**Section: Data**

Topics: Artificial Intelligence, Data Science, and Machine Learning.

Graphical user interface

Description automatically generated with low confidence

**Section: Cleaning and Preparing the Data**

Part 1: Training Data

A picture containing table

Description automatically generatedA picture containing table

Description automatically generated

Training Features DataFrame for DF1 (CountVectorizer) and DF2 (TfidfVectorizer with Stemming)

Part 2: Training Labels

Text

Description automatically generated

Training Target DataFrame for DF1 (CountVectorizer) and DF2 (TfidfVectorizer with Stemming)

Part 3: Testing Data

Calendar

Description automatically generatedTable

Description automatically generated with low confidence

Testing Features DataFrame for DF1 (CountVectorizer) and DF2 (TfidfVectorizer with Stemming)

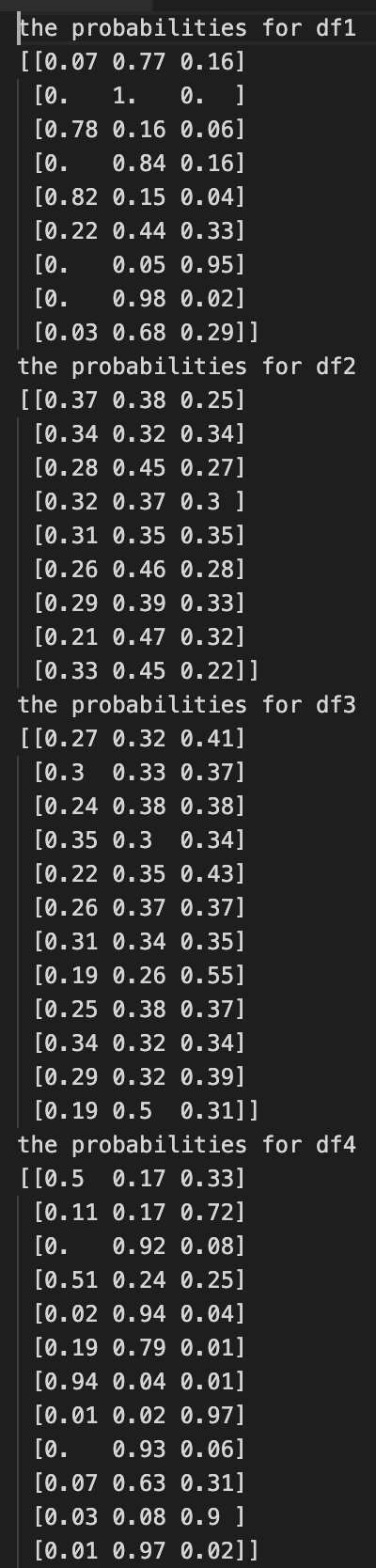
Part 4: Testing Labels

Graphical user interface, text

Description automatically generated

Testing Target DataFrame for DF1 (CountVectorizer) and DF2 (TfidfVectorizer with Stemming)

**Section: Naïve Bayes (NB) and Results**



Chart, bar chart

Description automatically generated

The prediction from NB1 is:

['Data\_Science' 'Data\_Science' 'Artificial\_Intelligence' 'Data\_Science'

'Artificial\_Intelligence' 'Data\_Science' 'Machine\_Learning'

'Data\_Science' 'Data\_Science']

The actual labels are:

8 Data\_Science

5 Data\_Science

2 Artificial\_Intelligence

6 Data\_Science

6 Artificial\_Intelligence

0 Artificial\_Intelligence

1 Machine\_Learning

0 Data\_Science

1 Data\_Science

Name: Label, dtype: object

The prediction from NB2 is:

['Data\_Science' 'Artificial\_Intelligence' 'Data\_Science' 'Data\_Science'

'Machine\_Learning' 'Data\_Science' 'Data\_Science' 'Data\_Science'

'Data\_Science']

The actual labels are:

7 Artificial\_Intelligence

8 Machine\_Learning

4 Data\_Science

1 Artificial\_Intelligence

2 Machine\_Learning

1 Data\_Science

5 Machine\_Learning

3 Machine\_Learning

0 Artificial\_Intelligence

Name: Label, dtype: object

The prediction from NB3 is:

['Machine\_Learning' 'Machine\_Learning' 'Machine\_Learning'

'Artificial\_Intelligence' 'Machine\_Learning' 'Machine\_Learning'

'Machine\_Learning' 'Machine\_Learning' 'Data\_Science' 'Machine\_Learning'

'Machine\_Learning' 'Data\_Science']

The actual labels are:

8 Machine\_Learning

2 Artificial\_Intelligence

5 Data\_Science

7 Artificial\_Intelligence

3 Data\_Science

7 Data\_Science

6 Artificial\_Intelligence

4 Data\_Science

9 Machine\_Learning

7 Machine\_Learning

4 Artificial\_Intelligence

0 Artificial\_Intelligence

Name: Label, dtype: object

The prediction from NB4 is:

['Artificial\_Intelligence' 'Machine\_Learning' 'Data\_Science'

'Artificial\_Intelligence' 'Data\_Science' 'Data\_Science'

'Artificial\_Intelligence' 'Machine\_Learning' 'Data\_Science'

'Data\_Science' 'Machine\_Learning' 'Data\_Science']

The actual labels are:

4 Artificial\_Intelligence

1 Artificial\_Intelligence

6 Data\_Science

0 Artificial\_Intelligence

8 Data\_Science

5 Data\_Science

2 Data\_Science

5 Machine\_Learning

0 Data\_Science

7 Data\_Science

9 Machine\_Learning

9 Data\_Science

Name: Label, dtype: object

**Section: Decision Trees (DT)**

Actual for DataFrame: 1

8 Data\_Science

5 Data\_Science

2 Artificial\_Intelligence

6 Data\_Science

6 Artificial\_Intelligence

0 Artificial\_Intelligence

1 Machine\_Learning

0 Data\_Science

1 Data\_Science

Name: Label, dtype: object

Prediction

['Data\_Science' 'Data\_Science' 'Artificial\_Intelligence' 'Data\_Science'

'Artificial\_Intelligence' 'Artificial\_Intelligence' 'Machine\_Learning'

'Data\_Science' 'Data\_Science']

Actual for DataFrame: 2

7 Artificial\_Intelligence

8 Machine\_Learning

4 Data\_Science

1 Artificial\_Intelligence

2 Machine\_Learning

1 Data\_Science

5 Machine\_Learning

3 Machine\_Learning

0 Artificial\_Intelligence

Name: Label, dtype: object

Prediction

['Machine\_Learning' 'Machine\_Learning' 'Data\_Science' 'Machine\_Learning'

'Machine\_Learning' 'Data\_Science' 'Machine\_Learning' 'Machine\_Learning'

'Artificial\_Intelligence']

Actual for DataFrame: 3

8 Machine\_Learning

2 Artificial\_Intelligence

5 Data\_Science

7 Artificial\_Intelligence

3 Data\_Science

7 Data\_Science

6 Artificial\_Intelligence

4 Data\_Science

9 Machine\_Learning

7 Machine\_Learning

4 Artificial\_Intelligence

0 Artificial\_Intelligence

Name: Label, dtype: object

Prediction

['Machine\_Learning' 'Artificial\_Intelligence' 'Data\_Science'

'Artificial\_Intelligence' 'Data\_Science' 'Machine\_Learning'

'Artificial\_Intelligence' 'Data\_Science' 'Machine\_Learning'

'Machine\_Learning' 'Artificial\_Intelligence' 'Artificial\_Intelligence']

Actual for DataFrame: 4

4 Artificial\_Intelligence

1 Artificial\_Intelligence

6 Data\_Science

0 Artificial\_Intelligence

8 Data\_Science

5 Data\_Science

2 Data\_Science

5 Machine\_Learning

0 Data\_Science

7 Data\_Science

9 Machine\_Learning

9 Data\_Science

Name: Label, dtype: object

Prediction

['Artificial\_Intelligence' 'Artificial\_Intelligence' 'Data\_Science'

'Data\_Science' 'Data\_Science' 'Data\_Science' 'Artificial\_Intelligence'

'Machine\_Learning' 'Data\_Science' 'Data\_Science' 'Machine\_Learning'

'Data\_Science']

The confusion matrix df1 is:

[[3 0 0]

[0 4 1]

[0 0 1]]

The confusion matrix df2 is:

[[1 0 2]

[0 2 0]

[0 0 4]]

The confusion matrix df3 is:

[[5 0 0]

[0 3 1]

[0 0 3]]

The confusion matrix df4 is:

[[2 1 0]

[1 6 0]

[0 0 2]]

DF1

Df2



DF3

DF4

**Section: Support Vector Machines (SVMs)**

Linear Kernel:

The confusion matrix df1 is:

[[3 0 0]

[0 5 0]

[0 0 1]]

The confusion matrix df2 is:

[[1 1 1]

[0 2 0]

[0 0 4]]

The confusion matrix df3 is:

[[2 1 2]

[0 2 2]

[1 1 1]]

The confusion matrix df4 is:

[[3 0 0]

[0 7 0]

[0 0 2]]

RBF Kernel:

The confusion matrix df1 is:

[[0 0 3]

[0 0 5]

[0 0 1]]

The confusion matrix df2 is:

[[0 3 0]

[0 2 0]

[0 4 0]]

The confusion matrix df3 is:

[[0 0 5]

[0 0 4]

[0 0 3]]

The confusion matrix df4 is:

[[0 0 3]

[0 0 7]

[0 0 2]]

POLY Kernel:

The confusion matrix df1 is:

[[0 0 3]

[0 0 5]

[0 0 1]]

The confusion matrix df2 is:

[[0 3 0]

[0 2 0]

[0 4 0]]

[LibSVM]

The confusion matrix df3 is:

[[0 0 5]

[0 0 4]

[0 0 3]]

The confusion matrix df4 is:

[[0 0 3]

[0 0 7]

[0 0 2]]

**Section: Conclusions**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | DF1 (CV) | DF2 (TF) | DF3 (CV lemming) | DF4 (TF stemming) |
| NB | 0.88 | 0.33 | 0.25 | 0.83 |
| DT | 0.88 | 0.77 | 0.91 | 0.83 |
| SVM (linear) | 1.0 | 0.77 | 0.41 | 1.0 |
| SVM (RBF) | 0.11 | 0.22 | 0.25 | 0.16 |
| SVM (POLY) | 0.11 | 0.22 | 0.25 | 0.16 |

The best is using Decision tree for Counter Vectorizer with lemming with accuracy 91%. Although SVM with linear kernel have 100% accuracy but it leads to overfitting issues so not take into consideration. And also compared different data frame, decision tree performs relatively similar performance compared to other models.