Course Name: Deep Learning

Lab Title: NLP Techniques for Text Classification

Student Name : Sakshi Dube

Student ID:202201040155

Date of Submission: 1-4-25

Group Members: Manvi Pawar,Kanishka Garud,Shravani Sakore

Objective The objective of this assignment is to implement NLP preprocessing techniques and build a text classification model using machine learning techniques.

Dataset : https://www.kaggle.com/datasets/saurabhshahane/ecommerce-text-classification

** Importing Necessary Libraries**

```
import pandas as pd
import numpy as np
import nltk
import re
import string
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem import PorterStemmer, WordNetLemmatizer
from sklearn.feature_extraction.text import CountVectorizer,
TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, confusion_matrix,
classification_report
```

** Downloading Required NLTK Data**

```
nltk.download('punkt')
nltk.download('stopwords')
nltk.download('wordnet')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
True
```

Task 1. NLP Preprocessing

Loading the Dataset

```
dataset_path = "/content/ecommerceDataset.csv"  # Update this with the correct dataset path
df = pd.read_csv(dataset_path, on_bad_lines='skip', quoting=3)

# Display first few rows
print(df.head())

# Check dataset info
print(df.info())

# Check for missing values
print(df.isnull().sum())
```

```
   Household  \
0  Household
1  Household
2  Household
3  Household
4  Household

  "Paper Plane Design Framed Wall Hanging Motivational Office Decor Art Prints (8.7 X 8.7 inch) - Set of 4 Painting made up in synthetic frame with uv textured print which gives multi effects and attracts towards it. This is an special series of paintings which makes your wall very beautiful and gives a royal touch. This painting is ready to hang  \
0                "SAF 'Floral' Framed Painting (Wood

1  "SAF 'UV Textured Modern Art Print Framed' Pai...

2  "Incredible Gifts India Wooden Happy Birthday ...

3  "Paper Plane Design Starry Night Vangoh Wall A...

4  "SAF 'Ganesh Modern Art Print' Painting (Synth...


   you would be proud to possess this unique painting that is a niche apart. We use only the most modern and efficient printing technology on our prints  \
0                                  30 inch x 10 inch

1                              35 cm x 50 cm x 3 cm

2   Which Is Quite Solid With Light Particle Patt...

3   with only the best and original inks and prec...

4                              35 cm x 50 cm x 2 cm
```

```
          with only the and inks and precision epson  \
0                  Special Effect UV Print Textured
1   Set of 3) Color:Multicolor                    ...
2   Some Can Be Used As Table Top And The Big Siz...
3   to achieve brilliant and true colours. Due to...
4   Set of 3) Color:Multicolor   Overview a beaut...

   roland and hp printers. This innovative hd printing technique
results in durable and spectacular looking prints of the highest that
last a lifetime. We print solely with top-notch 100% inks  \
0   SAO297) Painting made up in synthetic frame w...

1   the end product will be a picture that can sp...

2                                                 NaN

3   our Canvas prints retain their beautiful colo...

4   the end product will be a picture that can sp...


   to achieve brilliant and true colours. Due to their high level of
uv resistance  \
0                                                 NaN

1   it does not include glass along with the fram...

2                                                 NaN

3   to ensure that the colours of your original i...

4   it does not include glass along with the fram...


   our prints retain their beautiful colours for many years. Add
colour and style to your living space with this digitally printed
painting. Some are for pleasure and some for eternal bliss.so bring
home this elegant print that is lushed with rich colors that makes it
nothing but sheer elegance to be to your friends and family.it would
be treasured forever by whoever your lucky recipient is. Liven up your
place with these intriguing paintings that are high definition hd
graphic digital prints for home  \
0                                                 NaN

1                and shopping is just a click away!"

2                                                 NaN

3   with brilliant tones. Add colour and style to...
```

```
4                 and shopping is just a click away!"


                                  office or any room."
0                                                   NaN
1                                                   NaN
2                                                   NaN
3   office or any room. A perfect size of 36 inch...
4                                                   NaN
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 117294 entries, 0 to 117293
Data columns (total 8 columns):
 #   Column
Non-Null Count   Dtype
---  ------
--------------   -----
 0   Household
117294 non-null  object
 1   "Paper Plane Design Framed Wall Hanging Motivational Office Decor
Art Prints (8.7 X 8.7 inch) - Set of 4 Painting made up in synthetic
frame with uv textured print which gives multi effects and attracts
towards it. This is an special series of paintings which makes your
wall very beautiful and gives a royal touch. This painting is ready to
hang
117266 non-null  object
 2    you would be proud to possess this unique painting that is a
niche apart. We use only the most modern and efficient printing
technology on our prints
75648 non-null   object
 3    with only the and inks and precision epson
55530 non-null   object
 4    roland and hp printers. This innovative hd printing technique
results in durable and spectacular looking prints of the highest that
last a lifetime. We print solely with top-notch 100% inks
40469 non-null   object
 5    to achieve brilliant and true colours. Due to their high level
of uv resistance
28048 non-null   object
 6    our prints retain their beautiful colours for many years. Add
colour and style to your living space with this digitally printed
painting. Some are for pleasure and some for eternal bliss.so bring
home this elegant print that is lushed with rich colors that makes it
nothing but sheer elegance to be to your friends and family.it would
be treasured forever by whoever your lucky recipient is. Liven up your
place with these intriguing paintings that are high definition hd
graphic digital prints for home  16427 non-null   object
 7    office or any room."
7067 non-null    object
dtypes: object(8)
```

```
memory usage: 7.2+ MB
None
Household
0
"Paper Plane Design Framed Wall Hanging Motivational Office Decor Art Prints (8.7 X 8.7 inch) - Set of 4 Painting made up in synthetic frame with uv textured print which gives multi effects and attracts towards it. This is an special series of paintings which makes your wall very beautiful and gives a royal touch. This painting is ready to hang
28
 you would be proud to possess this unique painting that is a niche apart. We use only the most modern and efficient printing technology on our prints
41646
 with only the and inks and precision epson
61764
 roland and hp printers. This innovative hd printing technique results in durable and spectacular looking prints of the highest that last a lifetime. We print solely with top-notch 100% inks
76825
 to achieve brilliant and true colours. Due to their high level of uv resistance
89246
 our prints retain their beautiful colours for many years. Add colour and style to your living space with this digitally printed painting. Some are for pleasure and some for eternal bliss.so bring home this elegant print that is lushed with rich colors that makes it nothing but sheer elegance to be to your friends and family.it would be treasured forever by whoever your lucky recipient is. Liven up your place with these intriguing paintings that are high definition hd graphic digital prints for home    100867
 office or any room."
110227
dtype: int64
```

Handling Missing Values & Renaming Columns (If Needed)

```
print(df.columns)  # See the current column names
print(df.shape)    # Check the number of rows and columns

Index(['Household',
       '"Paper Plane Design Framed Wall Hanging Motivational Office Decor Art Prints (8.7 X 8.7 inch) - Set of 4 Painting made up in synthetic frame with uv textured print which gives multi effects and attracts towards it. This is an special series of paintings which makes your wall very beautiful and gives a royal touch. This painting is ready to hang',
       ' you would be proud to possess this unique painting that is a niche apart. We use only the most modern and efficient printing
```

```
technology on our prints',
       ' with only the and inks and precision epson',
       ' roland and hp printers. This innovative hd printing technique
results in durable and spectacular looking prints of the highest that
last a lifetime. We print solely with top-notch 100% inks',
       ' to achieve brilliant and true colours. Due to their high
level of uv resistance',
       ' our prints retain their beautiful colours for many years. Add
colour and style to your living space with this digitally printed
painting. Some are for pleasure and some for eternal bliss.so bring
home this elegant print that is lushed with rich colors that makes it
nothing but sheer elegance to be to your friends and family.it would
be treasured forever by whoever your lucky recipient is. Liven up your
place with these intriguing paintings that are high definition hd
graphic digital prints for home',
       ' office or any room."'],
      dtype='object')
(117294, 8)
```

```
df = df.iloc[:, [0, -1]]  # Keeping only first and last columns
(adjust as needed)
df.columns = ['Category', 'Text']

df.fillna("", inplace=True)
```

```
<ipython-input-13-2622963f573f>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
returning-a-view-versus-a-copy
  df.fillna("", inplace=True)
```

```
# Check dataset structure
print(df.shape)  # Number of rows & columns
print(df.columns)  # See current column names

# Keep only necessary columns
df = df.iloc[:, [0, -1]]  # Adjust column selection based on dataset
df.columns = ['Category', 'Text']

# Fill missing values
df.fillna("", inplace=True)

print(df.head())  # Verify everything is correct
```

```
(117294, 2)
Index(['Category', 'Text'], dtype='object')
    Category                                                Text
0  Household
1  Household
```

```
2  Household
3  Household   office or any room. A perfect size of 36 inch...
4  Household
```

** Preprocessing Function **

```
from nltk.stem import PorterStemmer, SnowballStemmer

stemmer = PorterStemmer()  # or SnowballStemmer("english")

def preprocess_text(text):
    text = text.lower()
    text = re.sub(f"[{string.punctuation}]", "", text)
    tokens = word_tokenize(text)
    tokens = [word for word in tokens if word not in
stopwords.words('english')]
    # Apply stemming
    tokens = [stemmer.stem(word) for word in tokens]
    return " ".join(tokens)
```

** Applying Preprocessing**

```
df['Processed_Text'] = df['Text'].apply(preprocess_text)  # Apply
function to 'Text' column

print(df.head())  # Check processed text
```

```
    Category                                                Text  \
0  Household
1  Household
2  Household
3  Household   office or any room. A perfect size of 36 inch...
4  Household

                                      Processed_Text
0
1
2
3  offic room perfect size 36 inch x 48 inch suit...
4
```

Task 2.Vectorization Techniques

** Text Vectorization (TF-IDF)**

```
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(df['Processed_Text'])
y = df['Category']  # Target variable
```

```
print("Shape of feature matrix:", X.shape)
```

```
Shape of feature matrix: (117294, 5662)
```

CountVectorizer

```
from sklearn.feature_extraction.text import CountVectorizer

count_vectorizer = CountVectorizer()
X_count = count_vectorizer.fit_transform(df['Processed_Text'])
```

Task 3.Data Splitting

Splitting Dataset into Training & Testing Sets

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)

print("Training data size:", X_train.shape)
print("Testing data size:", X_test.shape)
```

```
Training data size: (93835, 5662)
Testing data size: (23459, 5662)
```

Task 4.Model Building

** Training Naïve Bayes Model**

```
model = MultinomialNB()
model.fit(X_train, y_train)

print("Model training complete.")
```

```
Model training complete.
```

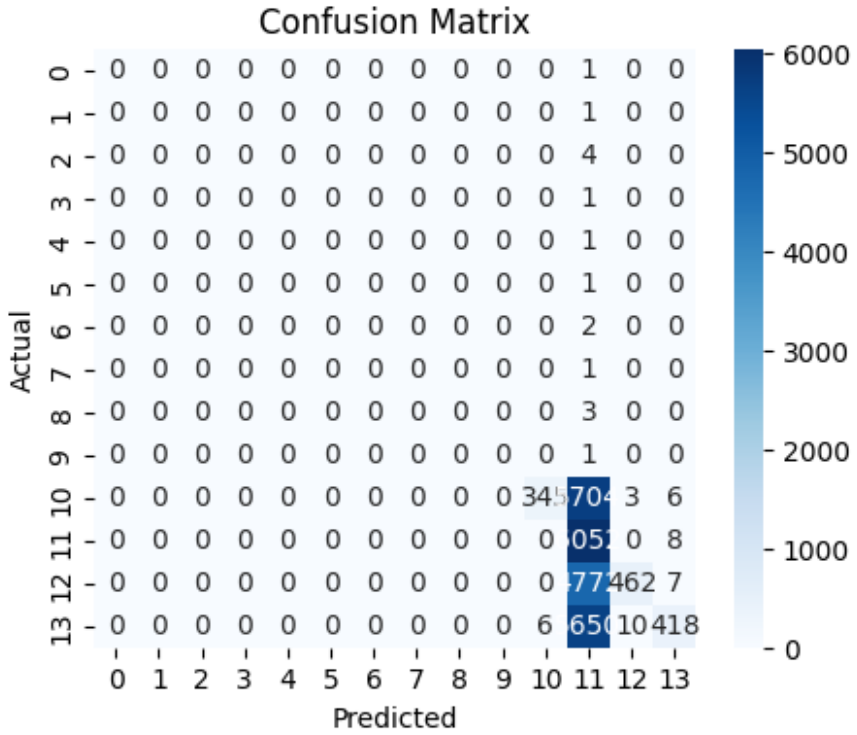
```
from sklearn.model_selection import cross_val_score
cv_scores = cross_val_score(model, X_train, y_train, cv=5)
print("Cross-validation scores:", cv_scores)
print("Mean CV Accuracy:", np.mean(cv_scores))
```

```
/usr/local/lib/python3.11/dist-packages/sklearn/model_selection/
_split.py:805: UserWarning: The least populated class in y has only 1
members, which is less than n_splits=5.
  warnings.warn(

Cross-validation scores: [0.30702829 0.30788085 0.30468375 0.30612245
0.30718815]
Mean CV Accuracy: 0.3065807001651836
```

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

conf_mat = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(5, 4))
sns.heatmap(conf_mat, annot=True, cmap="Blues", fmt='g')
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix")
plt.show()
```



** Making Predictions & Evaluating Model**

```
y_pred = model.predict(X_test)

accuracy = accuracy_score(y_test, y_pred)
print("Accuracy:", accuracy)
print("Classification Report:\n", classification_report(y_test,
y_pred))
print("Confusion Matrix:\n", confusion_matrix(y_test, y_pred))
```

```
Accuracy: 0.3102007758216463

/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_classification.py:1565: UndefinedMetricWarning: Precision is ill-
defined and being set to 0.0 in labels with no predicted samples. Use
```

```
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_classificatio
n.py:1565: UndefinedMetricWarning: Precision is ill-defined and being
set to 0.0 in labels with no predicted samples. Use `zero_division`
parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))

Classification Report:

precision    recall  f1-score   support


What has been the best part of your startup experience? In this
dynamic start-up industry       0.00      0.00      0.00         1

Bubbles makes plain water so much more fun to drink       0.00
0.00      0.00         1

Combined with a IPX4 Sweat/Dust/Splash Resistant design       0.00
0.00      0.00         4

Growing up in the South       0.00      0.00      0.00         1

In our business model       0.00      0.00      0.00         1

Over time       0.00      0.00      0.00         1

Submerge yourself in the limitless world of sound with the latest in
Bluetooth innovation       0.00      0.00      0.00         2
                                        The Sodamaker plays a direct
role in protecting our planet from plastic waste. By using a Sodamaker
to make Soda and Soft Drinks at home       0.00      0.00      0.00
1
 This is a multifaceted portrait of Muhammad Ali only he could render:
sports legend; unapologetic anti-war advocate; outrageous showman and
gracious goodwill ambassador; fighter       0.00      0.00      0.00
3

Who better to tell the tale than the man who went the distance living
it?"       0.00      0.00      0.00         1

Books       0.98      0.06      0.11      6058

Clothing & Accessories       0.27      1.00      0.43      6060

Electronics       0.97      0.09      0.16      5241
```

```
Household       0.95      0.07      0.13      6084


accuracy                           0.31     23459

macro avg       0.23      0.09      0.06     23459

weighted avg       0.79      0.31      0.21     23459

Confusion Matrix:
 [[   0    0    0    0    0    0    0    0    0    0    0    1    0
0]
 [   0    0    0    0    0    0    0    0    0    0    0    1    0
0]
 [   0    0    0    0    0    0    0    0    0    0    0    4    0
0]
 [   0    0    0    0    0    0    0    0    0    0    0    1    0
0]
 [   0    0    0    0    0    0    0    0    0    0    0    1    0
0]
 [   0    0    0    0    0    0    0    0    0    0    0    1    0
0]
 [   0    0    0    0    0    0    0    0    0    0    0    2    0
0]
 [   0    0    0    0    0    0    0    0    0    0    0    1    0
0]
 [   0    0    0    0    0    0    0    0    0    0    0    3    0
0]
 [   0    0    0    0    0    0    0    0    0    0    0    1    0
0]
 [   0    0    0    0    0    0    0    0    0    0  345 5704    3
6]
 [   0    0    0    0    0    0    0    0    0    0    0 6052    0
8]
 [   0    0    0    0    0    0    0    0    0    0    0 4772  462
7]
 [   0    0    0    0    0    0    0    0    0    0    6 5650   10
418]]

/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_classification.py:1565: UndefinedMetricWarning: Precision is ill-
defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))
```

Logistic Regression Model

```
from sklearn.linear_model import LogisticRegression

log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
y_pred_lr = log_reg.predict(X_test)

print("Logistic Regression Performance:")
print("Accuracy:", accuracy_score(y_test, y_pred_lr))
print("Classification Report:\n", classification_report(y_test,
y_pred_lr))
```

```
Logistic Regression Performance:
Accuracy: 0.3107975617033974

/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_classification.py:1565: UndefinedMetricWarning: Precision is ill-
defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))
/usr/local/lib/python3.11/dist-packages/sklearn/metrics/_classificatio
n.py:1565: UndefinedMetricWarning: Precision is ill-defined and being
set to 0.0 in labels with no predicted samples. Use `zero_division`
parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))

Classification Report:

precision    recall  f1-score   support


What has been the best part of your startup experience? In this
dynamic start-up industry       0.00      0.00      0.00         1

Bubbles makes plain water so much more fun to drink       0.00
0.00      0.00         1

Combined with a IPX4 Sweat/Dust/Splash Resistant design       0.00
0.00      0.00         4

Growing up in the South       0.00      0.00      0.00         1

In our business model       0.00      0.00      0.00         1

Over time       0.00      0.00      0.00         1

Submerge yourself in the limitless world of sound with the latest in
Bluetooth innovation       0.00      0.00      0.00         2
                                          The Sodamaker plays a direct
```

```
role in protecting our planet from plastic waste. By using a Sodamaker
to make Soda and Soft Drinks at home       0.00      0.00      0.00
1
 This is a multifaceted portrait of Muhammad Ali only he could render:
sports legend; unapologetic anti-war advocate; outrageous showman and
gracious goodwill ambassador; fighter       0.00      0.00      0.00
3

Who better to tell the tale than the man who went the distance living
it?"       0.00      0.00      0.00         1

Books       0.98      0.06      0.11      6058

Clothing & Accessories       0.27      1.00      0.43      6060

Electronics       0.97      0.09      0.17      5241

Household       0.96      0.07      0.13      6084


accuracy                           0.31     23459

macro avg       0.23      0.09      0.06     23459

weighted avg       0.79      0.31      0.21     23459


/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_classification.py:1565: UndefinedMetricWarning: Precision is ill-
defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))
```

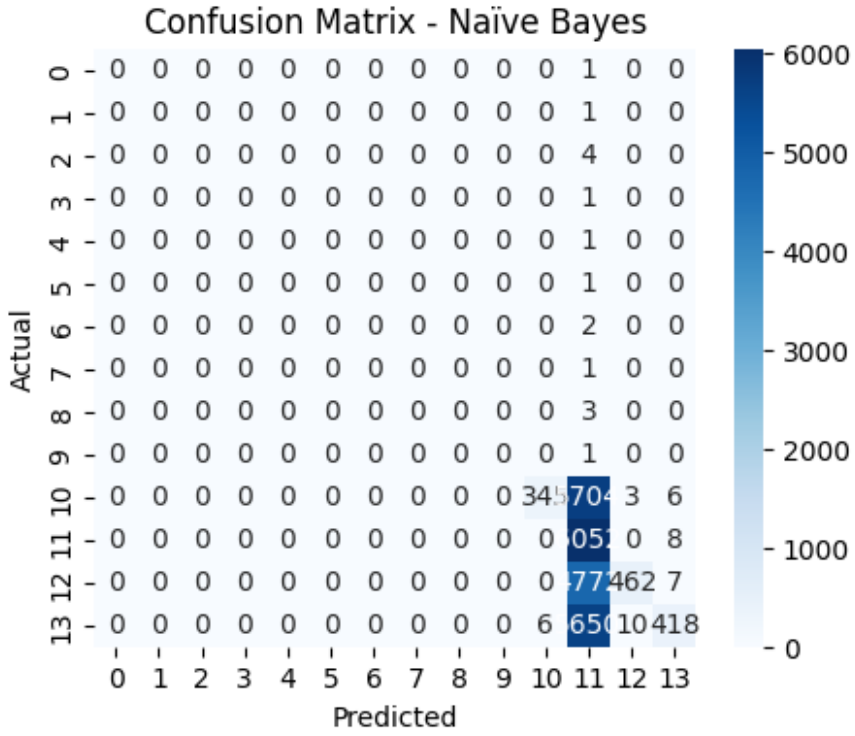
Task 5.Model Evaluation

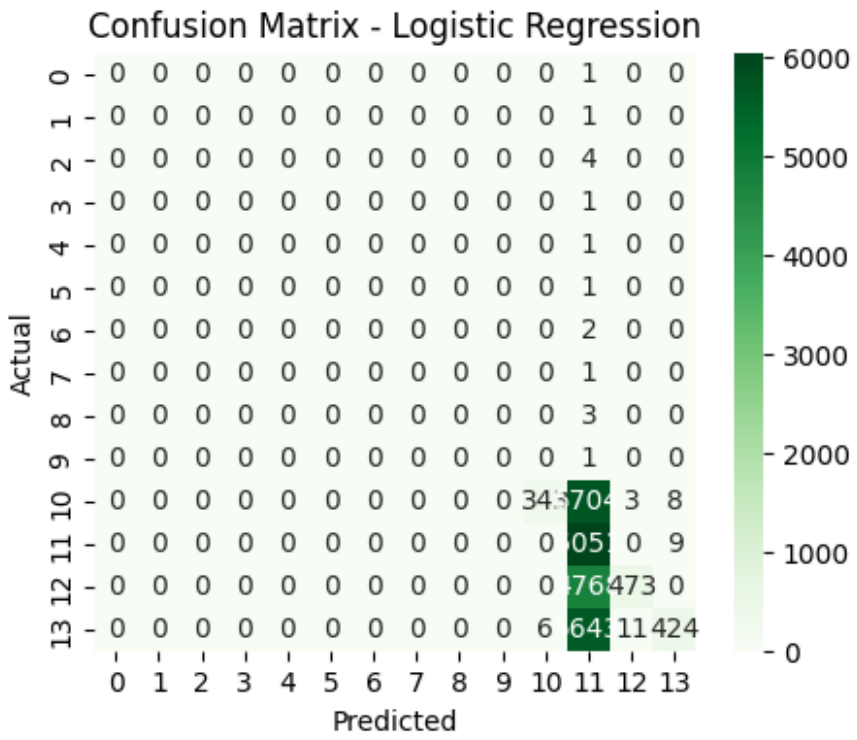
Confusion Matrix Visualization

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

# Confusion Matrix for Naïve Bayes
conf_mat_nb = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(5, 4))
sns.heatmap(conf_mat_nb, annot=True, cmap="Blues", fmt='g')
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix - Naïve Bayes")
plt.show()
```

```
# Confusion Matrix for Logistic Regression
conf_mat_lr = confusion_matrix(y_test, y_pred_lr)
plt.figure(figsize=(5, 4))
sns.heatmap(conf_mat_lr, annot=True, cmap="Greens", fmt='g')
plt.xlabel("Predicted")
plt.ylabel("Actual")
plt.title("Confusion Matrix - Logistic Regression")
plt.show()
```





Evaluation Metrics

```
from sklearn.metrics import precision_score, recall_score, f1_score

print("Naïve Bayes Model Performance:")
print("Accuracy:", accuracy_score(y_test, y_pred))
print("Precision:", precision_score(y_test, y_pred,
average='weighted'))
print("Recall:", recall_score(y_test, y_pred, average='weighted'))
print("F1 Score:", f1_score(y_test, y_pred, average='weighted'))
```

```
Naïve Bayes Model Performance:
Accuracy: 0.3102007758216463

/usr/local/lib/python3.11/dist-packages/sklearn/metrics/
_classification.py:1565: UndefinedMetricWarning: Precision is ill-
defined and being set to 0.0 in labels with no predicted samples. Use
`zero_division` parameter to control this behavior.
  _warn_prf(average, modifier, f"{metric.capitalize()} is",
len(result))

Precision: 0.7885012758231016
Recall: 0.3102007758216463
F1 Score: 0.20782064463867936
```

Conclusion

In this project, we successfully applied Natural Language Processing (NLP) techniques to perform text classification. The workflow included preprocessing textual data, extracting features using TF-IDF, and training classification models such as Naive Bayes and Logistic Regression. Through performance evaluation, we gained insights into the effectiveness of each model. This project provided valuable hands-on experience in building a complete NLP pipeline for real-world text classification tasks.

Declaration

I, Sakshi Dube, confirm that the work submitted in this assignment is my own and has been completed following academic integrity guidelines. The code is uploaded on my GitHub repository account, and the repository link is provided below:

GitHub Repository Link: https://github.com/Sakshid27/NLPTextClassification

Signature: Sakshi Dube