

```
# IMPORTANT: RUN THIS CELL IN ORDER TO IMPORT YOUR KAGGLE DATA SOURCES,
# THEN FEEL FREE TO DELETE THIS CELL.
# NOTE: THIS NOTEBOOK ENVIRONMENT DIFFERS FROM KAGGLE'S PYTHON
# ENVIRONMENT SO THERE MAY BE MISSING LIBRARIES USED BY YOUR
# NOTEBOOK.
import kagglehub
lakshmi25npathi_imdb_dataset_of_50k_movie_reviews_path = kagglehub.dataset_download('lakshmi25npathi/imdb-dataset-of-50k-mov

print('Data source import complete.')
```

```
!pip install nltk
```

```
Requirement already satisfied: nltk in /usr/local/lib/python3.12/dist-packages (3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.12/dist-packages (from nltk) (8.3.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.12/dist-packages (from nltk) (1.5.3)
Requirement already satisfied: regex<=2021.8.3 in /usr/local/lib/python3.12/dist-packages (from nltk) (2025.11.3)
Requirement already satisfied: tqdm in /usr/local/lib/python3.12/dist-packages (from nltk) (4.67.3)
```

```
import pandas as pd
import numpy as np
import re
import nltk

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, classification_report

stop_words = set([
'i','me','my','myself','we','our','ours','ourselves','you','your','yours',
'yourself','yourselves','he','him','his','himself','she','her','hers',
'herself','it','its','itself','they','them','their','theirs','themselves',
'what','which','who','whom','this','that','these','those','am','is','are',
'was','were','be','been','being','have','has','had','having','do','does',
'did','doing','a','an','the','and','but','if','or','because','as','until',
'while','of','at','by','for','with','about','against','between','into',
'through','during','before','after','above','below','to','from','up',
'down','in','out','on','off','over','under','again','further','then',
'once','here','there','when','where','why','how','all','any','both',
'each','few','more','most','other','some','such','no','nor','not',
'only','own','same','so','than','too','very','can','will','just','don'
])
```

```
import os

for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))
```

```
/kaggle/input/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/IMDB Dataset.csv
```

```
df = pd.read_csv('/kaggle/input/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/IMDB Dataset.csv')
df.head()
```

```
def clean_text(text):
    text = re.sub(r'<.*?>', '', text)
    text = re.sub(r'^a-zA-Z', ' ', text)
    text = text.lower()
    text = text.split()
    text = [word for word in text if word not in stopwords.words('english')]
    return ' '.join(text)

df['review'] = df['review'].apply(clean_text)

df.head()
```

```
df['sentiment'] = df['sentiment'].map({'positive':1, 'negative':0})

df['sentiment'].value_counts()
```

```
sentiment
1    25000
```

```
0    25000
Name: count, dtype: int64
```

```
X_train, X_test, y_train, y_test = train_test_split(
    df['review'], df['sentiment'], test_size=0.2, random_state=42
)
```

```
vectorizer = TfidfVectorizer(max_features=5000)

X_train_vec = vectorizer.fit_transform(X_train)
X_test_vec = vectorizer.transform(X_test)
```

```
model = LogisticRegression()

model.fit(X_train_vec, y_train)
```

```
y_pred = model.predict(X_test_vec)

print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

Accuracy: 0.8915

Classification Report:

	precision	recall	f1-score	support
0	0.90	0.88	0.89	4961
1	0.88	0.91	0.89	5039
accuracy			0.89	10000
macro avg	0.89	0.89	0.89	10000
weighted avg	0.89	0.89	0.89	10000

```
from sklearn.naive_bayes import MultinomialNB

nb_model = MultinomialNB()
nb_model.fit(X_train_vec, y_train)

y_pred_nb = nb_model.predict(X_test_vec)

print("Naive Bayes Accuracy:", accuracy_score(y_test, y_pred_nb))
```

Naive Bayes Accuracy: 0.8553