Movie Rating Analysis (Beginner Friendly)

This notebook performs sentiment analysis on movie reviews using a simple machine learning model.

In [1]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix, classification_report
import warnings
warnings.filterwarnings('ignore')
```

In [2]:

```python
# Load the dataset
df = pd.read_csv('movie.csv')
df.head()
```

Out[2]:

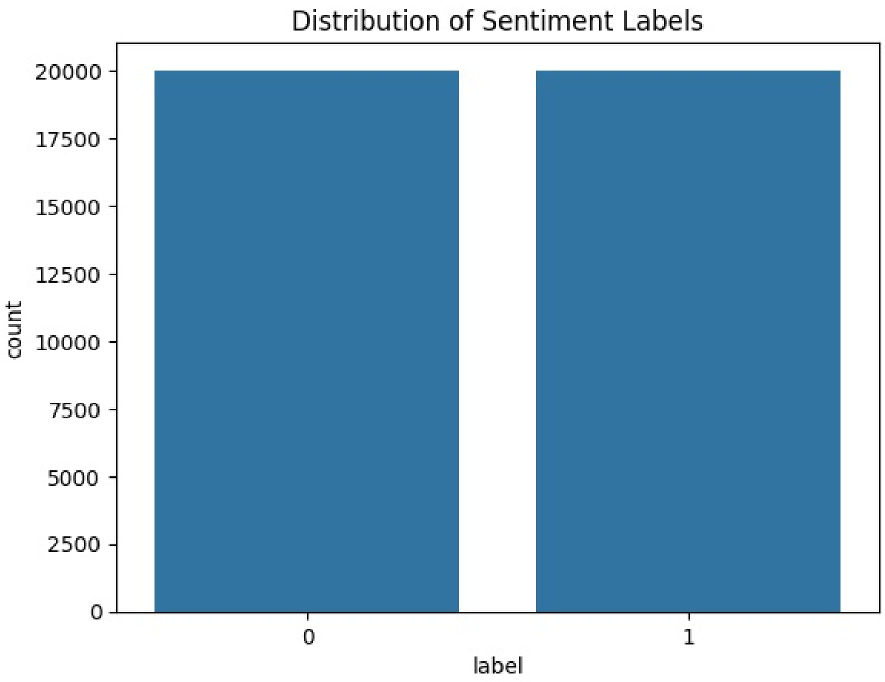
	text	label
0	I grew up (b. 1965) watching and loving the Th...	0
1	When I put this movie in my DVD player, and sa...	0
2	Why do people who do not know what a particula...	0
3	Even though I have great interest in Biblical ...	0
4	Im a die hard Dads Army fan and nothing will e...	1

In [3]:

```python
# Check for null values and basic statistics
print(df.info())
print('\nLabel distribution:')
print(df['label'].value_counts())
sns.countplot(x='label', data=df)
plt.title('Distribution of Sentiment Labels')
plt.show()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40000 entries, 0 to 39999
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype 
---  ------  --------------  ----- 
 0   text    40000 non-null  object
 1   label   40000 non-null  int64 
dtypes: int64(1), object(1)
memory usage: 625.1+ KB
None

Label distribution:
label
0    20019
1    19981
Name: count, dtype: int64
```



In [4]:

```python
import re
def preprocess_text(text):
    text = text.lower()
    text = re.sub(r'[^a-zA-Z\s]', '', text)
    return text

df['clean_text'] = df['text'].apply(preprocess_text)
df[['text', 'clean_text']].head()
```

Out[4]:

	text	clean_text
0	I grew up (b. 1965) watching and loving the Th...	i grew up b watching and loving the thunderbi...
1	When I put this movie in my DVD player, and sa...	when i put this movie in my dvd player and sat...
2	Why do people who do not know what a particula...	why do people who do not know what a particula...
3	Even though I have great interest in Biblical ...	even though i have great interest in biblical ...
4	Im a die hard Dads Army fan and nothing will e...	im a die hard dads army fan and nothing will e...

In [5]:

```python
# Split the data
X = df['clean_text']
y = df['label']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [6]:

```python
# Convert text data to numerical vectors
vectorizer = CountVectorizer()
X_train_vec = vectorizer.fit_transform(X_train)
X_test_vec = vectorizer.transform(X_test)
```

In [7]:

```
# Train a logistic regression model
model = LogisticRegression()
model.fit(X_train_vec, y_train)
y_pred = model.predict(X_test_vec)
```

In [8]:

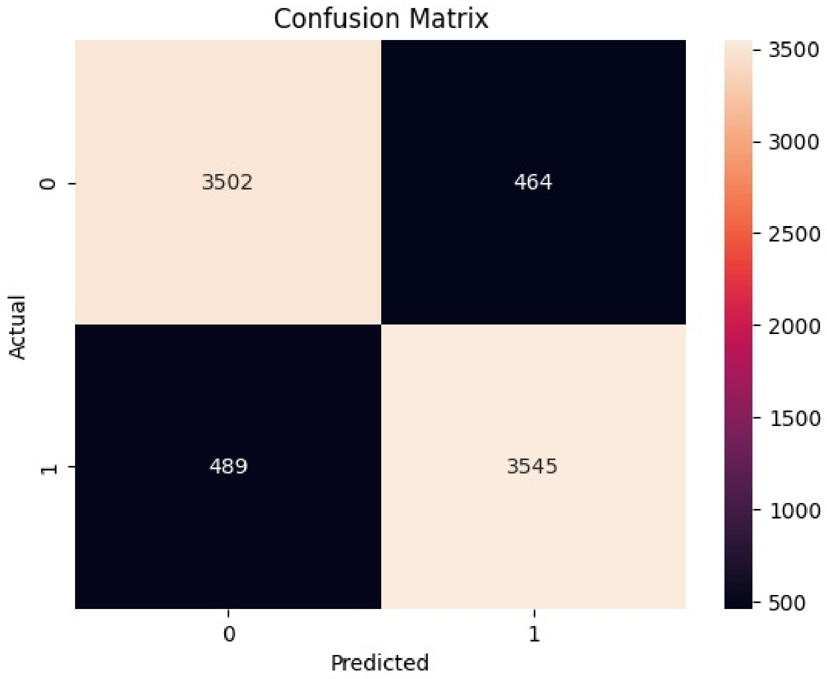
```
# Evaluate the model
print('Accuracy:', accuracy_score(y_test, y_pred))
print('\nClassification Report:')
print(classification_report(y_test, y_pred))
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d')
plt.title('Confusion Matrix')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```

```
Accuracy: 0.880875

Classification Report:
              precision    recall  f1-score   support

           0       0.88      0.88      0.88      3966
           1       0.88      0.88      0.88      4034

    accuracy                           0.88      8000
   macro avg       0.88      0.88      0.88      8000
weighted avg       0.88      0.88      0.88      8000
```



Loading [MathJax]/jax/output/CommonHTML/fonts/TeX/fontdata.js