Instructions: Install the requirements using **pip install -r requirements.txt** . Run the **main.py** or the **Language Detection.ipynb** notebook to train and test the models

Models	Macro F1	Micro F1
Multinomial Naive Bayes	0.9862	0.9830
Random Forest	0.9117	0.9211
Logistic Regression	0.9564	0.9541
Support Vector Machine	0.9266	0.9352
Voting Classifier (with the above four models)	0.9591	0.9627
BiLSTM	0.9294	0.9329
BERT + Linear	0.9732	0.9840
DistilBERT + Linear	0.9900	0.9888

Hyperparameters: I used the default parameters for sklearn models. For the BiLSTM model, I used batch size 32, the number of epochs 10, and a learning rate of 0.001. For both BERT and DistilBERT-based models, I used batch size 8, number epochs 2, and learning rate 0.0001. I did not tune the hyperparameters. Tuning the hyperparameters of each of the models will result in a performance boost. Tuning takes time which is why I did not include them in the code.

Data: I have used the language detection dataset from Kaggle.

Analysis:

I wanted to see the performance of different types of models linear models like logistic regression, ensemble models like the random forest, support vector machine, and probabilistic models like multinomial naive bayes, BiLSTM, language model (BERT and DistilBERT) based models. From the result, it is clear that the DistilBERT-based model is the best model for this experiment with a macro F1 score of .99. The second-best model is the multinomial naive bayes model. To my surprise, the random forest, support vector machine could not perform as I anticipated. If I could tune the hyperparameters of those models, I hope that their scores would be increased significantly. After running the code, scores for every class or language for every model will be printed. From the DistilBERT language-specific scores it is clear that the languages which don't share scripts or share very little with other languages (Arabic, Hindi Tamil etc.) get higher F1 scores. But languages like English, Spanish, German, Dutch, and Danish get lower F1 scores than those languages.