Natural Language Processing

Author identification using Naive Bayes

(Artificial Intelligence – AI 502)

(Fall-2025)

Submitted By: Syed Nouman (2025MSAI104)
To: Dr. Faisal Shahzad

University of Engineering and Technology, Lahore

Compute perplexity on 10 unseen sentences from Reuters.

Naive Bayes + TF-IDF

```
Naive Bayes (TF-IDF)
Accuracy: 0.143
                precision    recall  f1-score   support

Charles Dickens       0.17      0.67      0.27         3
Herman Melville       0.00      0.00      0.00         3
    Jane Austen       0.00      0.00      0.00         3
     Mark Twain       0.00      0.00      0.00         3
   Mary Shelley       0.33      0.33      0.33         3
    Oscar Wilde       0.00      0.00      0.00         3
    Shakespeare       0.00      0.00      0.00         3

       accuracy                           0.14        21
      macro avg       0.07      0.14      0.09        21
   weighted avg       0.07      0.14      0.09        21
```

Naive Bayes + CountVectorizer

```
Naive Bayes (CountVectorizer)
Accuracy: 0.19
                precision    recall  f1-score   support

Charles Dickens       0.21      1.00      0.35         3
Herman Melville       0.00      0.00      0.00         3
    Jane Austen       0.00      0.00      0.00         3
     Mark Twain       0.00      0.00      0.00         3
   Mary Shelley       0.33      0.33      0.33         3
    Oscar Wilde       0.00      0.00      0.00         3
    Shakespeare       0.00      0.00      0.00         3

       accuracy                           0.19        21
      macro avg       0.08      0.19      0.10        21
   weighted avg       0.08      0.19      0.10        21
```

Logistic Regression + TF-IDF

```
Logistic Regression (TF-IDF)
Accuracy: 0.095
                precision    recall  f1-score   support

Charles Dickens       0.14      0.33      0.20         3
Herman Melville       0.00      0.00      0.00         3
    Jane Austen       0.00      0.00      0.00         3
     Mark Twain       0.00      0.00      0.00         3
   Mary Shelley       0.33      0.33      0.33         3
    Oscar Wilde       0.00      0.00      0.00         3
    Shakespeare       0.00      0.00      0.00         3

       accuracy                           0.10        21
      macro avg       0.07      0.10      0.08        21
   weighted avg       0.07      0.10      0.08        21
```

Logistic Regression + CountVectorizer

```
Logistic Regression (CountVectorizer)
Accuracy: 0.095
                precision    recall  f1-score   support

Charles Dickens       0.17      0.33      0.22         3
Herman Melville       0.00      0.00      0.00         3
    Jane Austen       0.00      0.00      0.00         3
     Mark Twain       0.00      0.00      0.00         3
   Mary Shelley       0.00      0.00      0.00         3
    Oscar Wilde       0.17      0.33      0.22         3
    Shakespeare       0.00      0.00      0.00         3

       accuracy                           0.10        21
      macro avg       0.05      0.10      0.06        21
   weighted avg       0.05      0.10      0.06        21
```

Test sample for Naive Bayes (CountVectorizer)

```
Sample Predictions:

Text: O Romeo, Romeo! wherefore art thou Romeo?
→ Predicted Author: Charles Dickens

Text: It was a bright cold day in April, and the clocks were striking thirteen.
→ Predicted Author: Charles Dickens

Text: Vanity and pride are different things, though the words are often used synonymously.
→ Predicted Author: Jane Austen

Text: To be, or not to be, that is the question.
→ Predicted Author: Shakespeare
```

Test sample for Logistic Regression (CountVectorizer)

```
Sample Predictions:

Text: O Romeo, Romeo! wherefore art thou Romeo?
→ Predicted Author: Oscar Wilde

Text: It was a bright cold day in April, and the clocks were striking thirteen.
→ Predicted Author: Charles Dickens

Text: Vanity and pride are different things, though the words are often used synonymously.
→ Predicted Author: Jane Austen

Text: To be, or not to be, that is the question.
→ Predicted Author: Shakespeare
```