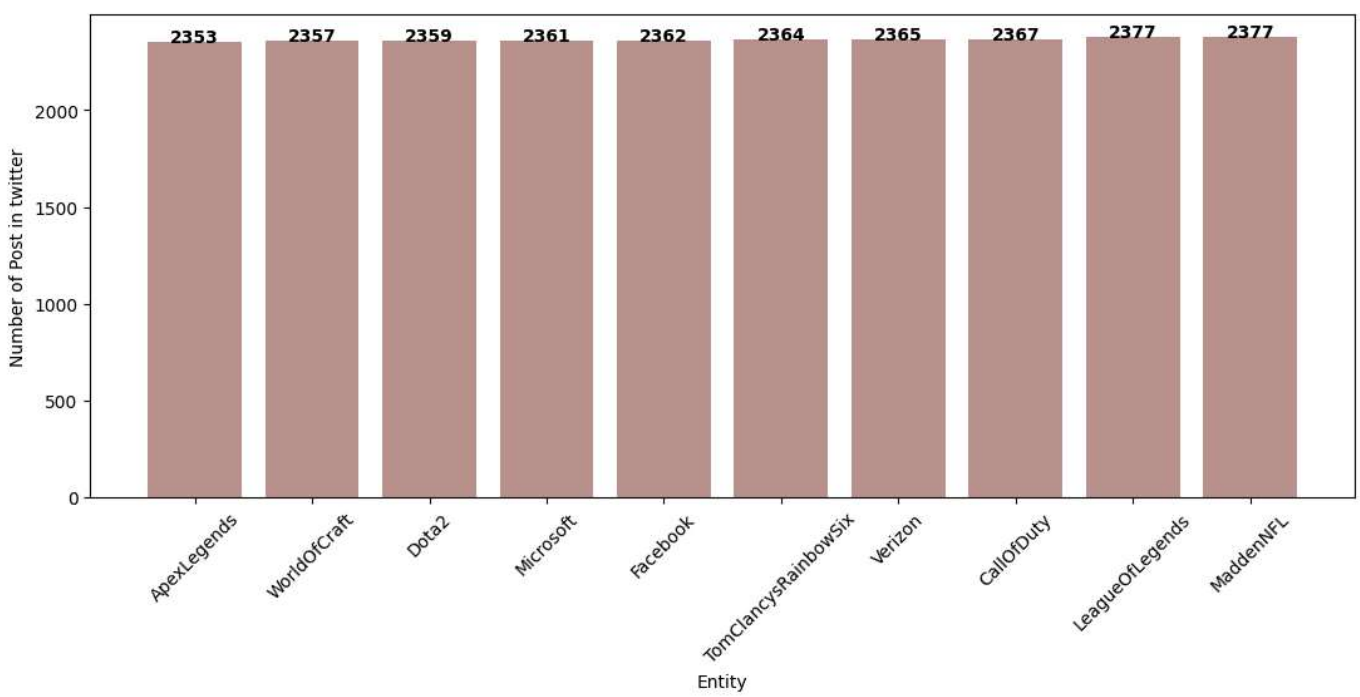
Entity	
MaddenNFL	2377
LeagueOfLegends	2377
CallOfDuty	2376
Verizon	2365
TomClancysRainbowSix	2364
Facebook	2362
Microsoft	2361
Dota2	2359
WorldOfCraft	2357
ApexLegends	2353
NBA2K	2343
CallOfDutyBlackopsColdWar	2343
FIFA	2324
johnson&johnson	2324
TomClancysGhostRecon	2321
Battlefield	2316
Overwatch	2316
GrandTheftAuto(GTA)	2293
HomeDepot	2292
PlayStation5(PS5)	2291
Hearthstone	2286
CS-GO	2284
Xbox(Xseries)	2283
Borderlands	2280
Amazon	2276
Google	2274
Nvidia	2271
Cyberpunk2077	2262
RedDeadRedemption(RDR)	2249
Fortnite	2249
PlayerUnknownsBattlegrounds(PUBG)	2234
AssassinsCreed	2234

dtype: int64

```
Entity_sort=Entity_count.sort_values(ascending=False)


Entity_top10=Entity_sort.head(10)
Entity_top10
```

Entity	count
MaddenNFL	2377
LeagueOfLegends	2377
CallOfDuty	2376
Verizon	2365
TomClancysRainbowSix	2364
Facebook	2362
Microsoft	2361
Dota2	2359
WorldOfCraft	2357
ApexLegends	2353

```
Entity_index=Entity_top10.index


plt.figure(figsize=(13,5))

x=['ApexLegends' , 'WorldOfCraft' , 'Dota2' , 'Microsoft' , 'Facebook' , 'TomClancysRainbowSix' , 'Verizon'
y=[2353,2357,2359,2361,2362,2364,2365,2367,2377,2377]

plt.bar( x , y , alpha=0.7 , color='#A2625D')

for i,v in enumerate(y):
    plt.text(i,v,str(v),ha='center',weight='bold' )

plt.xticks(rotation=45)
plt.xlabel('Entity')
plt.ylabel('Number of Post in twitter')
plt.show()
```




```
Entity_top3_df=Entity_sort.head(3)
Entity_top3_df
```

Entity	count
MaddenNFL	2377
LeagueOfLegends	2377
CallOfDuty	2376

```
Entity_top3=Entity_top3_df.index.tolist()
Entity_top3
```

```
['MaddenNFL', 'LeagueOfLegends', 'CallOfDuty']
```

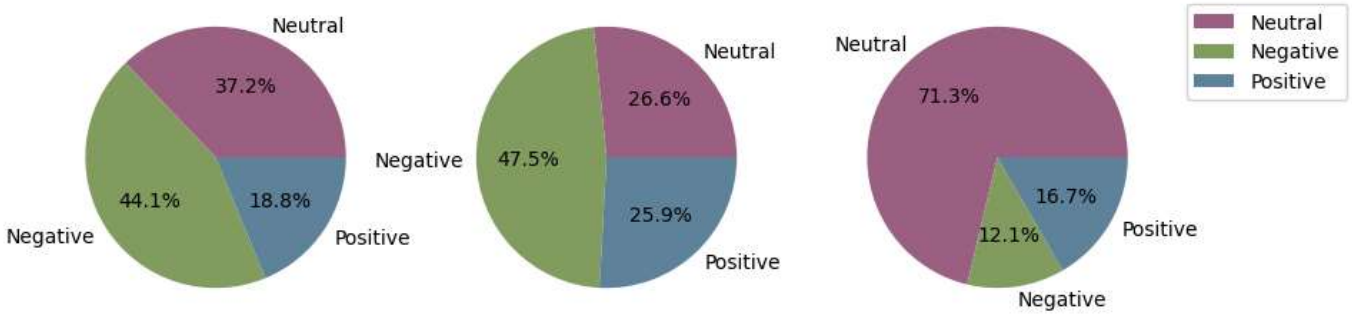
```
sentiment_by_entity=df_train.loc[df_train['Entity'].isin(Entity_top3)].groupby('Entity')['Sentiment'].value
sentiment_by_entity
```

Entity	Sentiment	count
CallOfDuty	Negative	883
	Neutral	1047
	Positive	446
LeagueOfLegends	Negative	632
	Neutral	1130
	Positive	615
MaddenNFL	Negative	1694
	Neutral	287
	Positive	396

dtype: int64

```
##model


plt.figure(figsize=(10,5))

y=['Neutral' , 'Negative' , 'Positive']
color=['#9C6383' , '#839C63' , '#63839C']

plt.subplot(1,3,1)
plt.pie(sentiment_by_entity[:3] , labels=y , autopct='%0.1f%%' , textprops={'fontsize':10} , colors=color)

plt.subplot(1,3,2)
plt.pie(sentiment_by_entity[3:6] , labels=y , autopct='%0.1f%%' , textprops={'fontsize':10} , colors=color)

plt.subplot(1,3,3)
plt.pie(sentiment_by_entity[6:] , labels=y , autopct='%0.1f%%' , textprops={'fontsize':10} , colors=color)

plt.legend(bbox_to_anchor=(1.05, 1), loc='upper left' , fontsize="10")
```

```
<matplotlib.legend.Legend at 0x7f225c4b8c70>
```




```
df_train
```

	Id	Entity	Sentiment	Content
0	2401	Borderlands	Positive	im getting on borderlands and i will murder yo...
1	2401	Borderlands	Positive	I am coming to the borders and I will kill you...
2	2401	Borderlands	Positive	im getting on borderlands and i will kill you ...
3	2401	Borderlands	Positive	im coming on borderlands and i will murder you...
4	2401	Borderlands	Positive	im getting on borderlands 2 and i will murder ...
...	...	...	...	...
74677	9200	Nvidia	Positive	Just realized that the Windows partition of my...
74678	9200	Nvidia	Positive	Just realized that my Mac window partition is ...
74679	9200	Nvidia	Positive	Just realized the windows partition of my Mac ...
74680	9200	Nvidia	Positive	Just realized between the windows partition of...
74681	9200	Nvidia	Positive	Just like the windows partition of my Mac is l...

Next steps: Generate code with df_train View recommended plots New interactive sheet

73996 rows × 4 columns

```
df_train.drop(['Id'] , axis=1 , inplace=True)
```

```
df_test.drop(['Id'] , axis=1 , inplace=True)
```

```
#train test split
X_train=df_train.drop(['Sentiment'] , axis=1)
X_test=df_test.drop(['Sentiment'] , axis=1)
y_train=df_train['Sentiment']
y_test=df_test['Sentiment']
```

```
df_train.Sentiment.unique()
```

```
array(['Positive', 'Neutral', 'Negative'], dtype=object)
```

```
#count the no of words in a sentence
from sklearn.feature_extraction.text import CountVectorizer
```

```
v=CountVectorizer()
X_train_count=v.fit_transform(X_train.Content)
```

```
#label Encoding
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
y_train=le.fit_transform(y_train)
y_test=le.fit_transform(y_test)
```