

twitter

June 4, 2024

```
[3]: import pandas as pd
import numpy as np
from nltk.tokenize import sent_tokenize, word_tokenize
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.svm import SVC
from sklearn.datasets import fetch_20newsgroups
from nltk.corpus import stopwords
import string
from nltk import pos_tag
from nltk.stem import WordNetLemmatizer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn import preprocessing
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
```

```
[2]: import nltk
nltk.download('stopwords')
```

```
[nltk_data] Error loading stopwords: <urlopen error [WinError 10060] A
[nltk_data]      connection attempt failed because the connected party
[nltk_data]      did not properly respond after a period of time, or
[nltk_data]      established connection failed because connected host
[nltk_data]      has failed to respond>
```

```
[2]: False
```

```
[7]: data = pd.read_csv('C:/Users/SRAVANTHI/prodigy infoTech/Task 04/
↳twitter_training.csv/twitter_training.csv')
v_data = pd.read_csv('C:/Users/SRAVANTHI/prodigy infoTech/Task 04/
↳twitter_validation.csv')
```

```
[8]: data
```

```
[8]:      2401  Borderlands  Positive  \
0      2401  Borderlands  Positive
1      2401  Borderlands  Positive
2      2401  Borderlands  Positive
3      2401  Borderlands  Positive
4      2401  Borderlands  Positive
...      ...      ...      ...
74676  9200      Nvidia  Positive
74677  9200      Nvidia  Positive
74678  9200      Nvidia  Positive
74679  9200      Nvidia  Positive
74680  9200      Nvidia  Positive

      im getting on borderlands and i will murder you all ,
0      I am coming to the borders and I will kill you...
1      im getting on borderlands and i will kill you ...
2      im coming on borderlands and i will murder you...
3      im getting on borderlands 2 and i will murder ...
4      im getting into borderlands and i can murder y...
...      ...
74676  Just realized that the Windows partition of my...
74677  Just realized that my Mac window partition is ...
74678  Just realized the windows partition of my Mac ...
74679  Just realized between the windows partition of...
74680  Just like the windows partition of my Mac is l...

[74681 rows x 4 columns]
```

```
[9]: v_data
```

```
[9]:      3364      Facebook  Irrelevant  \
0      352      Amazon      Neutral
1      8312      Microsoft  Negative
2      4371      CS-GO      Negative
3      4433      Google      Neutral
4      6273      FIFA      Negative
..      ...      ...      ...
994  4891  GrandTheftAuto(GTA)  Irrelevant
995  4359      CS-GO      Irrelevant
996  2652      Borderlands  Positive
997  8069      Microsoft  Positive
998  6960      johnson&johnson  Neutral
```

I mentioned on Facebook that I was struggling for motivation to go for a run the other day, which has been translated by Tom's great auntie as 'Hayley can't

get out of bed' and told to his grandma, who now thinks I'm a lazy, terrible person

```
0    BBC News - Amazon boss Jeff Bezos rejects clai...
1    @Microsoft Why do I pay for WORD when it funct...
2    CSGO matchmaking is so full of closet hacking,...
3    Now the President is slapping Americans in the...
4    Hi @EAHelp I've had Madeleine McCann in my cel...
..
994   Toronto is the arts and culture capital of ...
995   tHIS IS ACTUALLY A GOOD MOVE TOT BRING MORE VI...
996   Today sucked so it's time to drink wine n play...
997   Bought a fraction of Microsoft today. Small wins.
998   Johnson & Johnson to stop selling talc baby po...
```

[999 rows x 4 columns]

```
[10]: data.columns = ['id', 'game', 'sentiment', 'text']
      v_data.columns = ['id', 'game', 'sentiment', 'text']
```

```
[11]: data
```

```
[11]:      id      game sentiment \
0      2401  Borderlands  Positive
1      2401  Borderlands  Positive
2      2401  Borderlands  Positive
3      2401  Borderlands  Positive
4      2401  Borderlands  Positive
...    ...      ...      ...
74676  9200      Nvidia  Positive
74677  9200      Nvidia  Positive
74678  9200      Nvidia  Positive
74679  9200      Nvidia  Positive
74680  9200      Nvidia  Positive

                                     text
0      I am coming to the borders and I will kill you...
1      im getting on borderlands and i will kill you ...
2      im coming on borderlands and i will murder you...
3      im getting on borderlands 2 and i will murder ...
4      im getting into borderlands and i can murder y...
...
74676  Just realized that the Windows partition of my...
74677  Just realized that my Mac window partition is ...
74678  Just realized the windows partition of my Mac ...
74679  Just realized between the windows partition of...
74680  Just like the windows partition of my Mac is l...
```

[74681 rows x 4 columns]

```
[12]: v_data
```

```
[12]:      id      game  sentiment \
0    352      Amazon    Neutral
1   8312    Microsoft    Negative
2   4371      CS-GO    Negative
3   4433      Google    Neutral
4   6273      FIFA    Negative
..    ...      ...      ...
994  4891  GrandTheftAuto(GTA)  Irrelevant
995  4359      CS-GO  Irrelevant
996  2652    Borderlands    Positive
997  8069    Microsoft    Positive
998  6960  johnson&johnson    Neutral

                                     text
0  BBC News - Amazon boss Jeff Bezos rejects clai...
1  @Microsoft Why do I pay for WORD when it funct...
2  CSGO matchmaking is so full of closet hacking,...
3  Now the President is slapping Americans in the...
4  Hi @EAHelp I've had Madeleine McCann in my cel...
..
994  Toronto is the arts and culture capital of ...
995  tHIS IS ACTUALLY A GOOD MOVE TOT BRING MORE VI...
996  Today sucked so it's time to drink wine n play...
997  Bought a fraction of Microsoft today. Small wins.
998  Johnson & Johnson to stop selling talc baby po...
```

[999 rows x 4 columns]

```
[13]: data.shape
```

```
[13]: (74681, 4)
```

```
[14]: data.columns
```

```
[14]: Index(['id', 'game', 'sentiment', 'text'], dtype='object')
```

```
[15]: data.describe(include='all')
```

```
[15]:
```

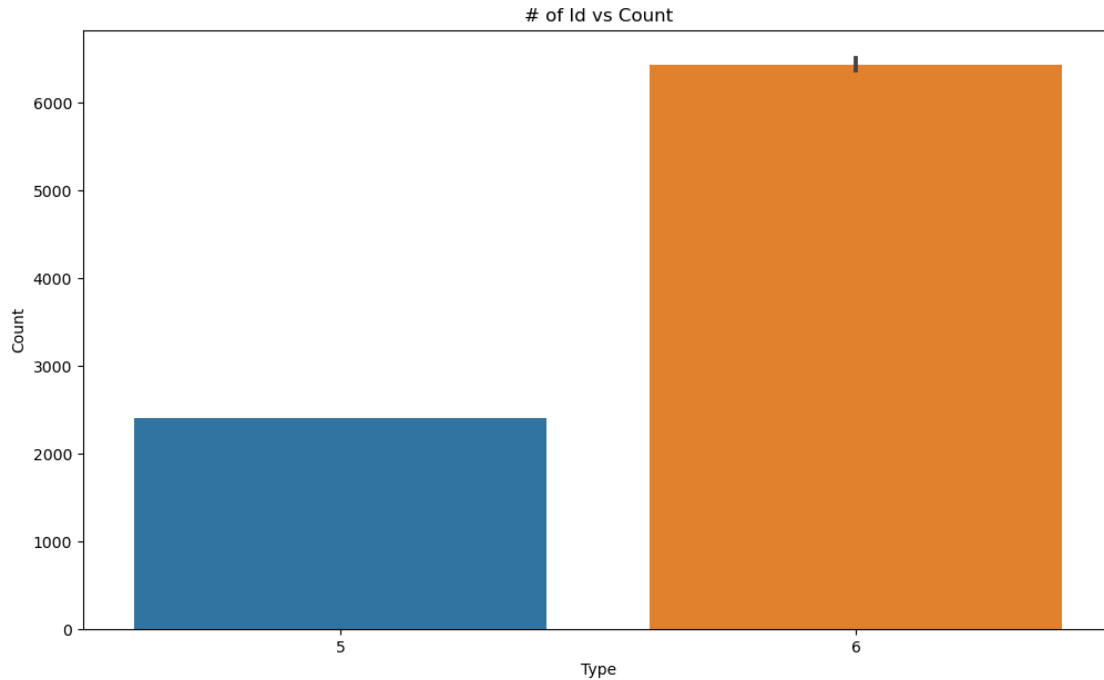
	id	game	sentiment	text
count	74681.000000	74681	74681	73995
unique	NaN	32	4	69490
top	NaN	TomClancysRainbowSix	Negative	
freq	NaN	2400	22542	172

mean	6432.640149	NaN	NaN	NaN
std	3740.423819	NaN	NaN	NaN
min	1.000000	NaN	NaN	NaN
25%	3195.000000	NaN	NaN	NaN
50%	6422.000000	NaN	NaN	NaN
75%	9601.000000	NaN	NaN	NaN
max	13200.000000	NaN	NaN	NaN

```
[16]: id_types = data['id'].value_counts()
      id_types
```

```
[16]: id
5203    6
6164    6
6141    6
6142    6
6143    6
..
4678    6
4679    6
4680    6
4681    6
2401    5
Name: count, Length: 12447, dtype: int64
```

```
[17]: plt.figure(figsize=(12,7))
      sns.barplot(y=id_types.index, x=id_types.values)
      plt.xlabel('Type')
      plt.ylabel('Count')
      plt.title('# of Id vs Count')
      plt.show()
```



```
[18]: game_types = data['game'].value_counts()
      game_types
```

```
[18]: game
      TomClancysRainbowSix      2400
      MaddenNFL              2400
      Microsoft              2400
      LeagueOfLegends        2394
      CallOfDuty             2394
      Verizon                2382
      CallOfDutyBlackopsColdWar 2376
      ApexLegends            2376
      Facebook               2370
      WorldOfCraft           2364
      Dota2                  2364
      NBA2K                  2352
      TomClancysGhostRecon    2346
      Battlefield            2346
      FIFA                   2340
      Xbox(Xseries)          2334
      Overwatch              2334
      johnson&johnson         2328
      Amazon                 2316
      PlayStation5(PS5)      2310
      HomeDepot              2310
```

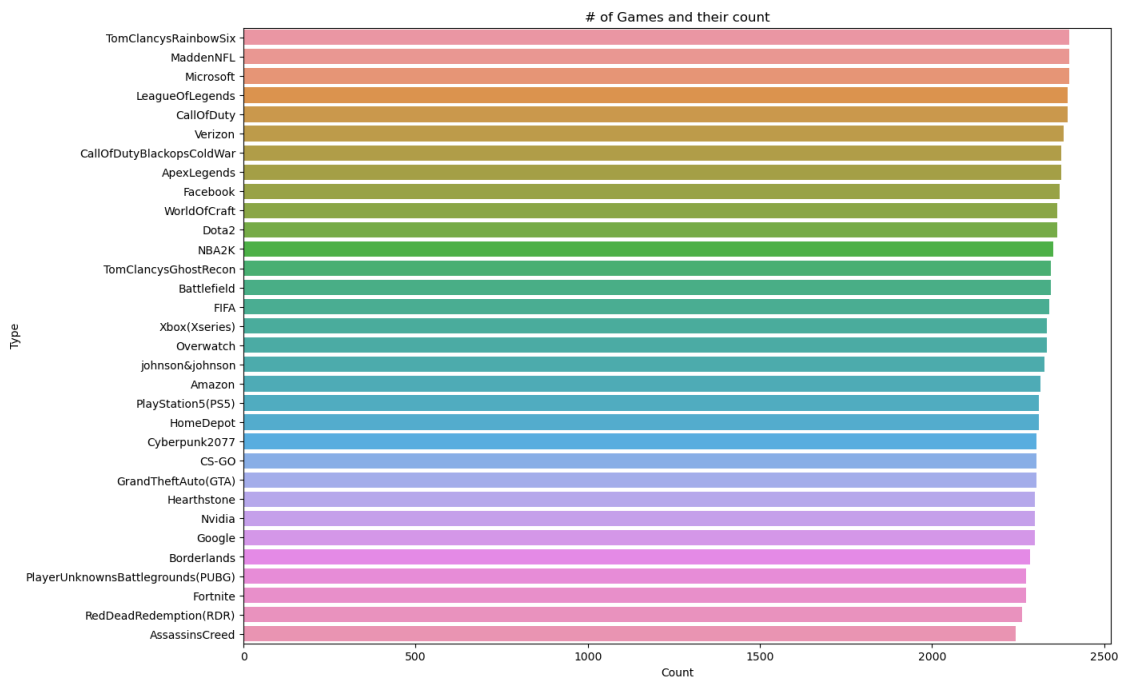
Cyberpunk2077	2304
CS-GO	2304
GrandTheftAuto(GTA)	2304
Hearthstone	2298
Nvidia	2298
Google	2298
Borderlands	2285
PlayerUnknownsBattlegrounds(PUBG)	2274
Fortnite	2274
RedDeadRedemption(RDR)	2262
AssassinsCreed	2244

Name: count, dtype: int64

```
[19]: plt.figure(figsize=(14,10))

sns.barplot(x=game_types.values,y=game_types.index)
plt.title('# of Games and their count')
plt.ylabel('Type')
plt.xlabel('Count')

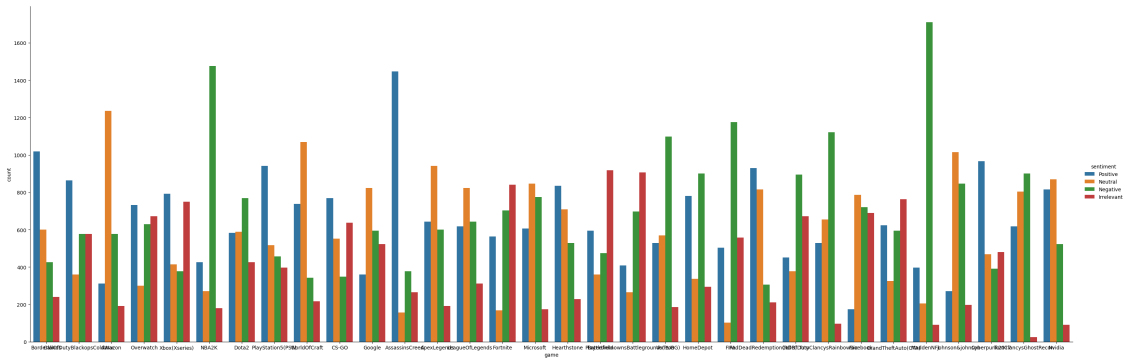
plt.show()
```



```
[20]: sns.catplot(x="game",hue="sentiment", kind="count",height=10,aspect=3,
↪data=data)
```

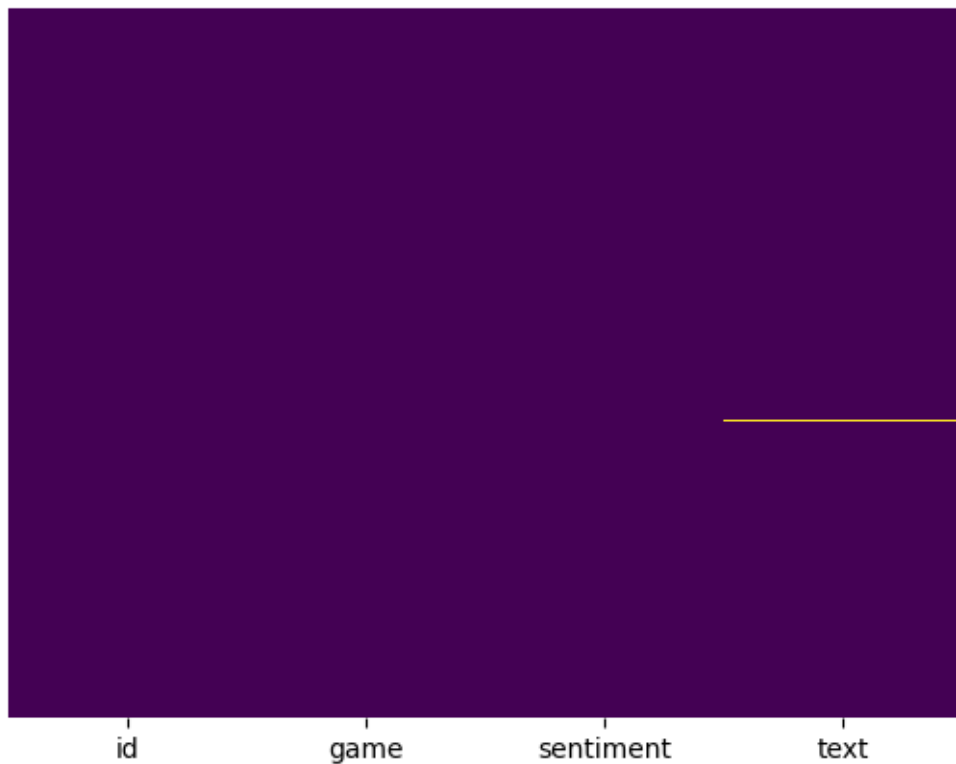
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning:
The figure layout has changed to tight
self._figure.tight_layout(*args, **kwargs)

[20]: <seaborn.axisgrid.FacetGrid at 0x1d52d7ba6d0>



[21]: sns.heatmap(data.isnull(),yticklabels=False,cbar=False,cmap='viridis')

[21]: <Axes: >




```
[22]: total_null=data.isnull().sum().sort_values(ascending=False)
percent = ((data.isnull().sum()/data.isnull().count()*100).
↳sort_values(ascending = False)
print("Total records = ", data.shape[0])
missing_data = pd.concat([total_null,percent.round(2)],axis=1,keys=['Total_
↳Missing','In Percent'])
missing_data.head(10)
```

Total records = 74681

```
[22]:
```

	Total Missing	In Percent
text	686	0.92
id	0	0.00
game	0	0.00
sentiment	0	0.00

```
[23]: data.dropna(subset=['text'],inplace=True)

total_null=data.isnull().sum().sort_values(ascending=False)
percent = ((data.isnull().sum()/data.isnull().count()*100).
↳sort_values(ascending = False)
print("Total records = ", data.shape[0])
missing_data = pd.concat([total_null,percent.round(2)],axis=1,keys=['Total_
↳Missing','In Percent'])
missing_data.head(10)
```

Total records = 73995

```
[23]:
```

	Total Missing	In Percent
id	0	0.0
game	0	0.0
sentiment	0	0.0
text	0	0.0

```
[24]: train0=data[data['sentiment']=="Negative"]
train1=data[data['sentiment']=="Positive"]
train2=data[data['sentiment']=="Irrelevant"]
train3=data[data['sentiment']=="Neutral"]
```

```
[25]: train0.shape, train1.shape, train2.shape, train3.shape
```

```
[25]: ((22358, 4), (20654, 4), (12875, 4), (18108, 4))
```

```
[26]: train0=train0[:int(train0.shape[0]/12)]
train1=train1[:int(train1.shape[0]/12)]
train2=train2[:int(train2.shape[0]/12)]
train3=train3[:int(train3.shape[0]/12)]
```

```
[27]: train0.shape, train1.shape, train2.shape, train3.shape
```

```
[27]: ((1863, 4), (1721, 4), (1072, 4), (1509, 4))
```

```
[28]: data=pd.concat([train0,train1,train2,train3],axis=0)
data
```

```
[28]:
```

	id	game	sentiment	\
23	2405	Borderlands	Negative	
24	2405	Borderlands	Negative	
25	2405	Borderlands	Negative	
26	2405	Borderlands	Negative	
27	2405	Borderlands	Negative	
...	
5603	165	Amazon	Neutral	
5604	165	Amazon	Neutral	
5605	165	Amazon	Neutral	
5606	165	Amazon	Neutral	
5607	165	Amazon	Neutral	

	text
23	the biggest dissappointment in my life came out...
24	The biggest disappointment of my life came a y...
25	The biggest disappointment of my life came a y...
26	the biggest dissappointment in my life coming o...
27	For the biggest male dissapoinment in my life...
...	...
5603	An amazing read aloud book for you and your ch...
5604	An amazing reading book for you and your child...
5605	An amazing book to read aloud for you and your...
5606	An amazing read aloud book for you and your ch...
5607	and An amazing read aloud book for you and you...

[6165 rows x 4 columns]

```
[29]: id_types = data['id'].value_counts()
id_types
```

```
[29]: id
2405    6
1810    6
1748    6
1754    6
1760    6
..
1602    3
1880    3
```

```

333      3
9388      2
9267      2
Name: count, Length: 1040, dtype: int64

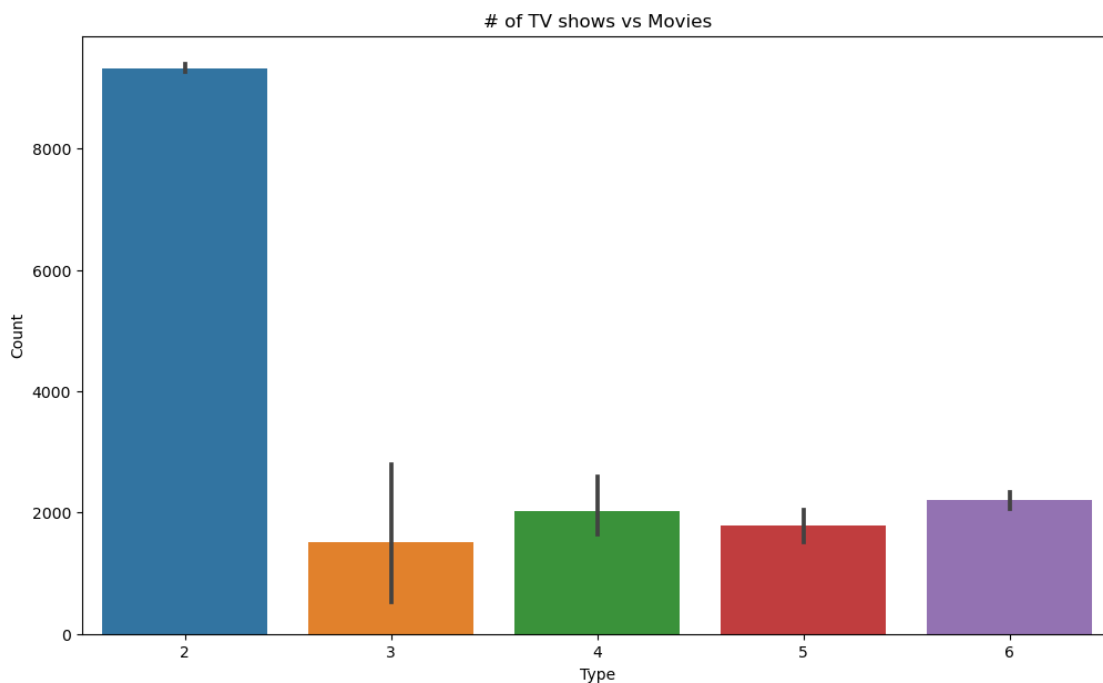
```

```

[30]: plt.figure(figsize=(12,7))
      sns.barplot(x=id_types.values,y=id_types.index)

      plt.xlabel('Type')
      plt.ylabel('Count')
      plt.title('# of TV shows vs Movies')
      plt.show()

```



```

[31]: game_types = data['game'].value_counts()
      game_types

```

```

[31]: game
      Borderlands      2279
      CallOfDutyBlackopsColdWar  2192
      Amazon      1321
      Overwatch      373
Name: count, dtype: int64

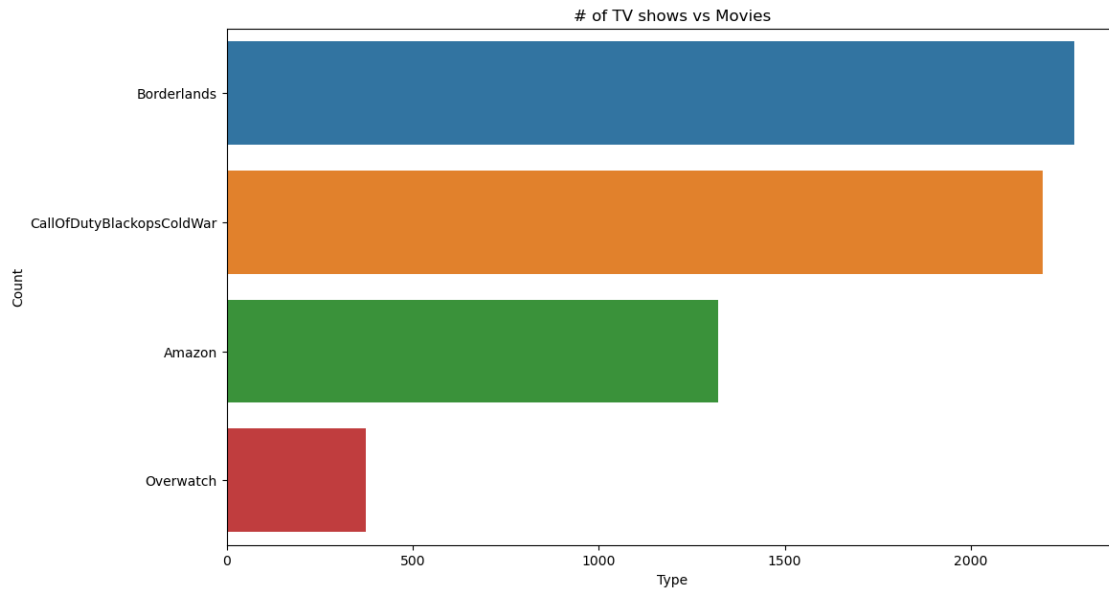
```

```

[32]: plt.figure(figsize=(12,7))
      sns.barplot(x=game_types.values,y=game_types.index)

```

```
plt.xlabel('Type')
plt.ylabel('Count')
plt.title('# of TV shows vs Movies')
plt.show()
```

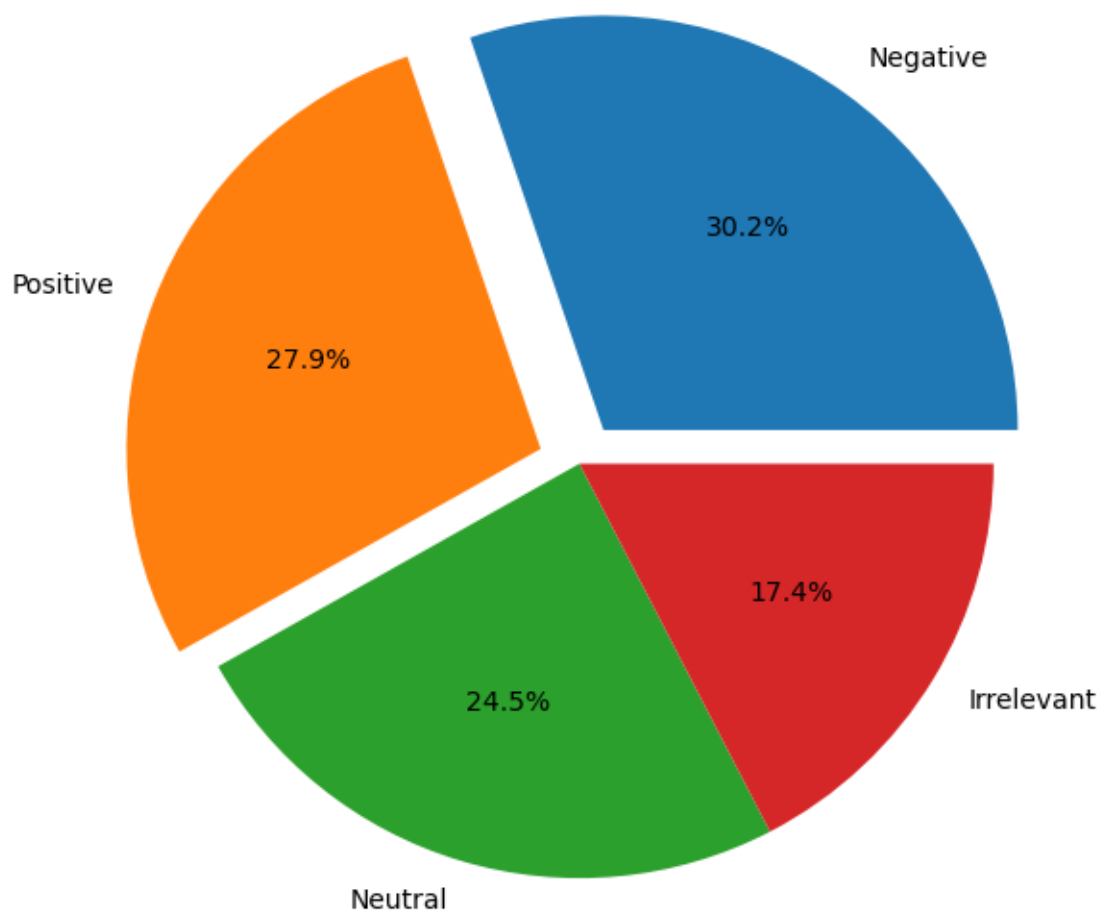


```
[33]: sentiment_types = data['sentiment'].value_counts()
sentiment_types
```

```
[33]: sentiment
Negative      1863
Positive      1721
Neutral       1509
Irrelevant    1072
Name: count, dtype: int64
```

```
[34]: plt.figure(figsize=(12,7))
plt.pie(x=sentiment_types.values, labels=sentiment_types.index, autopct='%.\
↪1f%%', explode=[0.1, 0.1,0,0])
plt.title('The Difference in the Type of Contents')
plt.show()
```

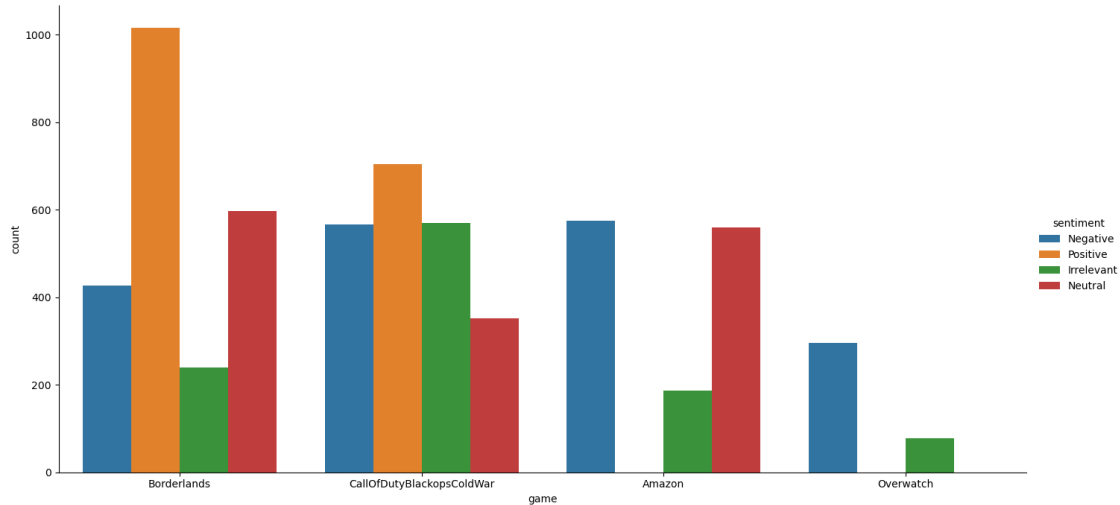
The Difference in the Type of Contents



```
[35]: sns.catplot(x='game',hue='sentiment',kind='count',height=7,aspect=2,data=data)
```

```
C:\ProgramData\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning:  
The figure layout has changed to tight  
  self._figure.tight_layout(*args, **kwargs)
```

```
[35]: <seaborn.axisgrid.FacetGrid at 0x1d52c323490>
```



```
[36]: from sklearn import preprocessing
label_encoder = preprocessing.LabelEncoder()
```

```
[37]: data['sentiment']=label_encoder.fit_transform(data['sentiment'])
data['game']=label_encoder.fit_transform(data['game'])
v_data['sentiment']=label_encoder.fit_transform(v_data['sentiment'])
v_data['game']=label_encoder.fit_transform(v_data['game'])
```

```
[38]: data = data.drop(['id'],axis=1)

data
```

```
[38]:
```

	game	sentiment	text
23	1	1	the biggest dissappointment in my life came out...
24	1	1	The biggest disappointment of my life came a y...
25	1	1	The biggest disappointment of my life came a y...
26	1	1	the biggest dissappointment in my life coming o...
27	1	1	For the biggest male dissappointment in my life...
...
5603	0	2	An amazing read aloud book for you and your ch...
5604	0	2	An amazing reading book for you and your child...
5605	0	2	An amazing book to read aloud for you and your...
5606	0	2	An amazing read aloud book for you and your ch...
5607	0	2	and An amazing read aloud book for you and you...

[6165 rows x 3 columns]

```
[39]: data.nunique()
```

```
[39]: game          4  
      sentiment     4  
      text        5854  
      dtype: int64
```

```
[40]: v_data.nunique()
```

```
[40]: id            999  
      game          32  
      sentiment     4  
      text         998  
      dtype: int64
```