Natural Language Processing Challenge

Eloisa Salinas

PROJECT GOAL

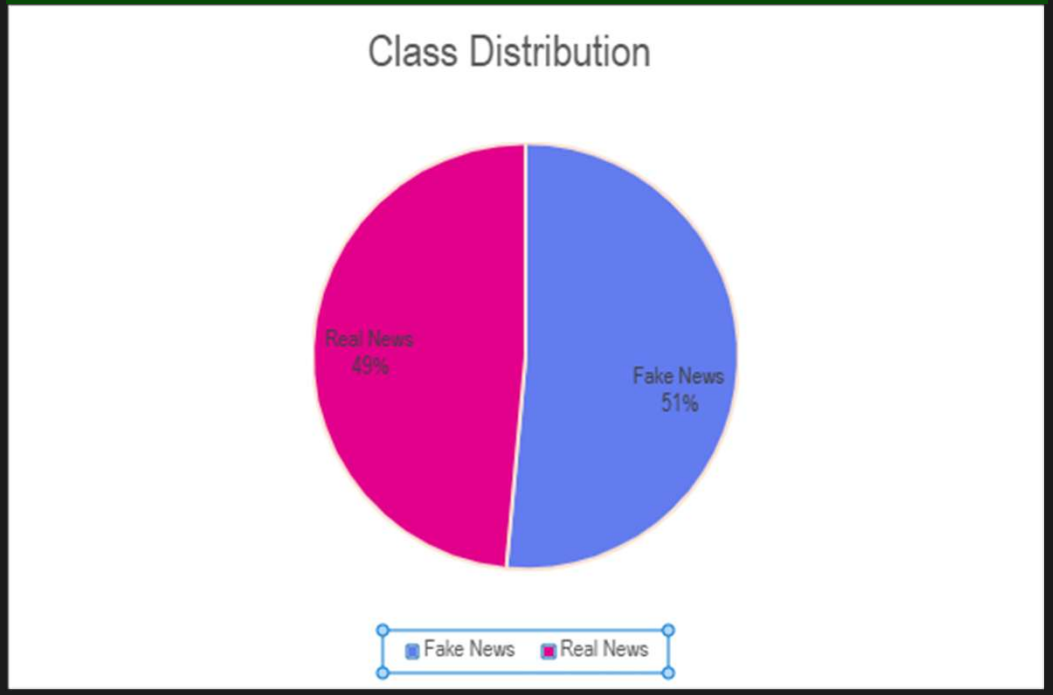
A dataset containing news headlines tagged either as fake or real news.
The goal is to build a classifier that is able to distinguish between both of them.
Use the classifier to predict the labels of a completely different data set.

Presentation overview

- Explore Data Set
- Work path definition
- Evaluation Metrics comparison
- Study Case Final Conclusions
- Key Highlights

Training Data

Class Distribution



Work path definition

Data Preprocessing

Vectorization
Tokenization
StopWord

Models Training

List of Models to train:
.Linear SVM
.Logistics Regression
.Random Forest
.XGBoots
.Multinomial NB
Hyperparameter Tuning of some of the Models above

Model Evaluation

.Comparison of Evaluation metrics for all the models (before and after tuning)

Pretrained Model

.Usage of pretrained model

Model Evaluation

.Pretrained model performance and final conclusions.

Data Prepocessing

1. Convert to lowercase
2. Remove special characters and numbers
3. - Tokenizing
4. - Remove stop words

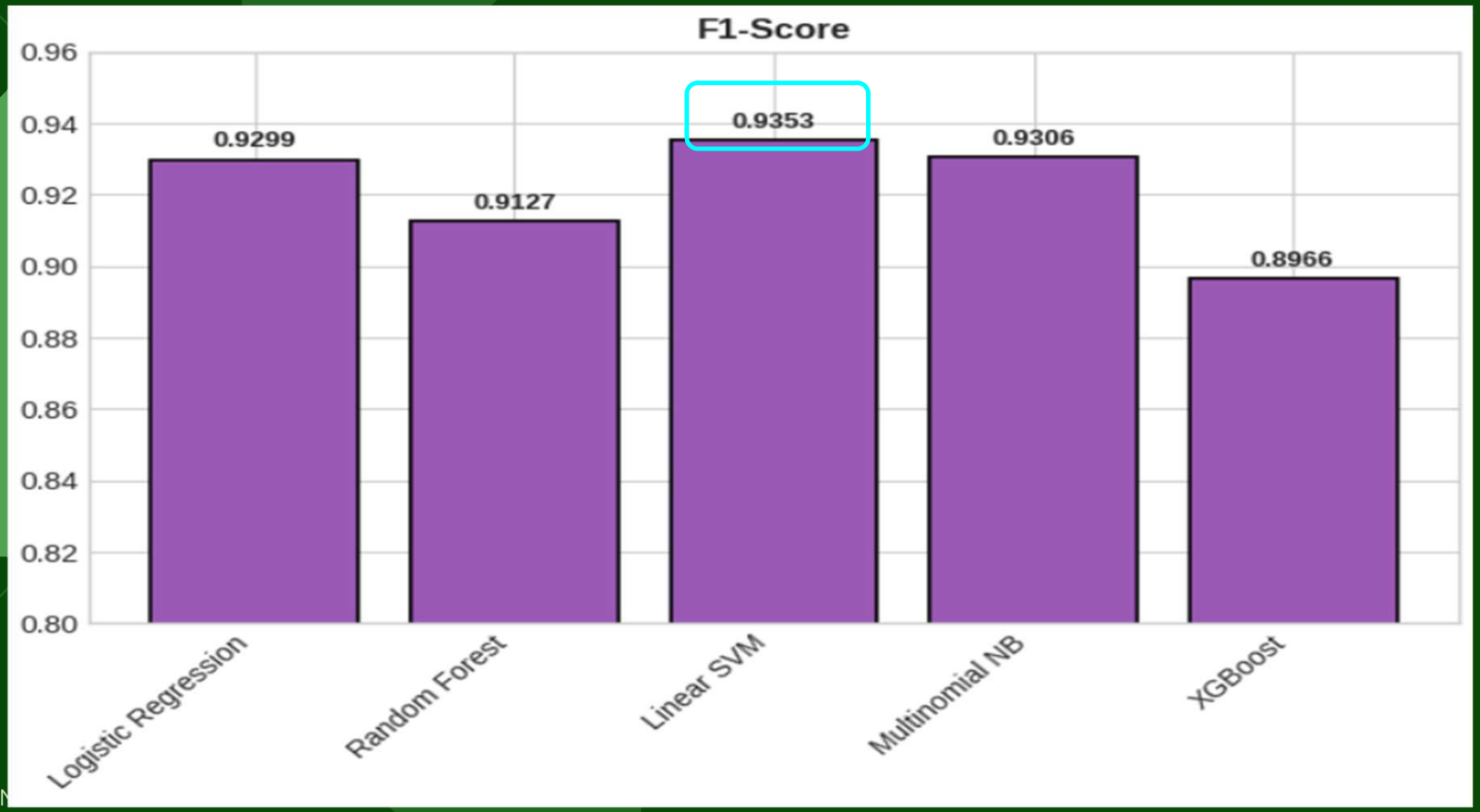
5. **Vectorization** ← Selected
   - TF-IDF (1,2-gram)
   - TF-IDF (1,2,3-gram)
   - Count Vectorizer

Models

List of models without tuning
.Linear SVM
.Logistics Regression
.Multinomial NB
.XGBoost
.Random Forest

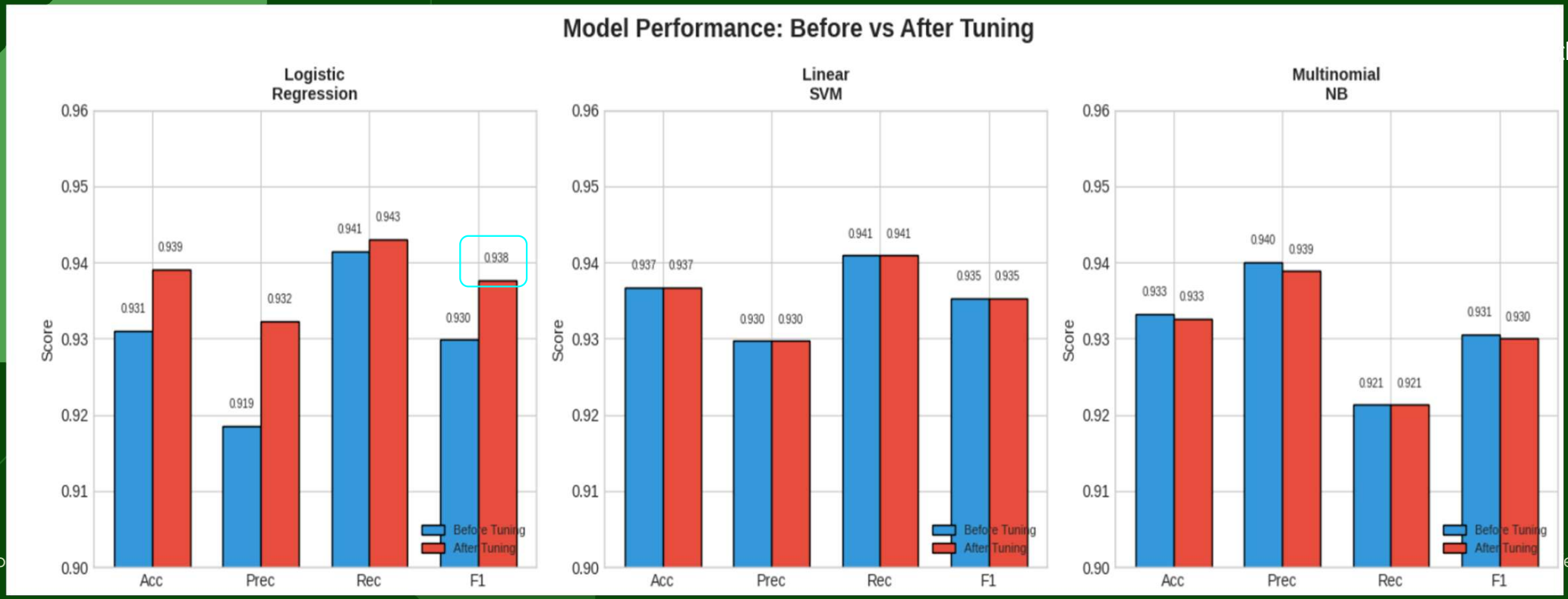
List of Models tuned:
.Linear SVM
.Logistics Regression
.Multinomial NB

Performance comparison (without tunning)



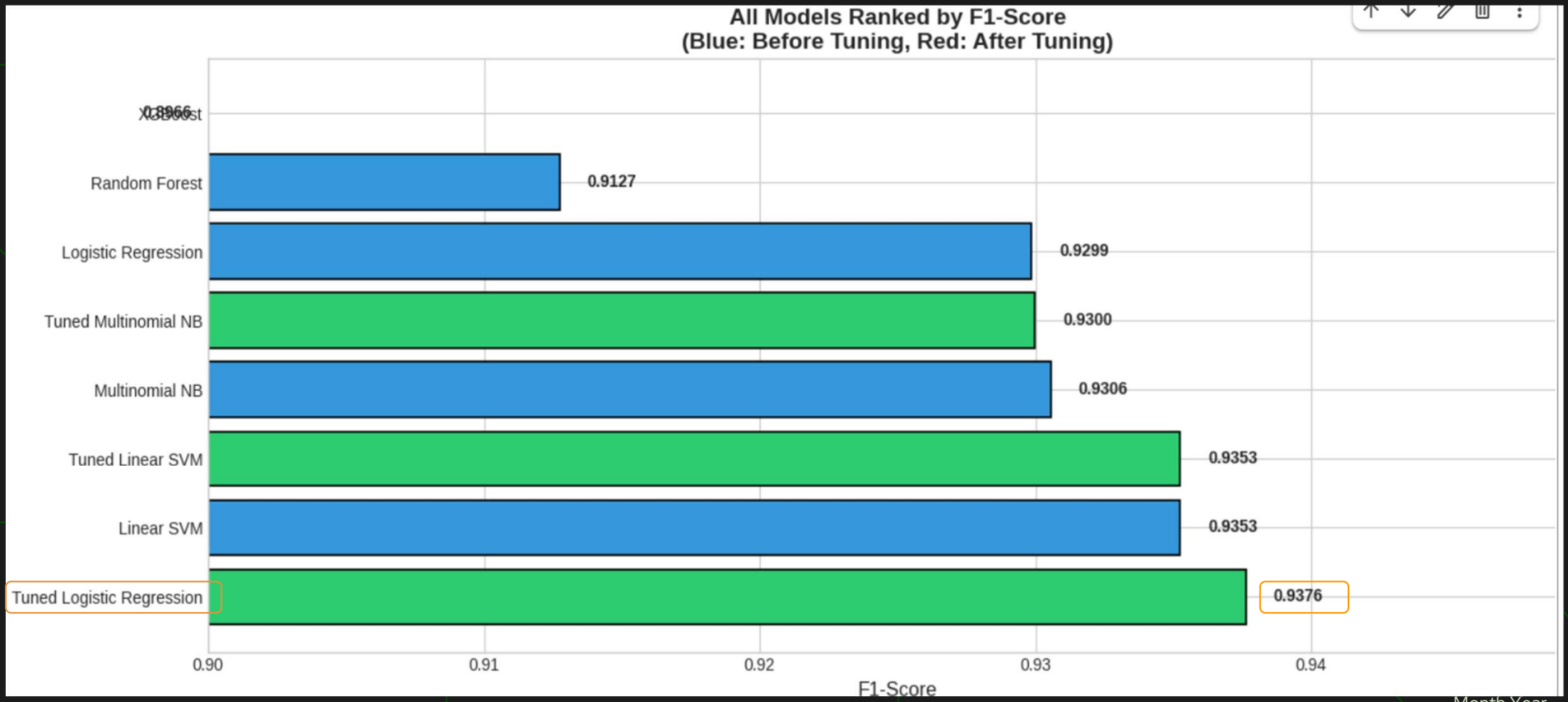
Linear SVM has the best performance with F1-Score of 0.9353

Before & after Performance comparison (available only for 3 models)



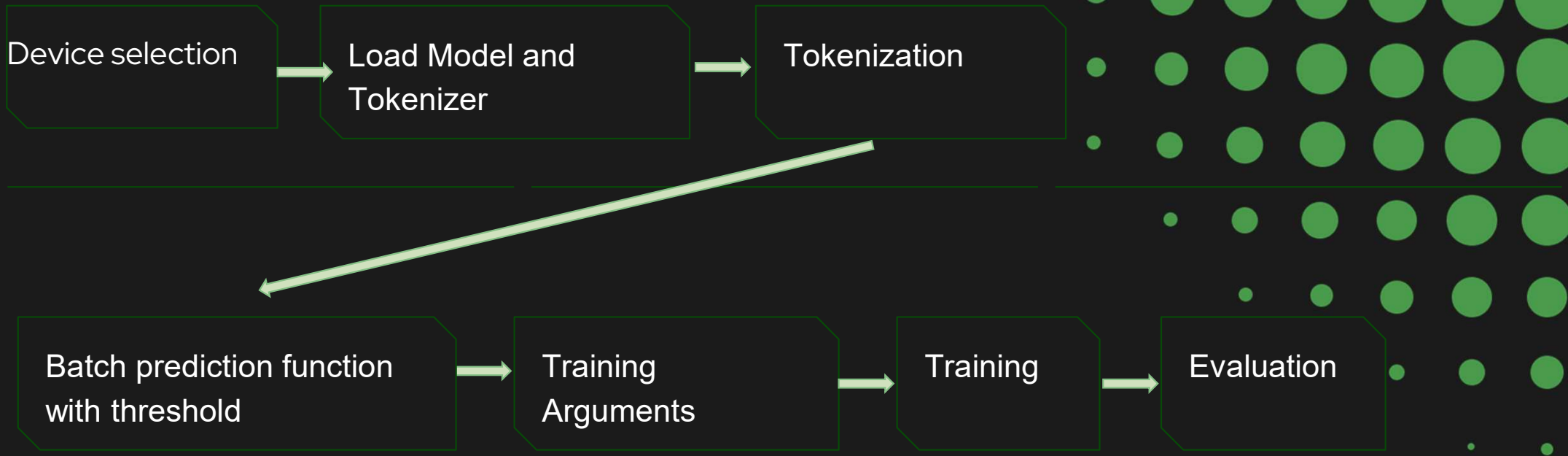
All models comparison

Tuned Logistic Regression shows the best performance of all the models



Month Year

Pretrained Model



Pretrained Model Performance

Pre-trained BERT for fake news detection

Classification report:

	precision	recall	f1-score	support
FAKE (0)	0.9887	0.9832	0.9859	3205
REAL (1)	0.9834	0.9889	0.9861	3237
accuracy			0.9860	6442
macro avg	0.9861	0.9860	0.9860	6442
weighted avg	0.9860	0.9860	0.9860	6442

Highlights

Performance (F1 Score)

Performance for most of the models went above the 90% for F1 Score, slight tuning and attention on parameters definition can improve the performance around 2%.

Pretrained Model

Performance of pretrained model significantly better than the rest of the models, model require considerable more time than the rest of the models no pretrained.

Future assessments

For future assessments, complexity, size of data set and other factors have to be considered for the model selection.