UE23CS352A: MACHINE LEARNING

Week 12: Naive Bayes Classifier

Name : CHARAN M REDDY

SRN : PES2UG23CS146

Section : C

Date : 31/10/2025

---

1. Introduction

The purpose of this lab was to build and compare different text classification models using the PubMed dataset.

We trained models using three methods :

i) Count-based Naive Bayes
ii) TF-IDF based Naive Bayes
iii) Bayes Optimal Classifier.

The main goal was to understand how each method extracts features and improves accuracy in text classification.

---

2. Methodology

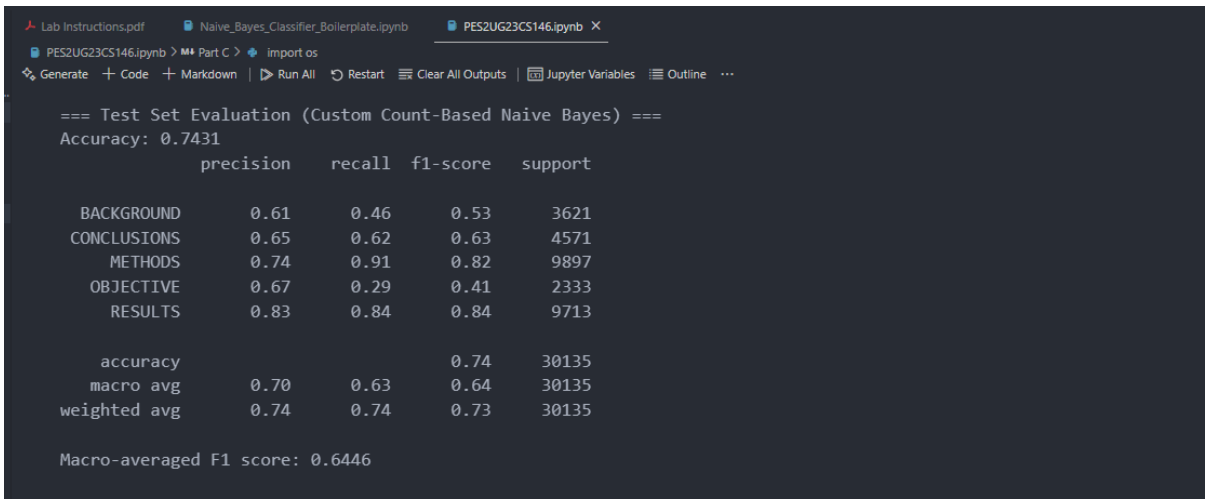
For the **Multinomial Naive Bayes (MNB)** model, we used both CountVectorizer and TF-IDF to convert text into numerical form and trained the model to classify the given text data.

For the **Bayes Optimal Classifier (BOC)**, we combined five models Naive Bayes, Logistic Regression, Random Forest, Decision Tree, and KNN , and gave each one a weight based on its performance. These models were then combined using soft voting to make the final prediction.

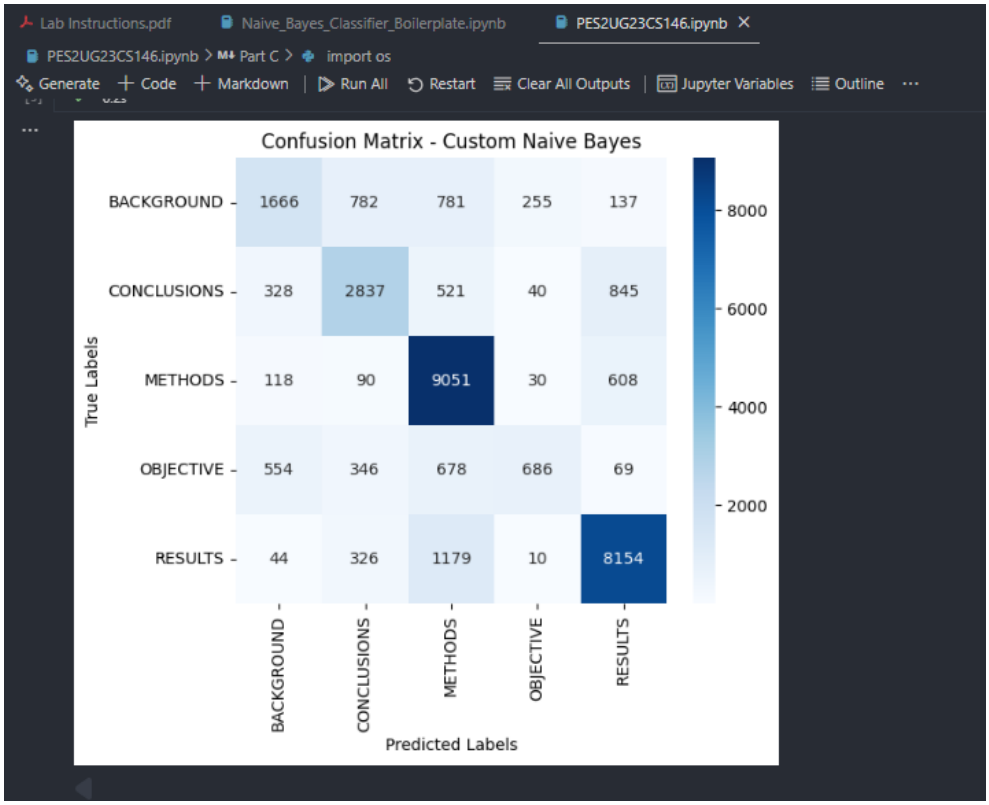
---

3. Results and Analysis

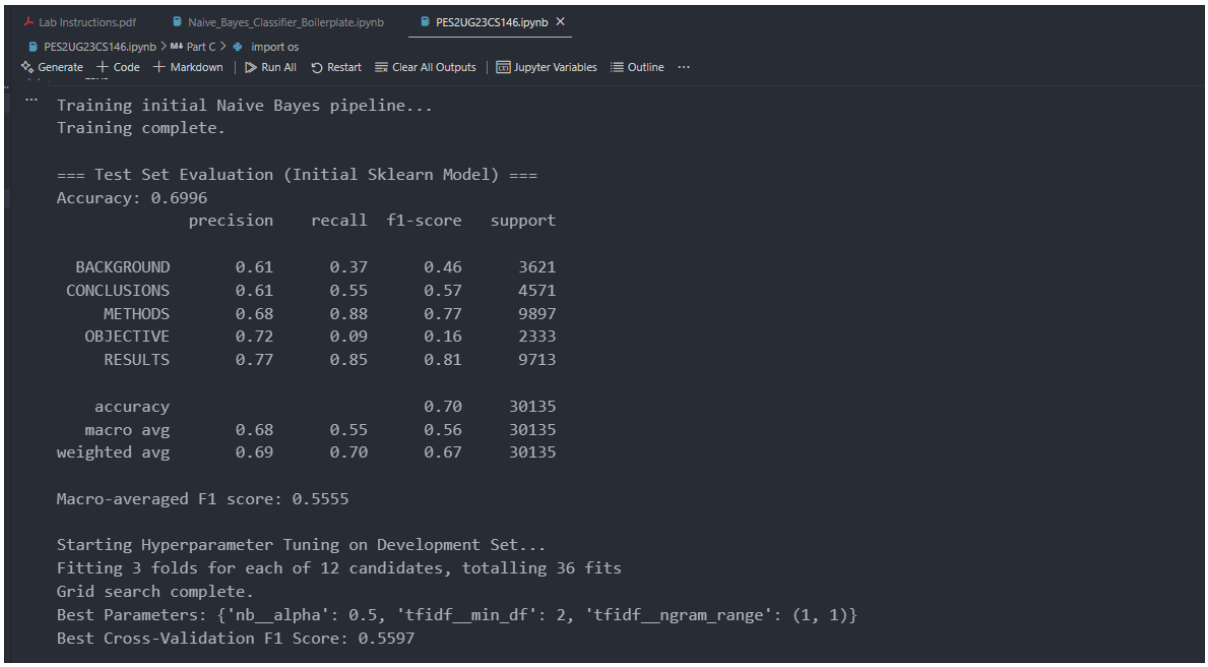
Part A: Screenshot of final test Accuracy, F1 Score



Part A: Confusion Matrix



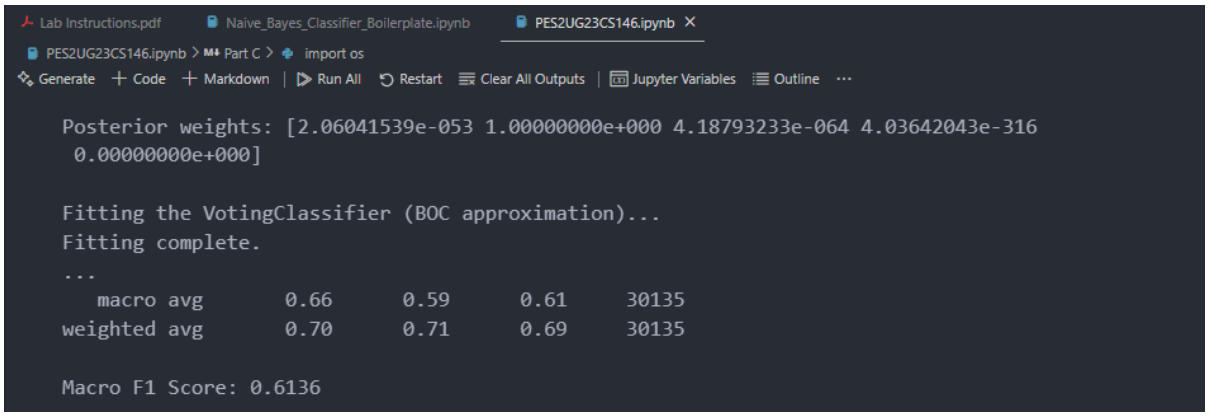
Part B: Screenshot of best hyperparameters found and their resulting F1 score.



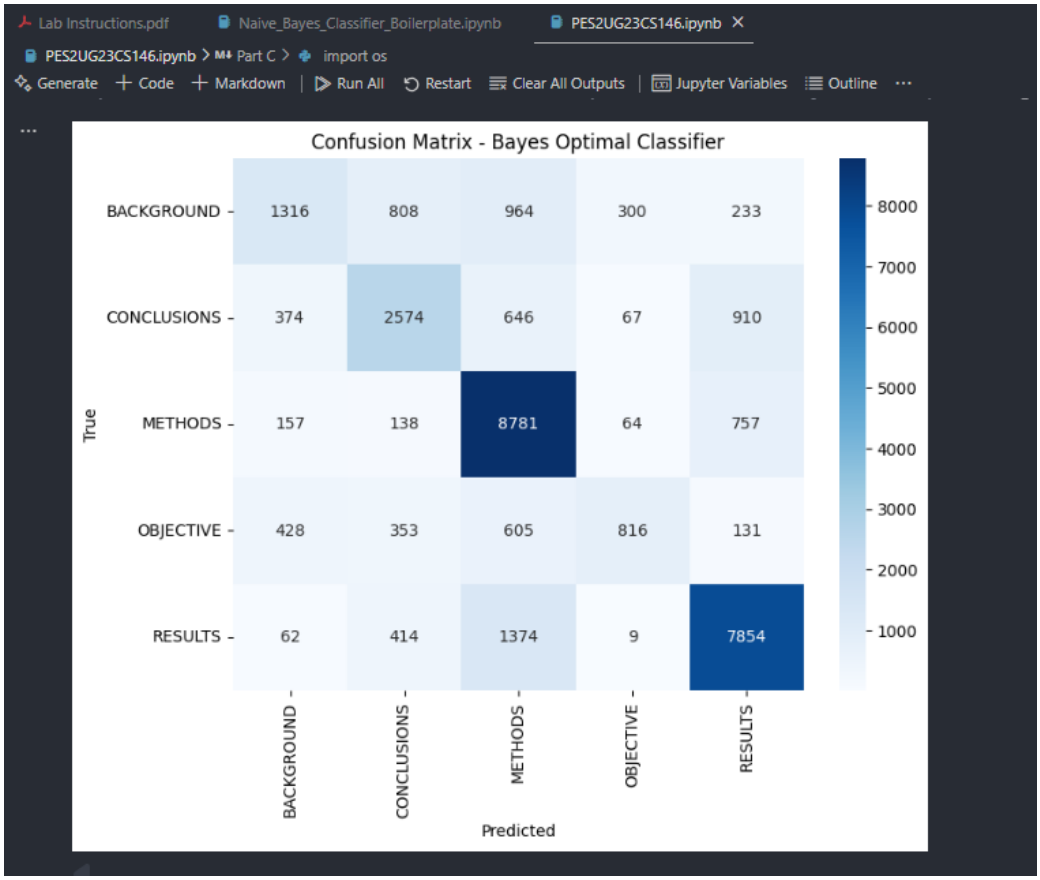
Part C: Screenshot of SRN and sample size.

```
Using dynamic sample size: 10146
Actual sampled training set size used: 10146

SRN : PES2UG23CS146


Training all base models...
Training NaiveBayes...
Training LogisticRegression...
Training RandomForest...
Training DecisionTree...
Training KNN...
```

Part C: Screenshot of BOC final Accuracy, F1 Score



Part C: Confusion Matrix.



4. Discussion

Comparison of Models:

In Part A (Count-Based Naive Bayes), the model gave the best accuracy of 74.31% and a macro F1 score of 0.64. It performed well because the count-based features captured frequent word patterns effectively.

In Part B (TF-IDF Sklearn Naive Bayes), the accuracy dropped to 69.96% with a macro F1 score of 0.55. This happened since TF-IDF reduces the effect of common words, which may have caused some loss of useful frequency information. Even after tuning, the improvement was very small.

In Part C (Bayes Optimal Classifier), the accuracy was around 71% with a macro F1 of 0.61. It performed better than the tuned TF-IDF model but slightly below the custom count-based one. The reason is that while it combined multiple models (Naive Bayes, Logistic Regression, Random Forest, etc.), most weight was given to Logistic Regression, limiting the advantage of others.