In [1]:
```
!pip install numpy
!pip install pandas
!pip install matplotlib
!pip install seaborn
!pip install -U scikit-learn
```

```
Requirement already satisfied: numpy in c:\users\dell\anaconda3\lib\site-packages (1.24.3)
Requirement already satisfied: pandas in c:\users\dell\anaconda3\lib\site-packages (2.0.3)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\dell\anaconda3\lib\site-packages (from pandas)
(2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\dell\anaconda3\lib\site-packages (from pandas) (2023.3.post
1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\dell\anaconda3\lib\site-packages (from pandas) (2023.3)
Requirement already satisfied: numpy>=1.21.0 in c:\users\dell\anaconda3\lib\site-packages (from pandas) (1.24.3)
Requirement already satisfied: six>=1.5 in c:\users\dell\anaconda3\lib\site-packages (from python-dateutil>=2.8.2->p
andas) (1.16.0)
Requirement already satisfied: matplotlib in c:\users\dell\anaconda3\lib\site-packages (3.7.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib)
(1.0.5)
Requirement already satisfied: cycler>=0.10 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib) (4.2
5.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib)
(1.4.4)
Requirement already satisfied: numpy>=1.20 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib) (1.24.3)
Requirement already satisfied: packaging>=20.0 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib) (23.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing<3.1,>=2.3.1 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib)
(3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib)
(2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\dell\anaconda3\lib\site-packages (from python-dateutil>=2.7->mat
plotlib) (1.16.0)
Requirement already satisfied: seaborn in c:\users\dell\anaconda3\lib\site-packages (0.12.2)
Requirement already satisfied: numpy!=1.24.0,>=1.17 in c:\users\dell\anaconda3\lib\site-packages (from seaborn) (1.2
4.3)
Requirement already satisfied: pandas>=0.25 in c:\users\dell\anaconda3\lib\site-packages (from seaborn) (2.0.3)
Requirement already satisfied: matplotlib!=3.6.1,>=3.1 in c:\users\dell\anaconda3\lib\site-packages (from seaborn)
(3.7.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib!=
3.6.1,>=3.1->seaborn) (1.0.5)
Requirement already satisfied: cycler>=0.10 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=
3.1->seaborn) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib!=
3.6.1,>=3.1->seaborn) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib!=
3.6.1,>=3.1->seaborn) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib!=
3.6.1,>=3.1->seaborn) (23.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=
```

```
3.1->seaborn) (9.4.0)
Requirement already satisfied: pyparsing<3.1,>=2.3.1 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib!=
3.6.1,>=3.1->seaborn) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\dell\anaconda3\lib\site-packages (from matplotlib!=
3.6.1,>=3.1->seaborn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\dell\anaconda3\lib\site-packages (from pandas>=0.25->seabor
n) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\dell\anaconda3\lib\site-packages (from pandas>=0.25->seabo
rn) (2023.3)
Requirement already satisfied: six>=1.5 in c:\users\dell\anaconda3\lib\site-packages (from python-dateutil>=2.7->mat
plotlib!=3.6.1,>=3.1->seaborn) (1.16.0)
Requirement already satisfied: scikit-learn in c:\users\dell\anaconda3\lib\site-packages (1.3.0)
Collecting scikit-learn
  Obtaining dependency information for scikit-learn from https://files.pythonhosted.org/packages/ae/20/6d1a0a61d468b
37a142fd90bb93c73bc1c2205db4a69ac630ed218c31612/scikit_learn-1.5.0-cp311-cp311-win_amd64.whl.metadata
  Downloading scikit_learn-1.5.0-cp311-cp311-win_amd64.whl.metadata (11 kB)
Requirement already satisfied: numpy>=1.19.5 in c:\users\dell\anaconda3\lib\site-packages (from scikit-learn) (1.2
4.3)
Requirement already satisfied: scipy>=1.6.0 in c:\users\dell\anaconda3\lib\site-packages (from scikit-learn) (1.1
1.1)
Requirement already satisfied: joblib>=1.2.0 in c:\users\dell\anaconda3\lib\site-packages (from scikit-learn)
(1.2.0)
Collecting threadpoolctl>=3.1.0 (from scikit-learn)
  Obtaining dependency information for threadpoolctl>=3.1.0 from https://files.pythonhosted.org/packages/4b/2c/ffbf7
a134b9ab11a67b0cf0726453cedd9c5043a4fe7a35d1cefa9a1bcfb/threadpoolctl-3.5.0-py3-none-any.whl.metadata
  Downloading threadpoolctl-3.5.0-py3-none-any.whl.metadata (13 kB)
Downloading scikit_learn-1.5.0-cp311-cp311-win_amd64.whl (11.0 MB)
   ---------------------------------------- 0.0/11.0 MB ? eta -:--:--
   ---------------------------------------- 0.0/11.0 MB ? eta -:--:--
   ---------------------------------------- 0.0/11.0 MB ? eta -:--:--
   ---------------------------------------- 0.0/11.0 MB ? eta -:--:--
   ---------------------------------------- 0.0/11.0 MB ? eta -:--:--
   ---------------------------------------- 0.0/11.0 MB ? eta -:--:--
   ---------------------------------------- 0.0/11.0 MB ? eta -:--:--
   ---------------------------------------- 0.1/11.0 MB 241.3 kB/s eta 0:00:46
   ---------------------------------------- 0.1/11.0 MB 241.3 kB/s eta 0:00:46
   ---------------------------------------- 0.1/11.0 MB 245.8 kB/s eta 0:00:45
    --------------------------------------- 0.2/11.0 MB 374.1 kB/s eta 0:00:29
    --------------------------------------- 0.3/11.0 MB 496.1 kB/s eta 0:00:22
   - -------------------------------------- 0.3/11.0 MB 563.8 kB/s eta 0:00:19
   - -------------------------------------- 0.4/11.0 MB 637.8 kB/s eta 0:00:17
   - -------------------------------------- 0.5/11.0 MB 684.7 kB/s eta 0:00:16
   -- ------------------------------------- 0.6/11.0 MB 815.7 kB/s eta 0:00:13
   -- ------------------------------------- 0.8/11.0 MB 975.9 kB/s eta 0:00:11
```

```
-- ---------------------------------- 0.8/11.0 MB 975.9 kB/s eta 0:00:11
-- ---------------------------------- 0.8/11.0 MB 975.9 kB/s eta 0:00:11
--- ---------------------------------- 1.0/11.0 MB 1.0 MB/s eta 0:00:10
--- ---------------------------------- 1.1/11.0 MB 1.1 MB/s eta 0:00:09
---- ---------------------------------- 1.2/11.0 MB 1.2 MB/s eta 0:00:09
----- ---------------------------------- 1.4/11.0 MB 1.3 MB/s eta 0:00:08
----- ---------------------------------- 1.6/11.0 MB 1.4 MB/s eta 0:00:07
------ ---------------------------------- 1.8/11.0 MB 1.5 MB/s eta 0:00:07
------ ---------------------------------- 1.9/11.0 MB 1.6 MB/s eta 0:00:06
------- ---------------------------------- 2.1/11.0 MB 1.7 MB/s eta 0:00:06
-------- ---------------------------------- 2.4/11.0 MB 1.8 MB/s eta 0:00:05
--------- ---------------------------------- 2.5/11.0 MB 1.8 MB/s eta 0:00:05
---------- ---------------------------------- 2.8/11.0 MB 2.0 MB/s eta 0:00:05
---------- ---------------------------------- 2.9/11.0 MB 1.9 MB/s eta 0:00:05
----------- ---------------------------------- 3.1/11.0 MB 2.1 MB/s eta 0:00:04
------------ ---------------------------------- 3.4/11.0 MB 2.2 MB/s eta 0:00:04
------------- ---------------------------------- 3.7/11.0 MB 2.2 MB/s eta 0:00:04
-------------- ---------------------------------- 4.0/11.0 MB 2.4 MB/s eta 0:00:03
-------------- ---------------------------------- 4.3/11.0 MB 2.5 MB/s eta 0:00:03
--------------- ---------------------------------- 4.5/11.0 MB 2.5 MB/s eta 0:00:03
---------------- ---------------------------------- 4.9/11.0 MB 2.7 MB/s eta 0:00:03
------------------ ---------------------------------- 5.0/11.0 MB 2.6 MB/s eta 0:00:03
------------------- ---------------------------------- 5.4/11.0 MB 2.8 MB/s eta 0:00:03
-------------------- ---------------------------------- 5.5/11.0 MB 2.8 MB/s eta 0:00:02
--------------------- ---------------------------------- 5.9/11.0 MB 2.9 MB/s eta 0:00:02
--------------------- ---------------------------------- 5.9/11.0 MB 2.9 MB/s eta 0:00:02
---------------------- ---------------------------------- 6.1/11.0 MB 2.8 MB/s eta 0:00:02
---------------------- ---------------------------------- 6.1/11.0 MB 2.8 MB/s eta 0:00:02
----------------------- ---------------------------------- 6.5/11.0 MB 2.9 MB/s eta 0:00:02
----------------------- ---------------------------------- 6.6/11.0 MB 2.9 MB/s eta 0:00:02
----------------------- ---------------------------------- 6.7/11.0 MB 2.9 MB/s eta 0:00:02
------------------------ ---------------------------------- 7.2/11.0 MB 3.0 MB/s eta 0:00:02
------------------------ ---------------------------------- 7.3/11.0 MB 3.0 MB/s eta 0:00:02
------------------------- ---------------------------------- 7.8/11.0 MB 3.1 MB/s eta 0:00:02
-------------------------- ---------------------------------- 8.1/11.0 MB 3.1 MB/s eta 0:00:01
-------------------------- ---------------------------------- 8.4/11.0 MB 3.2 MB/s eta 0:00:01
--------------------------- ---------------------------------- 9.0/11.0 MB 3.3 MB/s eta 0:00:01
--------------------------- ---------------------------------- 9.3/11.0 MB 3.4 MB/s eta 0:00:01
---------------------------- ---------------------------------- 9.6/11.0 MB 3.5 MB/s eta 0:00:01
---------------------------- ---------------------------------- 9.9/11.0 MB 3.5 MB/s eta 0:00:01
---------------------------- ---------------------------------- 9.9/11.0 MB 3.5 MB/s eta 0:00:01
---------------------------- ---------------------------------- 10.1/11.0 MB 3.4 MB/s eta 0:00:01
---------------------------- ---------------------------------- 10.6/11.0 MB 4.4 MB/s eta 0:00:01
---------------------------- ----------------------------------  10.9/11.0 MB 4.5 MB/s eta 0:00:01
```

```
   ---------------------------------------- 11.0/11.0 MB 4.5 MB/s eta 0:00:01
   ---------------------------------------- 11.0/11.0 MB 4.5 MB/s eta 0:00:01
   ---------------------------------------- 11.0/11.0 MB 4.2 MB/s eta 0:00:00
Downloading threadpoolctl-3.5.0-py3-none-any.whl (18 kB)
Installing collected packages: threadpoolctl, scikit-learn
  Attempting uninstall: threadpoolctl
    Found existing installation: threadpoolctl 2.2.0
    Uninstalling threadpoolctl-2.2.0:
      Successfully uninstalled threadpoolctl-2.2.0
  Attempting uninstall: scikit-learn
    Found existing installation: scikit-learn 1.3.0
    Uninstalling scikit-learn-1.3.0:
      Successfully uninstalled scikit-learn-1.3.0
Successfully installed scikit-learn-1.5.0 threadpoolctl-3.5.0
```

In [3]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import re
import nltk
import string
from nltk.corpus import stopwords
from nltk.stem import LancasterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report
```

In [5]:
```
train_path = "train_data.txt"
train_data = pd.read_csv(train_path, sep=':::', names=['Title', 'Genre', 'Description'], engine='python')
train_data
```

Out[5]:

	Title	Genre	Description
1	Oscar et la dame rose (2009)	drama	Listening in to a conversation between his do...
2	Cupid (1997)	thriller	A brother and sister with a past incestuous r...
3	Young, Wild and Wonderful (1980)	adult	As the bus empties the students for their fie...
4	The Secret Sin (1915)	drama	To help their unemployed father make ends mee...
5	The Unrecovered (2007)	drama	The film's title refers not only to the un-re...
...	...	...	...
54210	"Bonino" (1953)	comedy	This short-lived NBC live sitcom centered on ...
54211	Dead Girls Don't Cry (????)	horror	The NEXT Generation of EXPLOITATION. The sist...
54212	Ronald Goedemondt: Ze bestaan echt (2008)	documentary	Ze bestaan echt, is a stand-up comedy about g...
54213	Make Your Own Bed (1944)	comedy	Walter and Vivian live in the country and hav...
54214	Nature's Fury: Storm of the Century (2006)	history	On Labor Day Weekend, 1935, the most intense ...

54214 rows × 3 columns

In [6]:
```
test_path = "test_data.txt"
test_data = pd.read_csv(test_path, sep=':::', names=['Id', 'Title', 'Description'], engine='python')
test_data
```

Out[6]:

	Id	Title	Description
0	1	Edgar's Lunch (1998)	L.R. Brane loves his life - his car, his apar...
1	2	La guerra de papá (1977)	Spain, March 1964: Quico is a very naughty ch...
2	3	Off the Beaten Track (2010)	One year in the life of Albin and his family ...
3	4	Meu Amigo Hindu (2015)	His father has died, he hasn't spoken with hi...
4	5	Er nu zhai (1955)	Before he was known internationally as a mart...
...	...	...	...
54195	54196	"Tales of Light & Dark" (2013)	Covering multiple genres, Tales of Light & Da...
54196	54197	Der letzte Mohikaner (1965)	As Alice and Cora Munro attempt to find their...
54197	54198	Oliver Twink (2007)	A movie 169 years in the making. Oliver Twist...
54198	54199	Slipstream (1973)	Popular, but mysterious rock DJ Mike Mallard...
54199	54200	Curitiba Zero Grau (2010)	Curitiba is a city in movement, with rhythms ...

54200 rows × 3 columns

In [7]:
```
train_data.describe()
```

Out[7]:

	Title	Genre	Description
count	54214	54214	54214
unique	54214	27	54086
top	Oscar et la dame rose (2009)	drama	Grammy - music award of the American academy ...
freq	1	13613	12

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

Out[8]:
```
Title          0
Genre          0
Description    0
dtype: int64
```

In [9]:
```
test_data.isnull().sum()
```

Out[9]:
```
Id             0
Title          0
Description    0
dtype: int64
```

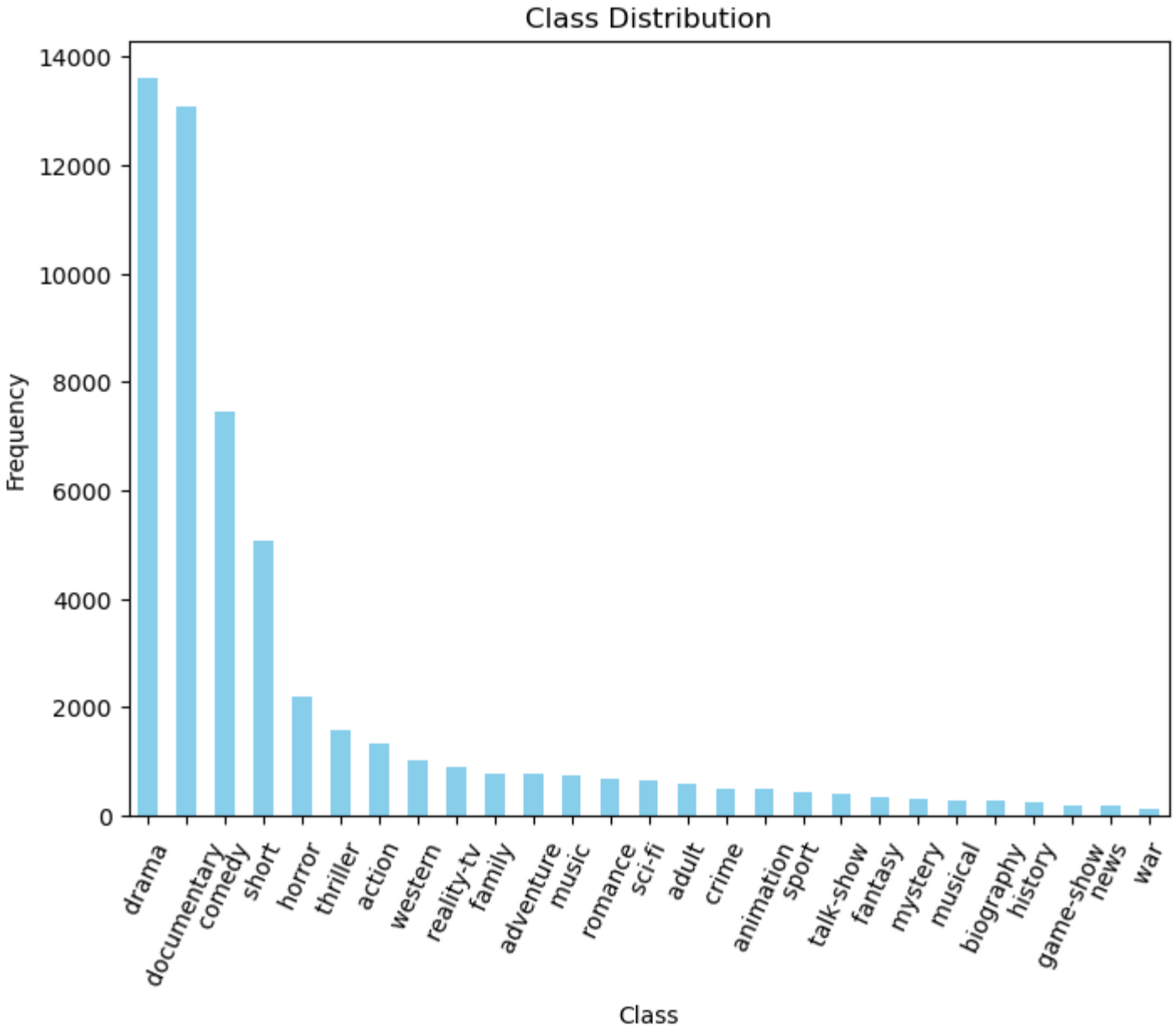
In [10]:
```
class_distribution = train_data['Genre'].value_counts()
print("Class Distribution:")
print(class_distribution)
```

```
Class Distribution:
Genre
 drama           13613
 documentary     13096
 comedy           7447
 short            5073
 horror           2204
 thriller         1591
 action           1315
 western          1032
 reality-tv        884
 family            784
 adventure         775
 music             731
 romance           672
 sci-fi            647
 adult             590
 crime             505
 animation         498
 sport             432
 talk-show         391
 fantasy           323
 mystery           319
 musical           277
 biography         265
 history           243
 game-show         194
 news              181
 war               132
Name: count, dtype: int64
```

In [11]:
```
imbalance_ratio = class_distribution.min() / class_distribution.max()
print("Imbalance Ratio:", imbalance_ratio)
```

```
Imbalance Ratio: 0.009696613531183427
```

In [12]:
```
plt.figure(figsize=(8, 6))
class_distribution.plot(kind='bar', color='skyblue')
plt.title('Class Distribution')
plt.xlabel('Class')
plt.ylabel('Frequency')
plt.xticks(rotation=65)
plt.show()
```


Class Distribution
Frequency
14000
12000
10000
8000
6000
4000
2000
0
drama
documentary
comedy
short
horror
thriller
action
western
reality-tv
family
adventure
music
romance
sci-fi
adult
crime
animation
sport
talk-show
fantasy
mystery
musical
biography
history
game-show
news
war
Class

In [13]:
```
tfidf_vectorizer = TfidfVectorizer(max_features=5000)
X_train_tfidf = tfidf_vectorizer.fit_transform(train_data['Description'])
y_train = train_data['Genre']


nb_classifier = MultinomialNB()
nb_classifier.fit(X_train_tfidf, y_train)


y_train_pred = nb_classifier.predict(X_train_tfidf)


print("Accuracy on training set:", accuracy_score(y_train, y_train_pred))
print("Classification Report on training set:\n", classification_report(y_train, y_train_pred))
```

```
Accuracy on training set: 0.5359132327443096
C:\Users\Dell\anaconda3\Lib\site-packages\sklearn\metrics\_classification.py:1517: UndefinedMetricWarning: Precision
is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control th
is behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
C:\Users\Dell\anaconda3\Lib\site-packages\sklearn\metrics\_classification.py:1517: UndefinedMetricWarning: Precision
is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control th
is behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

```
Classification Report on training set:
               precision    recall  f1-score   support

      action       0.70      0.09      0.16      1315
       adult       0.79      0.05      0.10       590
   adventure       0.76      0.05      0.10       775
   animation       0.00      0.00      0.00       498
   biography       0.00      0.00      0.00       265
      comedy       0.56      0.45      0.50      7447
       crime       0.00      0.00      0.00       505
 documentary       0.57      0.90      0.70     13096
       drama       0.47      0.84      0.60     13613
      family       1.00      0.00      0.01       784
     fantasy       0.00      0.00      0.00       323
   game-show       1.00      0.14      0.24       194
     history       0.00      0.00      0.00       243
      horror       0.78      0.36      0.50      2204
       music       0.90      0.16      0.27       731
     musical       0.00      0.00      0.00       277
     mystery       0.00      0.00      0.00       319
        news       0.00      0.00      0.00       181
  reality-tv       0.85      0.03      0.05       884
     romance       0.00      0.00      0.00       672
      sci-fi       0.85      0.04      0.09       647
       short       0.66      0.11      0.19      5073
       sport       0.80      0.11      0.19       432
   talk-show       1.00      0.01      0.02       391
    thriller       0.71      0.02      0.05      1591
         war       0.00      0.00      0.00       132
     western       0.97      0.59      0.73      1032

    accuracy                           0.54     54214
   macro avg       0.50      0.15      0.17     54214
weighted avg       0.57      0.54      0.46     54214

C:\Users\Dell\anaconda3\Lib\site-packages\sklearn\metrics\_classification.py:1517: UndefinedMetricWarning: Precision
is ill-defined and being set to 0.0 in labels with no predicted samples. Use `zero_division` parameter to control th
is behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is", len(result))
```

In [15]:
```
tfidf_vectorizer = TfidfVectorizer(max_features=5000)
X_test = tfidf_vectorizer.fit_transform(test_data['Description'])
```

In [16]:
```
X_test_predictions = nb_classifier.predict(X_test)
test_data['Predicted_Genre'] = X_test_predictions
```

In [17]:
```
test_data.to_csv('predicted_genres.csv', index=False)

print(test_data)
```

```
          Id                               Title  \
0          1             Edgar's Lunch (1998)
1          2         La guerra de papá (1977)
2          3      Off the Beaten Track (2010)
3          4           Meu Amigo Hindu (2015)
4          5                Er nu zhai (1955)
...      ...                               ...
54195  54196   "Tales of Light & Dark" (2013)
54196  54197      Der letzte Mohikaner (1965)
54197  54198              Oliver Twink (2007)
54198  54199                Slipstream (1973)
54199  54200        Curitiba Zero Grau (2010)

                                        Description Predicted_Genre
0       L.R. Brane loves his life - his car, his apar...          drama
1       Spain, March 1964: Quico is a very naughty ch...          drama
2       One year in the life of Albin and his family ...    documentary
3       His father has died, he hasn't spoken with hi...    documentary
4       Before he was known internationally as a mart...    documentary
...                                                   ...            ...
54195   Covering multiple genres, Tales of Light & Da...          drama
54196   As Alice and Cora Munro attempt to find their...          drama
54197   A movie 169 years in the making. Oliver Twist...    documentary
54198   Popular, but mysterious rock D.J Mike Mallard...          drama
54199   Curitiba is a city in movement, with rhythms ...          short

[54200 rows x 4 columns]
```

In [18]:
```
import pickle
with open('tfidf_vectorizer.pkl', 'wb') as file:
    pickle.dump(tfidf_vectorizer, file)
with open('nb_classifier.pkl', 'wb') as file:
    pickle.dump(nb_classifier, file)

print("Models pickled successfully.")
```

```
Models pickled successfully.
```

In [ ]: