OLUYEMI OLUWOYE ASSIGNMENT 13_ML

Double-click (or enter) to edit

```
import pandas as pd

df = pd.read_csv('IMDB Dataset.csv')
df
```

	review	sentiment
0	One of the other reviewers has mentioned that ...	positive
1	A wonderful little production. The...	positive
2	I thought this was a wonderful way to spend ti...	positive
3	Basically there's a family where a little boy ...	negative
4	Petter Mattei's "Love in the Time of Money" is...	positive
...	...	...
49995	I thought this movie did a down right good job...	positive
49996	Bad plot, bad dialogue, bad acting, idiotic di...	negative
49997	I am a Catholic taught in parochial elementary...	negative
49998	I'm going to have to disagree with the previou...	negative
49999	No one expects the Star Trek movies to be high...	negative

50000 rows × 2 columns

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

```
import nltk
from nltk.tokenize import word_tokenize

nltk.download('punkt_tab')

# Tokenize the reviews
df['tokens'] = df['review'].apply(word_tokenize)
df[['review', 'tokens']].head()
```

```
[nltk_data] Downloading package punkt_tab to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt_tab.zip.
```

	review	tokens
0	One of the other reviewers has mentioned that ...	[One, of, the, other, reviewers, has, mentione...
1	A wonderful little production. The...	[A, wonderful, little, production, ., <, br, /...
2	I thought this was a wonderful way to spend ti...	[I, thought, this, was, a, wonderful, way, to,...
3	Basically there's a family where a little boy ...	[Basically, there, 's, a, family, where, a, li...
4	Petter Mattei's "Love in the Time of Money" is...	[Petter, Mattei, 's, ``, Love, in, the, Time, ...

```
from nltk.corpus import stopwords
nltk.download('stopwords')

stop_words = set(stopwords.words('english'))
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
#  create 'tokens_no_stop'

df['tokens_no_stop'] = df['tokens'].apply(lambda x: [word for word in x if word.lower() not in stop_words])
```

```
df
```

	review	sentiment	tokens	tokens_no_stop
0	One of the other reviewers has mentioned that ...	positive	[One, of, the, other, reviewers, has, mentione...	[One, reviewers, mentioned, watching, 1, Oz, e...
1	A wonderful little production. The...	positive	[A, wonderful, little, production, ., <, br, /...	[wonderful, little, production, ., <, br, /, >...
2	I thought this was a wonderful way to spend ti...	positive	[I, thought, this, was, a, wonderful, way, to,...	[thought, wonderful, way, spend, time, hot, su...
3	Basically there's a family where a little boy ...	negative	[Basically, there, 's, a, family, where, a, li...	[Basically, 's, family, little, boy, (, Jake, ...
4	Petter Mattei's "Love in the Time of Money" is...	positive	[Petter, Mattei, 's, ``, Love, in, the, Time, ...	[Petter, Mattei, 's, ``, Love, Time, Money, "...
...	...	...	...	...
49995	I thought this movie did a down right good job...	positive	[I, thought, this, movie, did, a, down, right,...	[thought, movie, right, good, job, ., n't, cre...
49996	Bad plot, bad dialogue, bad acting, idiotic di...	negative	[Bad, plot, ,, bad, dialogue, ,, bad, acting, ...	[Bad, plot, ,, bad, dialogue, ,, bad, acting, ...
49997	I am a Catholic taught in parochial	negative	[I, am, a, Catholic, taught, in, parochial, el	[Catholic, taught, parochial, elementary,

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

```
from nltk.stem import PorterStemmer

stemmer = PorterStemmer()
df['stemmed'] = df['tokens_no_stop'].apply(lambda x: [stemmer.stem(word) for word in x])
df[['review', 'stemmed']].head()
```

	review	stemmed
0	One of the other reviewers has mentioned that ...	[one, review, mention, watch, 1, oz, episod, '...
1	A wonderful little production. The...	[wonder, littl, product, ., <, br, /, >, <, br...
2	I thought this was a wonderful way to spend ti...	[thought, wonder, way, spend, time, hot, summe...
3	Basically there's a family where a little boy ...	[basic, 's, famili, littl, boy, (, jake,), th...
4	Petter Mattei's "Love in the Time of Money" is...	[petter, mattei, 's, ``, love, time, money, "...

Part 2

```
import re

from nltk.stem import WordNetLemmatizer
from bs4 import BeautifulSoup


nltk.download('wordnet')
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data] Downloading package omw-1.4 to /root/nltk_data...
True
```

```
lemmatizer = WordNetLemmatizer()

def preprocess(text):
    text = BeautifulSoup(text, "html.parser").get_text()
    text = re.sub(r'[^a-zA-Z]', ' ', text.lower())
    tokens = word_tokenize(text)
    tokens = [lemmatizer.lemmatize(word) for word in tokens if word not in stop_words]
    return ' '.join(tokens)

df['clean_review'] = df['review'].apply(preprocess)


df
```

	review	sentiment	tokens	tokens_no_stop	stemmed	clean_review
0	One of the other reviewers has mentioned that ...	positive	[One, of, the, other, reviewers, has, mentione...	[One, reviewers, mentioned, watching, 1, Oz, e...	[one, review, mention, watch, 1, oz, episod, '...	one reviewer mentioned watching oz episode hoo...
1	A wonderful little production. The...	positive	[A, wonderful, little, production, ., <, br, /...	[wonderful, little, production, ., <, br, /, >...	[wonder, littl, product, ., <, br, /, >, <, br...	wonderful little production filming technique ...
2	I thought this was a wonderful way to spend ti...	positive	[I, thought, this, was, a, wonderful, way, to,...	[thought, wonderful, way, spend, time, hot, su...	[thought, wonder, way, spend, time, hot, summe...	thought wonderful way spend time hot summer we...
3	Basically there's a family where a little boy ...	negative	[Basically, there, 's, a, family, where, a, li...	[Basically, 's, family, little, boy, (, Jake, ...	[basic, 's, famili, littl, boy, (, jake,), th...	basically family little boy jake think zombie ...
4	Petter Mattei's "Love in the Time of Money" is...	positive	[Petter, Mattei, 's, ``, Love, in, the, Time, ...	[Petter, Mattei, 's, ``, Love, Time, Money, "...	[petter, mattei, 's, ``, love, time, money, "...	petter mattei love time money visually stunnin...
...	...	...	...	...	...	...
49995	I thought this movie did a down right good job...	positive	[I, thought, this, movie, did, a, down, right,...	[thought, movie, right, good, job, ., n't, cre...	[thought, movi, right, good, job, ., n't, crea...	thought movie right good job creative original...
49996	Bad plot, bad dialogue, bad acting, idiotic di...	negative	[Bad, plot, ,, bad, dialogue, ,, bad, acting, ...	[Bad, plot, ,, bad, dialogue, ,, bad, acting, ...	[bad, plot, ,, bad, dialogu, ,, bad, act, ,, i...	bad plot bad dialogue bad acting idiotic direc...
49997	I am a Catholic taught in parochial elementary...	negative	[I, am, a, Catholic, taught, in, parochial, el...	[Catholic, taught, parochial, elementary, scho...	[cathol, taught, parochi, elementari, school, ...	catholic taught parochial elementary school nu...
49998	I'm going to have to disagree with the previou...	negative	[I, 'm, going, to, have, to, disagree, with, t...	['m, going, disagree, previous, comment, side,...	['m, go, disagre, previou, comment, side, malt...	going disagree previous comment side maltin on...
	No one expects the Star		[No, one, expects, the	[one, expects, Star, Trek	[one, expect, star, trek	one expects star trek movie

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

```
# Label Encoding:

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

```
df
```

	review	sentiment	tokens	tokens_no_stop	stemmed	clean_review	label
0	One of the other reviewers has mentioned that ...	positive	[One, of, the, other, reviewers, has, mentione...	[One, reviewers, mentioned, watching, 1, Oz, e...	[one, review, mention, watch, 1, oz, episod, '...	one reviewer mentioned watching oz episode hoo...	1
1	A wonderful little production. The...	positive	[A, wonderful, little, production, ., <, br, /...	[wonderful, little, production, ., <, br, /, >...	[wonder, littl, product, ., <, br, /, >, <, br...	wonderful little production filming technique ...	1
2	I thought this was a wonderful way to spend ti...	positive	[I, thought, this, was, a, wonderful, way, to,...	[thought, wonderful, way, spend, time, hot, su...	[thought, wonder, way, spend, time, hot, summe...	thought wonderful way spend time hot summer we...	1
3	Basically there's a family where a little boy ...	negative	[Basically, there, 's, a, family, where, a, li...	[Basically, 's, family, little, boy, (, Jake, ...	[basic, 's, famili, littl, boy, (, jake,), th...	basically family little boy jake think zombie ...	0
4	Petter Mattei's "Love in the Time of Money" is...	positive	[Petter, Mattei, 's, ``, Love, in, the, Time, ...	[Petter, Mattei, 's, ``, Love, Time, Money, "...	[petter, mattei, 's, ``, love, time, money, "...	petter mattei love time money visually stunnin...	1
...	...	...	...	...	...	...	...
49995	I thought this movie did a down right good job...	positive	[I, thought, this, movie, did, a, down, right,...	[thought, movie, right, good, job, ., n't, cre...	[thought, movi, right, good, job, ., n't, crea...	thought movie right good job creative original...	1
49996	Bad plot, bad dialogue, bad acting, idiotic di...	negative	[Bad, plot, ,, bad, dialogue, ,, bad, acting, ...	[Bad, plot, ,, bad, dialogue, ,, bad, acting, ...	[bad, plot, ,, bad, dialogu, ,, bad, act, ,, i...	bad plot bad dialogue bad acting idiotic direc...	0
49997	I am a Catholic taught in parochial elementary...	negative	[I, am, a, Catholic, taught, in, parochial, el...	[Catholic, taught, parochial, elementary, scho...	[cathol, taught, parochi, elementari, school, ...	catholic taught parochial elementary school nu...	0
49998	I'm going to have to disagree with the previou...	negative	[I, 'm, going, to, have, to, disagree, with, t...	['m, going, disagree, previous, comment, side,...	['m, go, disagre, previou, comment, side, malt...	going disagree previous comment side maltin on...	0
	No one expects the		[No, one, expects		[one, expect, star	one expects star trek	

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

```
# spliting and vectorizing the data

from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer

X = df['clean_review']
y = df['label']

X_train, X_test, y_train, y_test = train_test_split(X, y, 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 training

from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

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

y_pred = model.predict(X_test_vec)
```

```
# evaluating

print("Accuracy:", accuracy_score(y_test, y_pred))
print(classification_report(y_test, y_pred))
```

```
Accuracy: 0.8545
              precision    recall  f1-score   support
```

```
           0       0.86      0.85      0.85      4961
           1       0.85      0.86      0.86      5039

    accuracy                           0.85     10000
   macro avg       0.85      0.85      0.85     10000
weighted avg       0.85      0.85      0.85     10000
```

** Interpretation of Results**

Accuracy: 85.45% The model correctly predicted sentiment for 85% of the 10,000 test reviews.

Class 0 (Negative Reviews)

- Precision: 0.86 of all reviews predicted as negative, 86% were actually negative.
- Recall: 0.85 of all true negative reviews, 85% were correctly identified.
- F1-Score: 0.85 Balanced performance between precision and recall.

Class 1 (Positive Reviews)

- Precision: 0.85, Recall: 0.86, F1-Score: 0.86 → Similar strong performance on positive reviews.

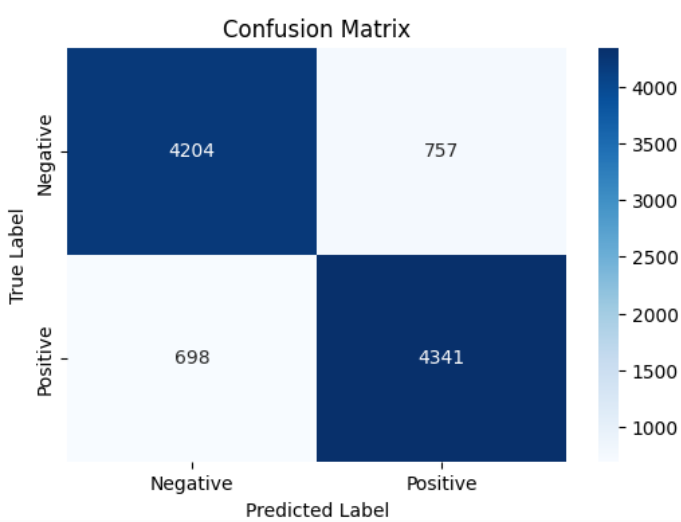
Macro avg / Weighted avg: 0.85 Shows overall balanced performance across both sentiment classes.

Conclusion: The model is performing well, with balanced and reliable prediction for both positive and negative sentiments

```
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix

# Generate confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Plot
plt.figure(figsize=(6, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues', xticklabels=['Negative', 'Positive'], yticklabels=['Negative', 'Positive'])
plt.title('Confusion Matrix')
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
plt.show()
```



Interpretation of the Confusion Matrix

True Negatives (Top-left: 4204) The model correctly predicted 4,204 negative reviews as negative.

False Positives (Top-right: 757) 757 negative reviews were incorrectly predicted as positive.

False Negatives (Bottom-left: 698) 698 positive reviews were incorrectly predicted as negative.

True Positives (Bottom-right: 4341) The model correctly predicted 4,341 positive reviews as positive.