Artificial Intelligence / SVM Project / Documentation

Yazdan Mastery Farahani / 99522059

1. Linear Kernel

- **Usage**:
    - When the data is linearly separable.
    - For large datasets, it is computationally less expensive compared to other kernels.
    - When the number of features is large compared to the number of samples.

2. Polynomial Kernel

- **Usage**:
    - When the data has polynomial relationships.
    - For moderately complex relationships where the linear kernel fails.
    - Tuning the degree dallows control over the flexibility of the decision boundary.

3. Radial Basis Function (RBF) Kernel / Gaussian Kernel

- **Usage**:
    - When there is no prior knowledge about the data.
    - Effective for non-linearly separable data.
    - Versatile and can handle complex decision boundaries.
    - Suitable when the relationship between features is not linear.

4. Sigmoid Kernel

- **Usage**:
    - Similar to a two-layer neural network.
    - For certain types of datasets where sigmoid function characteristics are beneficial.
    - Rarely used in practice compared to other kernels.

Part-1. Spam Emails

```
linear : 0.98
poly : 0.87
rbf : 0.87
sigmoid : 0.87
```

Best Kernel: *Linear*

Reason:

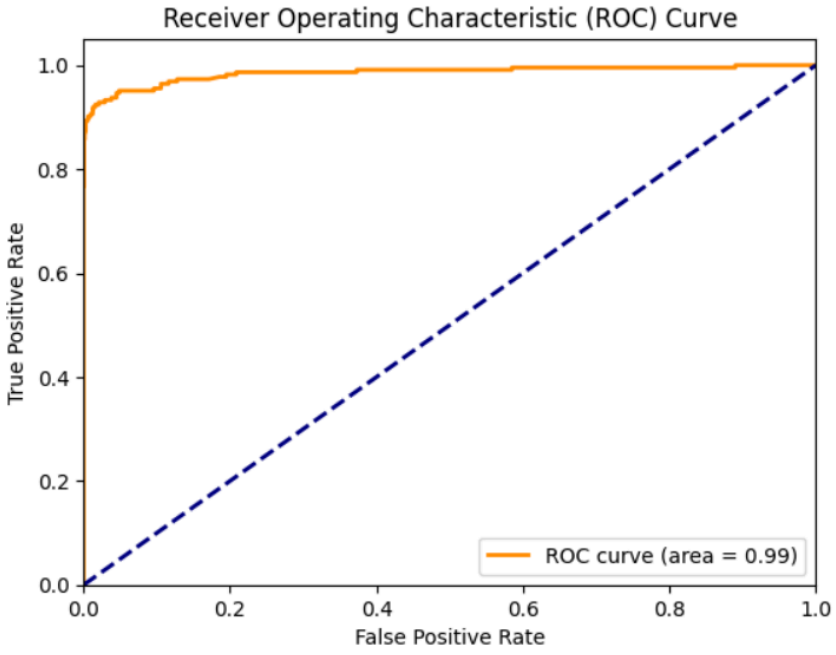
- *High Dimensionality of Text Data*
- *Simplicity*
- *Nature of the Dataset*

```
Accuracy: 0.98
Classification Report:
              precision    recall  f1-score   support

           0       0.98      1.00      0.99      1448
           1       0.98      0.87      0.92       224

    accuracy                           0.98      1672
   macro avg       0.98      0.93      0.96      1672
weighted avg       0.98      0.98      0.98      1672

Confusion Matrix:
[[1445    3]
 [  29  195]]
```



Part-2. Diabetes

```
linear : 77 %
rbf : 79 %
sigmoid : 53 %
```

Best Kernel: *RBF*

Reason:

- *Non-linear Relationships*
- *Automatic Adaptation*

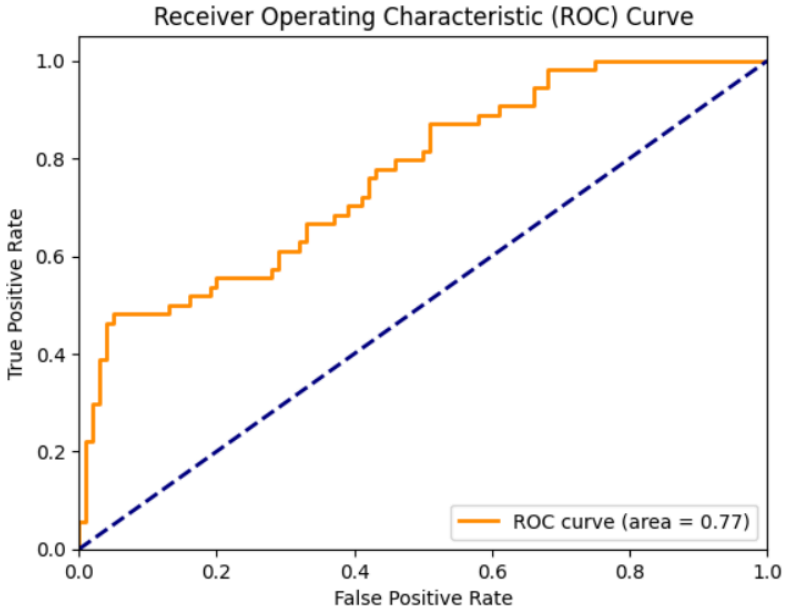
*Poly Kernel (Not Working): Likely due to high computational complexity and potential overfitting, leading to very slow or impractical training times.

```
Accuracy: 0.79
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.95      0.85       100
           1       0.84      0.48      0.61        54

    accuracy                           0.79       154
   macro avg       0.81      0.72      0.73       154
weighted avg       0.80      0.79      0.77       154

Confusion Matrix:
[[95  5]
 [28 26]]
```



Part-3. Boston Housing

Selected Kernel: *RBF*

Reason:

- *Non-linear Relationships*
- *Automatic Adaptation*

```
CV:  0.7826006932475 84
R2_score (train):  0.8742715669147498
R2_score (test):  0.8044070157723574
RMSE:  3.8388977809328355
```

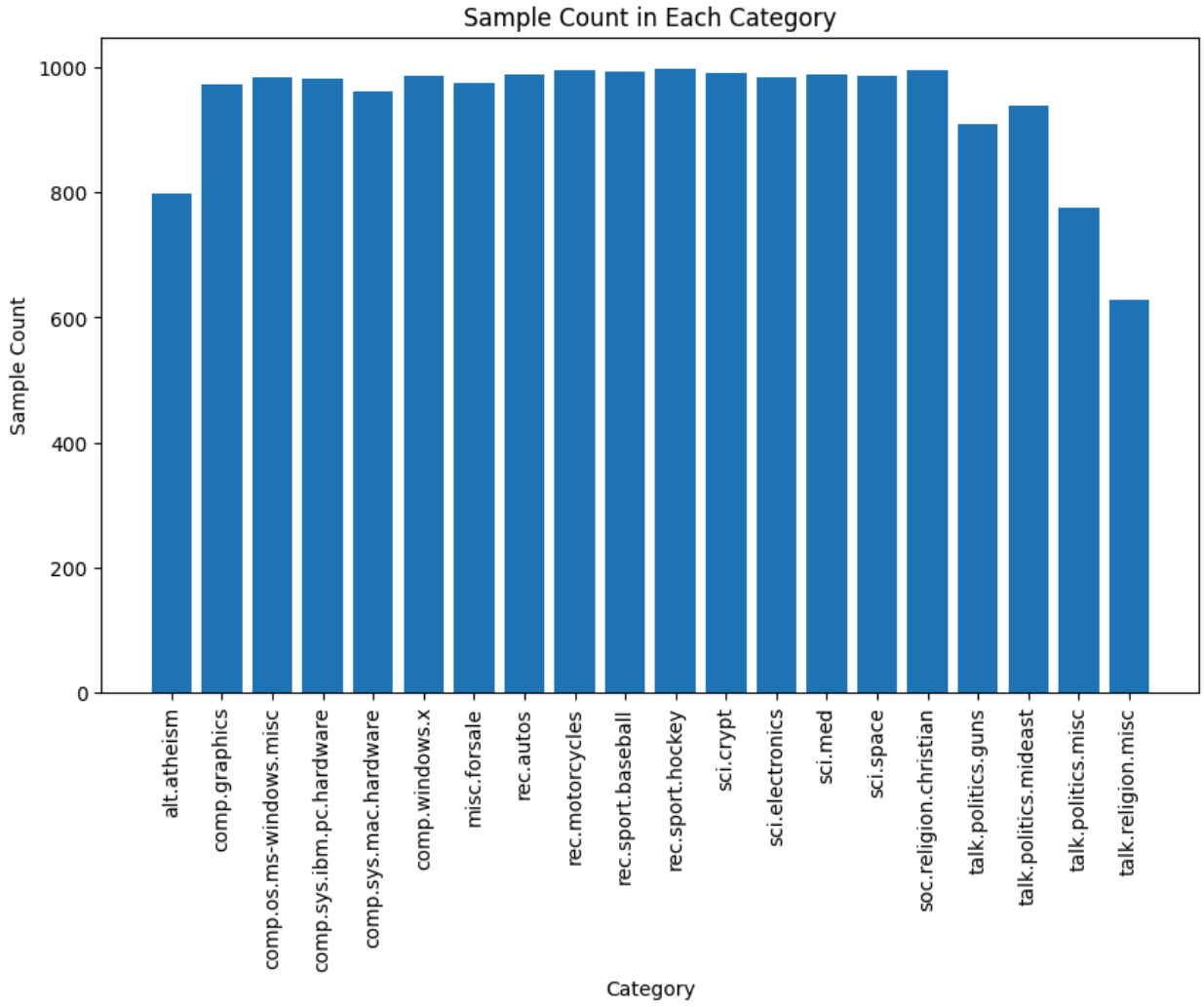
Part-4. 20NewsGroup

Best Kernel: *Linear*

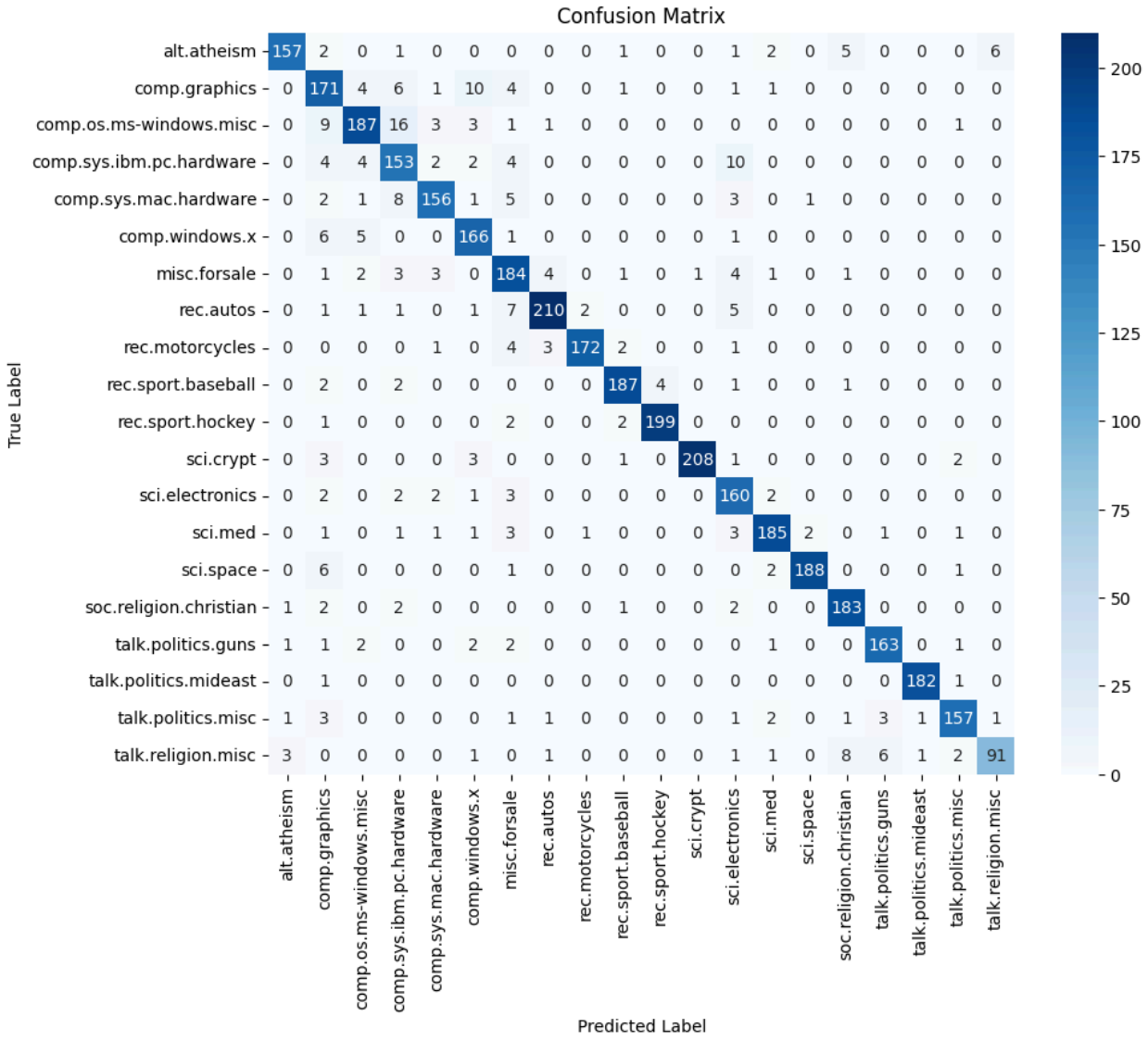
Reason:

- *Simplicity*
- *High Dimensionality of Text Data*
- *Nature of the Dataset*

```
linear : 69%
rbf : 45%
sigmoid : 67%
poly : 11%
```




```
SVM Accuracy: 91.751%
SVM Recall: 91.508%
SVM F1: 91.650%
```


Confusion Matrix
True Label
Predicted Label
alt.atheism
comp.graphics
comp.os.ms-windows.misc
comp.sys.ibm.pc.hardware
comp.sys.mac.hardware
comp.windows.x
misc.forsale
rec.autos
rec.motorcycles
rec.sport.baseball
rec.sport.hockey
sci.crypt
sci.electronics
sci.med
sci.space
soc.religion.christian
talk.politics.guns
talk.politics.mideast
talk.politics.misc
talk.religion.misc
200
175
150
125
100
75
50
25
0

Part5_HealthExaminationSurvay

```
Accuracy Score: 0.963294538943599
Prediction: [0. 0. 0. ... 0. 0. 0.]
Classification Report:
              precision    recall  f1-score   support

         0.0       0.96      1.00      0.98      1076
         1.0       0.00      0.00      0.00        41

    accuracy                           0.96      1117
   macro avg       0.48      0.50      0.49      1117
weighted avg       0.93      0.96      0.95      1117
```