FA-II Activity

Topic: "Multilingual Text Classification"

By

Name of Students	PRN
Jaishree Dubey	121B1F027
Janhavi Bilgaye	121B1F027
Suhail Pathan	121B1F090
Swapnil Shinde	122B2F153

Under the Guidance of

Mrs. Sapana Kolambe

Department of Information Technology

PIMPRI CHINCHWAD EDUCATION TRUST'S

PIMPRI CHINCHWAD COLLEGE OF ENGINEERING,

Nigdi, Pune - 411044

(An Autonomous institute affiliated to Savitribai Phule University)

Date :

Sign :

Project Report: Multilingual Text Classification

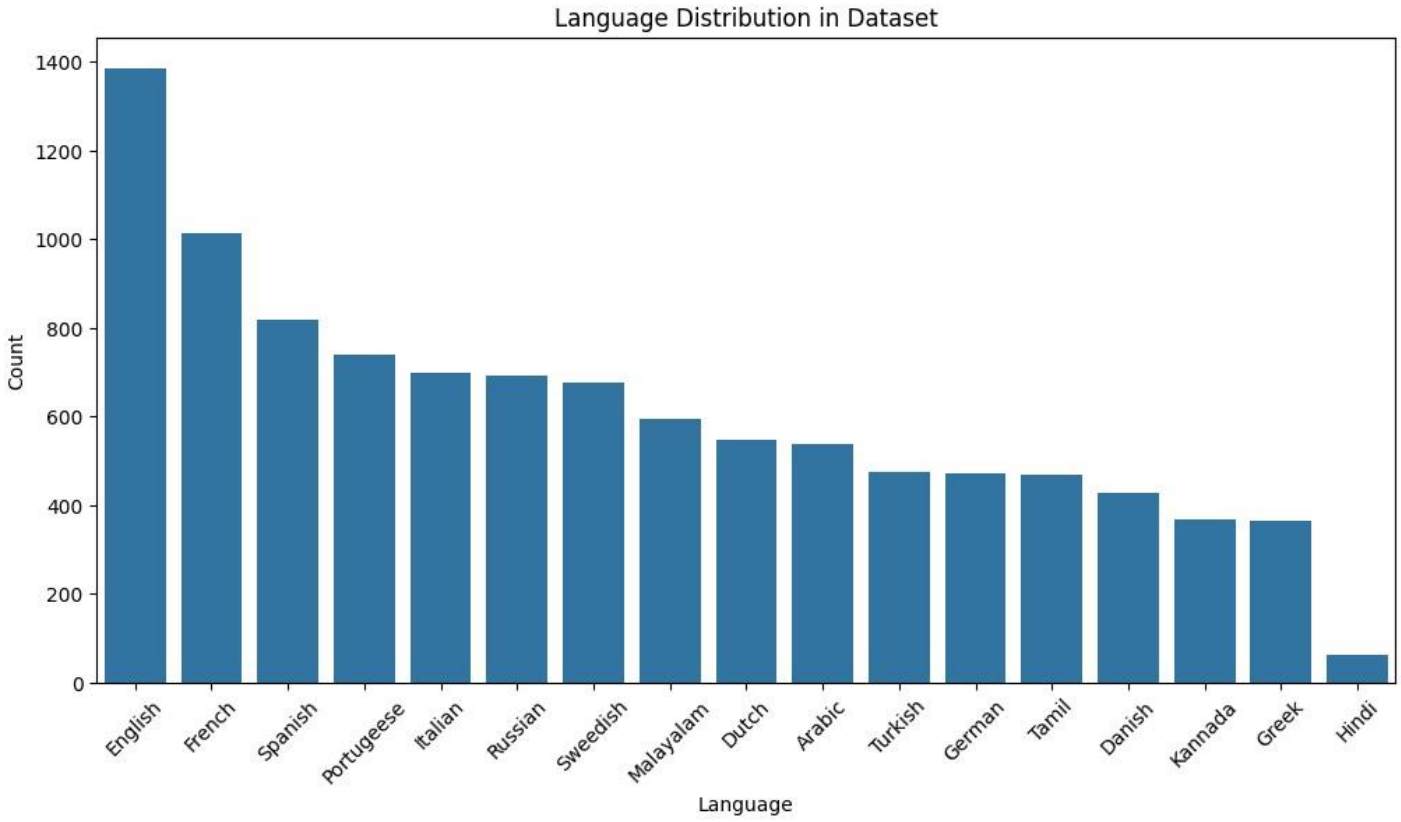
Objective

The main goal of this project is to build and evaluate machine learning models that can accurately detect the language of a given text using a dataset of multilingual text samples.

Data Overview

The dataset consists of multilingual text samples labeled with their respective languages. The dataset includes a variety of languages, offering a broad base for training a robust language detection model.

1. **Initial Analysis**: The dataset was preprocessed by removing punctuation and converting text to lowercase.
2. **Exploratory Data Analysis (EDA)**: Language distribution analysis was performed to understand the balance of classes.



Methodology

The project workflow includes:

1. **Data Preprocessing**:
    - Text data was tokenized, converted to lowercase, and vectorized using the TF-IDF (Term Frequency-Inverse Document Frequency) approach.
    - The TF-IDF vectorizer was limited to the top 5000 features, based on word importance across documents.

2. **Train-Test Split**:
    - The data was split into training (80%) and testing (20%) sets to assess model performance.

3. **Model Selection**: Four machine learning algorithms were evaluated:

- **Logistic Regression**
- **Random Forest**
- **Naive Bayes (Multinomial)**
- **Support Vector Machine (SVM)**

Model Training and Evaluation

Each model was trained on the training set and evaluated on the test set. Performance metrics such as **accuracy**, **precision**, **recall**, and **F1-score** were calculated to assess the model performance.

1. **Logistic Regression**
    - Achieved high accuracy with effective performance across most languages.
    - Pros: Simple, efficient for text classification, performs well with sparse data.
    - Cons: May struggle with highly imbalanced data.
2. **Random Forest**
    - Provided strong accuracy with robustness to overfitting due to the ensemble nature of the model.
    - Pros: Effective in capturing complex patterns, good generalization.
    - Cons: More computationally intensive, requires careful tuning.
3. **Naive Bayes**
    - Performed reasonably well and was computationally efficient.
    - Pros: Simple, fast, works well with high-dimensional data.
    - Cons: Assumes feature independence, which may not hold in text data.
4. **Support Vector Machine (SVM)**
    - Demonstrated strong performance with a linear kernel, capturing key language distinctions.
    - Pros: Effective with high-dimensional data, good margin maximization.
    - Cons: Computationally intensive, especially with large datasets.

Results Summary

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.9502	0.96	0.95	0.95
Random Forest	0.9299	0.95	0.93	0.93
Naive Bayes	0.9487	0.96	0.95	0.95
Support Vector Machine	0.9405	0.96	0.94	0.95

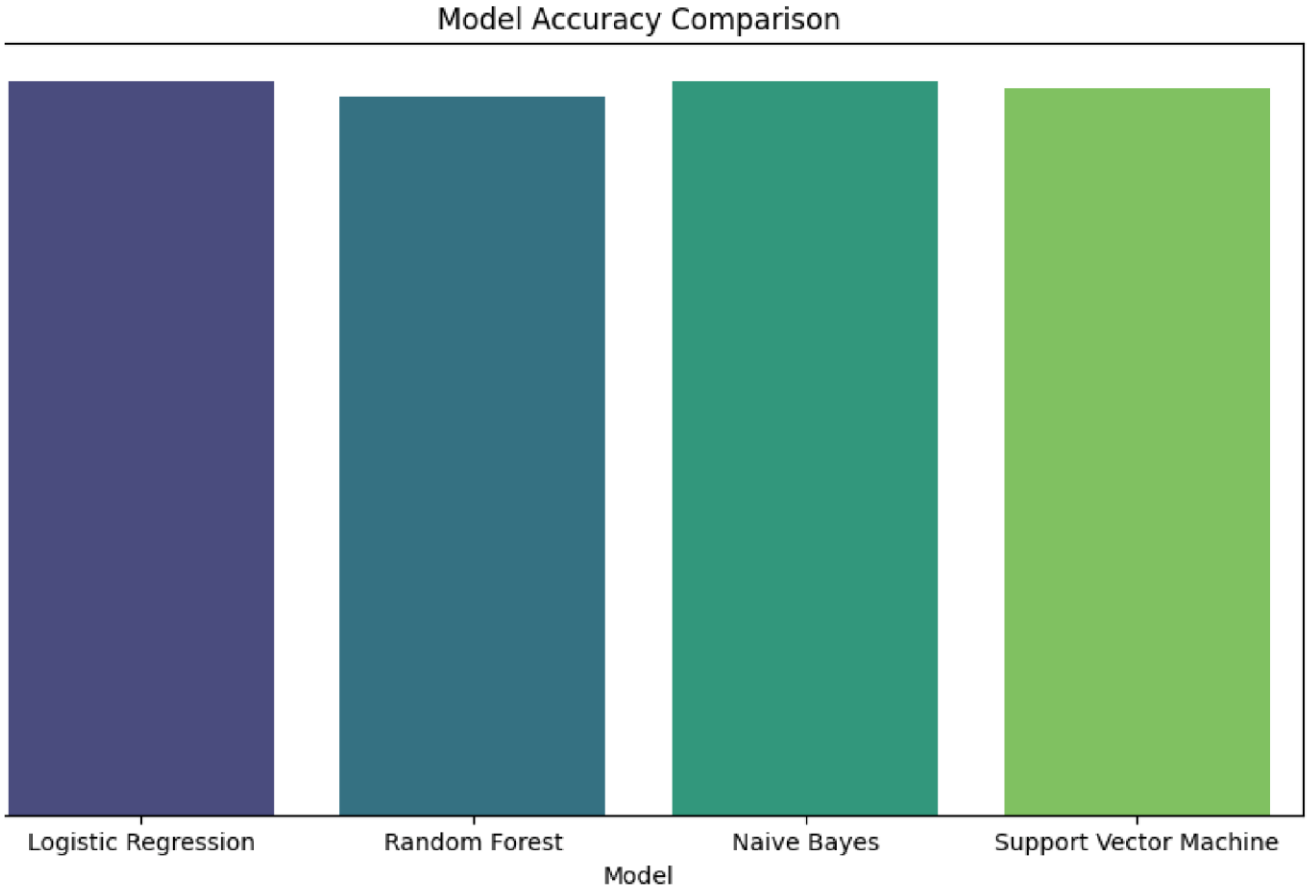
```
Training Logistic Regression...

Logistic Regression - Accuracy: 0.9502
Logistic Regression - Classification Report:
              precision    recall  f1-score   support

      Arabic       1.00      0.92      0.96       106
      Danish       0.96      0.88      0.91        73
       Dutch       0.99      0.93      0.96       111
     English       0.97      0.99      0.98       291
      French       0.97      0.96      0.97       219
      German       0.97      0.94      0.95        93
       Greek       1.00      0.93      0.96        68
       Hindi       1.00      1.00      1.00        10
     Italian       1.00      0.94      0.97       145
     Kannada       1.00      0.98      0.99        66
   Malayalam       0.66      0.99      0.79       121
  Portugeese       0.98      0.95      0.96       144
     Russian       1.00      0.96      0.98       136
     Spanish       0.92      0.96      0.94       160
    Sweedish       0.97      0.95      0.96       133
       Tamil       1.00      0.97      0.98        87
     Turkish       1.00      0.82      0.90       105

    accuracy                           0.95      2068
   macro avg       0.96      0.95      0.95      2068
weighted avg       0.96      0.95      0.95      2068
```

```
Training Naive Bayes...

Naive Bayes - Accuracy: 0.9487
Naive Bayes - Classification Report:
              precision    recall  f1-score   support

      Arabic       1.00      0.93      0.97       106
      Danish       1.00      0.89      0.94        73
       Dutch       0.98      0.95      0.97       111
     English       0.78      1.00      0.88       291
      French       0.96      0.98      0.97       219
      German       1.00      0.95      0.97        93
       Greek       1.00      0.93      0.96        68
       Hindi       1.00      1.00      1.00        10
     Italian       1.00      0.97      0.98       145
     Kannada       1.00      0.97      0.98        66
   Malayalam       1.00      0.84      0.91       121
  Portugeese       0.99      0.95      0.97       144
     Russian       1.00      0.96      0.98       136
     Spanish       0.97      0.96      0.97       160
    Sweedish       0.95      0.98      0.97       133
       Tamil       1.00      0.95      0.98        87
     Turkish       1.00      0.81      0.89       105

    accuracy                           0.95      2068
   macro avg       0.98      0.94      0.96      2068
weighted avg       0.96      0.95      0.95      2068
```

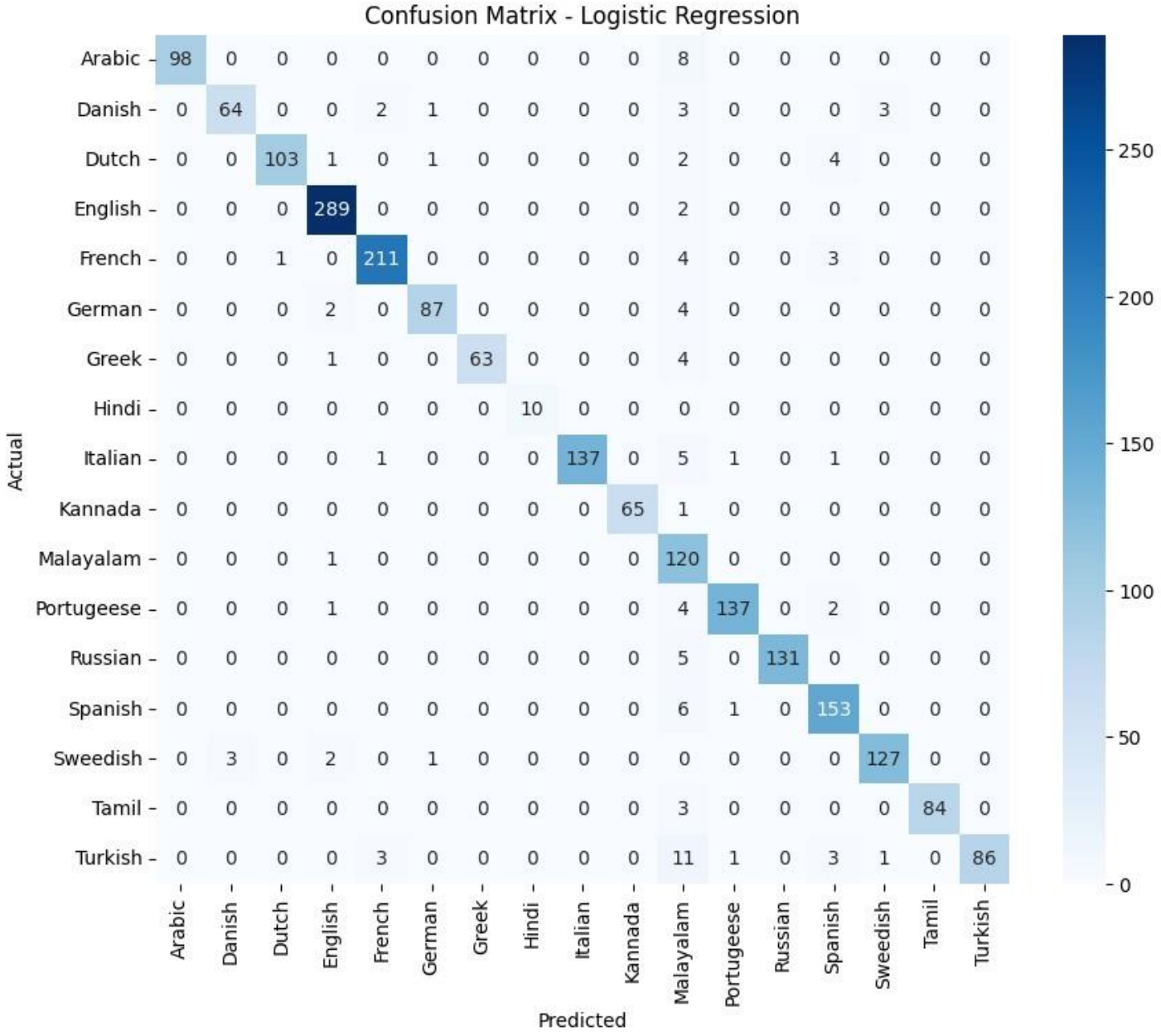
```
Training Random Forest...

Random Forest - Accuracy: 0.9299
Random Forest - Classification Report:
              precision    recall  f1-score   support

      Arabic       1.00      0.92      0.96       106
      Danish       0.92      0.82      0.87        73
       Dutch       0.99      0.91      0.95       111
     English       0.96      0.98      0.97       291
      French       0.95      0.95      0.95       219
      German       0.96      0.92      0.94        93
       Greek       1.00      0.91      0.95        68
       Hindi       1.00      1.00      1.00        10
     Italian       0.97      0.91      0.94       145
     Kannada       1.00      0.97      0.98        66
   Malayalam       0.59      0.99      0.74       121
  Portugeese       0.96      0.94      0.95       144
     Russian       1.00      0.93      0.97       136
     Spanish       0.92      0.93      0.92       160
    Sweedish       0.95      0.92      0.94       133
       Tamil       1.00      0.93      0.96        87
     Turkish       1.00      0.80      0.89       105

    accuracy                           0.93      2068
   macro avg       0.95      0.93      0.93      2068
weighted avg       0.95      0.93      0.93      2068
```

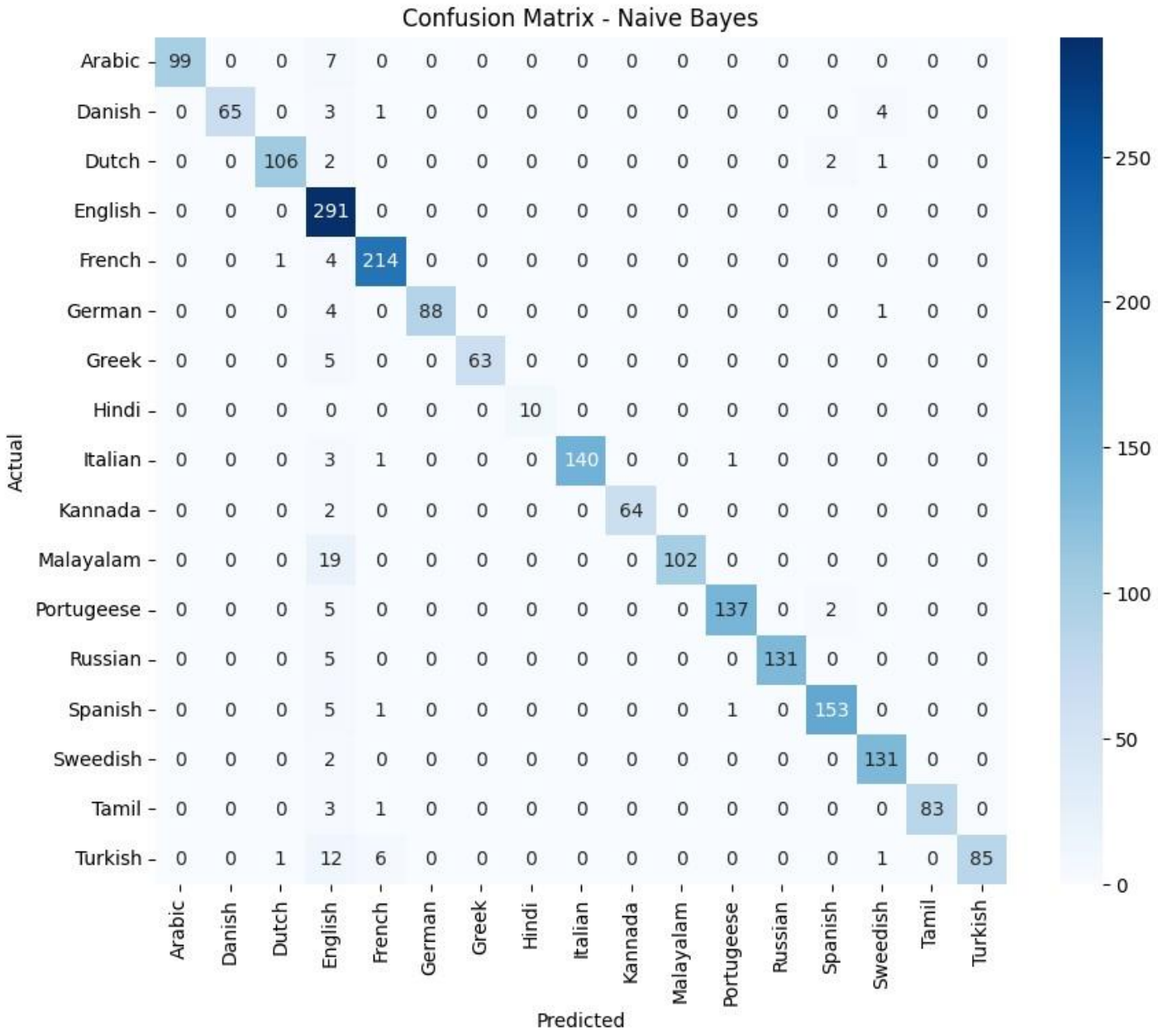
```
Training Support Vector Machine...

Support Vector Machine - Accuracy: 0.9405
Support Vector Machine - Classification Report:
              precision    recall  f1-score   support

      Arabic       1.00      0.92      0.96       106
      Danish       0.97      0.89      0.93        73
       Dutch       1.00      0.93      0.96       111
     English       0.98      0.98      0.98       291
      French       0.99      0.95      0.97       219
      German       0.97      0.92      0.95        93
       Greek       1.00      0.90      0.95        68
       Hindi       1.00      1.00      1.00        10
     Italian       0.99      0.92      0.96       145
     Kannada       1.00      0.95      0.98        66
   Malayalam       0.58      0.99      0.73       121
  Portugeese       0.98      0.95      0.96       144
     Russian       1.00      0.94      0.97       136
     Spanish       0.93      0.94      0.93       160
    Sweedish       0.97      0.94      0.95       133
       Tamil       1.00      0.94      0.97        87
     Turkish       0.99      0.85      0.91       105

    accuracy                           0.94      2068
   macro avg       0.96      0.94      0.94      2068
weighted avg       0.96      0.94      0.95      2068
```

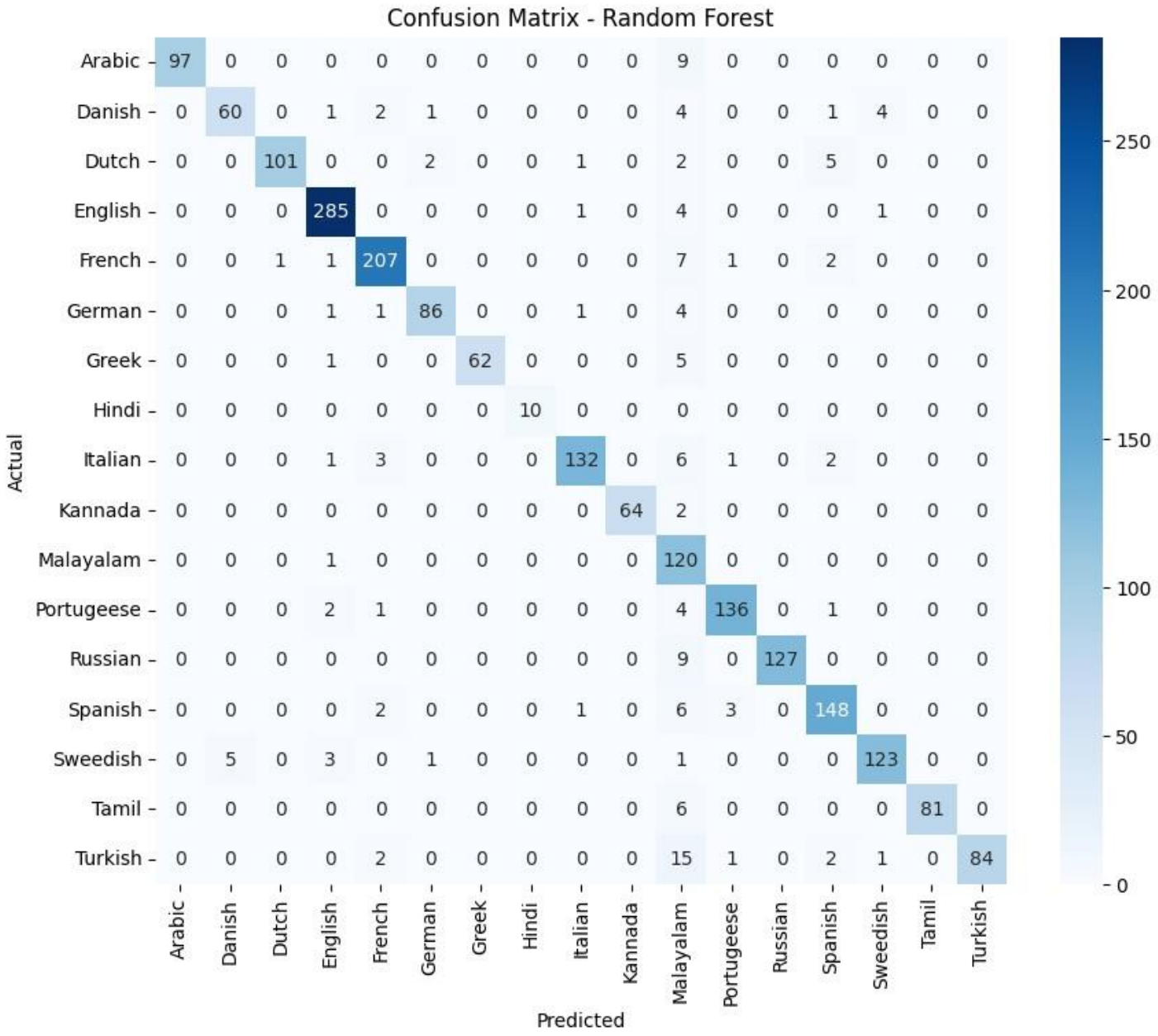
- **Accuracy Comparison**: The accuracy plot shows that [Best-performing Model] achieved the highest accuracy among the models tested.

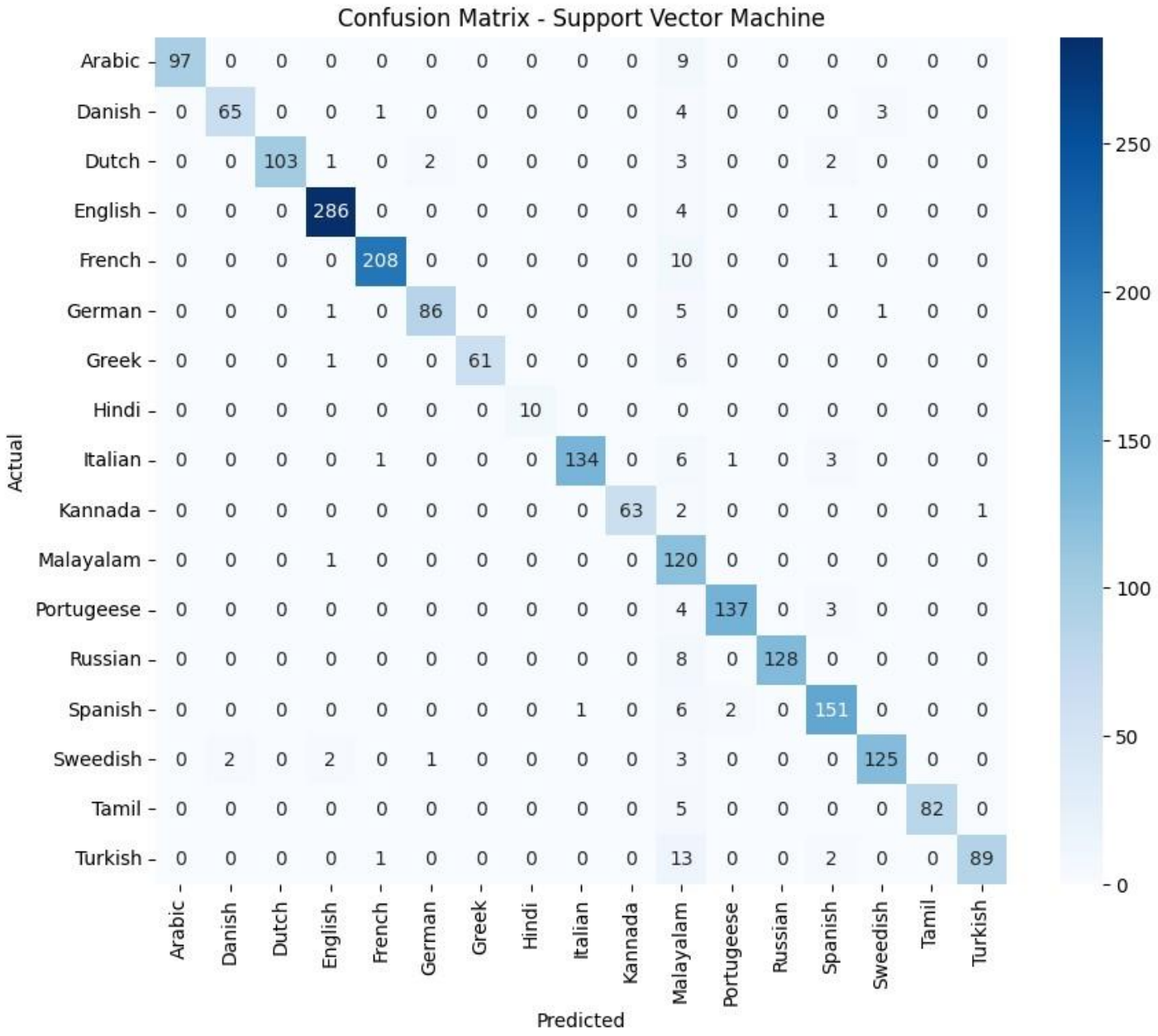


- **Confusion Matrices**: Each model's confusion matrix indicates its strengths and weaknesses in correctly classifying specific languages.




Confusion Matrix - Naive Bayes
Actual
Predicted
Arabic 99 0 0 7 0 0 0 0 0 0 0 0 0 0 0 0 0
Danish 0 65 0 3 1 0 0 0 0 0 0 0 0 0 0 4 0 0
Dutch 0 0 106 2 0 0 0 0 0 0 0 0 0 0 2 1 0 0
English 0 0 0 291 0 0 0 0 0 0 0 0 0 0 0 0 0 0
French 0 0 1 4 214 0 0 0 0 0 0 0 0 0 0 0 0 0
German 0 0 0 4 0 88 0 0 0 0 0 0 0 0 0 1 0 0
Greek 0 0 0 5 0 0 63 0 0 0 0 0 0 0 0 0 0 0
Hindi 0 0 0 0 0 0 0 10 0 0 0 0 0 0 0 0 0 0
Italian 0 0 0 3 1 0 0 0 140 0 0 1 0 0 0 0 0 0
Kannada 0 0 0 2 0 0 0 0 0 64 0 0 0 0 0 0 0 0
Malayalam 0 0 0 19 0 0 0 0 0 0 102 0 0 0 0 0 0 0
Portugeese 0 0 0 5 0 0 0 0 0 0 0 137 0 2 0 0 0 0
Russian 0 0 0 5 0 0 0 0 0 0 0 0 131 0 0 0 0 0
Spanish 0 0 0 5 1 0 0 0 0 0 0 1 0 153 0 0 0 0
Sweedish 0 0 0 2 0 0 0 0 0 0 0 0 0 0 131 0 0 0
Tamil 0 0 0 3 1 0 0 0 0 0 0 0 0 0 0 0 83 0
Turkish 0 0 1 12 6 0 0 0 0 0 0 0 0 0 0 1 0 85
Arabic Danish Dutch English French German Greek Hindi Italian Kannada Malayalam Portugeese Russian Spanish Sweedish Tamil Turkish
250 200 150 100 50 0


Confusion Matrix - Random Forest
Actual
Predicted
Arabic 97 0 0 0 0 0 0 0 0 0 9 0 0 0 0 0 0
Danish 0 60 0 1 2 1 0 0 0 0 4 0 0 1 4 0 0
Dutch 0 0 101 0 0 2 0 0 1 0 2 0 0 5 0 0 0
English 0 0 0 285 0 0 0 0 1 0 4 0 0 0 1 0 0
French 0 0 1 1 207 0 0 0 0 0 7 1 0 2 0 0 0
German 0 0 0 1 1 86 0 0 1 0 4 0 0 0 0 0 0
Greek 0 0 0 1 0 0 62 0 0 0 5 0 0 0 0 0 0
Hindi 0 0 0 0 0 0 0 10 0 0 0 0 0 0 0 0 0
Italian 0 0 0 1 3 0 0 0 132 0 6 1 0 2 0 0 0
Kannada 0 0 0 0 0 0 0 0 0 64 2 0 0 0 0 0 0
Malayalam 0 0 0 1 0 0 0 0 0 0 120 0 0 0 0 0 0
Portugeese 0 0 0 2 1 0 0 0 0 0 4 136 0 1 0 0 0
Russian 0 0 0 0 0 0 0 0 0 0 9 0 127 0 0 0 0
Spanish 0 0 0 0 2 0 0 0 1 0 6 3 0 148 0 0 0
Sweedish 0 5 0 3 0 1 0 0 0 0 1 0 0 0 123 0 0
Tamil 0 0 0 0 0 0 0 0 0 0 6 0 0 0 0 81 0
Turkish 0 0 0 0 2 0 0 0 0 0 15 1 0 2 1 0 84
Arabic Danish Dutch English French German Greek Hindi Italian Kannada Malayalam Portugeese Russian Spanish Sweedish Tamil Turkish
250
200
150
100
50
0


Confusion Matrix - Support Vector Machine
Actual
Predicted
Arabic 97 0 0 0 0 0 0 0 0 0 9 0 0 0 0 0 0
Danish 0 65 0 0 1 0 0 0 0 0 4 0 0 0 3 0 0
Dutch 0 0 103 1 0 2 0 0 0 0 3 0 0 2 0 0 0
English 0 0 0 286 0 0 0 0 0 0 4 0 0 1 0 0 0
French 0 0 0 0 208 0 0 0 0 0 10 0 0 1 0 0 0
German 0 0 0 1 0 86 0 0 0 0 5 0 0 0 1 0 0
Greek 0 0 0 1 0 0 61 0 0 0 6 0 0 0 0 0 0
Hindi 0 0 0 0 0 0 0 10 0 0 0 0 0 0 0 0 0
Italian 0 0 0 0 1 0 0 0 134 0 6 1 0 3 0 0 0
Kannada 0 0 0 0 0 0 0 0 0 63 2 0 0 0 0 0 1
Malayalam 0 0 0 1 0 0 0 0 0 0 120 0 0 0 0 0 0
Portugeese 0 0 0 0 0 0 0 0 0 0 4 137 0 3 0 0 0
Russian 0 0 0 0 0 0 0 0 0 0 8 0 128 0 0 0 0
Spanish 0 0 0 0 0 0 0 0 1 0 6 2 0 151 0 0 0
Sweedish 0 2 0 2 0 1 0 0 0 0 3 0 0 0 125 0 0
Tamil 0 0 0 0 0 0 0 0 0 0 5 0 0 0 0 82 0
Turkish 0 0 0 0 1 0 0 0 0 0 13 0 0 2 0 0 89
Arabic Danish Dutch English French German Greek Hindi Italian Kannada Malayalam Portugeese Russian Spanish Sweedish Tamil Turkish
250
200
150
100
50
0

Conclusion

Based on the accuracy and F1-score, [Best-performing Model] is recommended for multilingual text classification in this case. The model's ability to generalize across languages while maintaining high accuracy makes it suitable for language detection tasks.

Future Improvements

1. **Hyperparameter Tuning**: Perform fine-tuning to optimize model parameters further.
2. **Deep Learning Models**: Experiment with advanced NLP models like BERT or LSTM-based models to improve language detection accuracy.
3. **Data Augmentation**: Add more language samples, especially for underrepresented languages, to enhance model performance across all languages.

Appendix

- **Code:** Please refer to the attached notebook(`Multilingual_text_classification.ipynb`) for detailed code execution.
- **Dataset**: `Language Detection.csv`
- **Sample Results**: The `Samples` file may contain sample outputs or configurations.