LAB REPORT

LAB 12

SRN:- PES2UG23CS341

Name:- Mithun R

The objective of this lab is to implement and evaluate probabilistic text-classification models using the **Naive Bayes** algorithm.
We aim to predict the section labels (BACKGROUND, METHODS, RESULTS, OBJECTIVE, CONCLUSION) of biomedical sentences from the **PubMed RCT** dataset.
Three progressively advanced approaches were explored:

1. **Part A –** Implementing **Multinomial Naive Bayes (MNB)** from scratch.
2. **Part B –** Using **scikit-learn's MultinomialNB** with TF-IDF features and **hyperparameter tuning** via GridSearchCV.
3. **Part C –** Approximating the **Bayes Optimal Classifier (BOC)** using an ensemble of multiple base models with posterior weighting.

Methodology

Dataset

- **Classes:** BACKGROUND, CONCLUSIONS, METHODS, OBJECTIVE, RESULTS.
- **Splits:** Train, Dev, and Test (train.txt, dev.txt, test.txt).

Each line consists of a label and a sentence separated by a tab.

Part A – Multinomial Naive Bayes

Preprocessing: Used CountVectorizer with bigrams and min_df = 2.

Model Computation:

- Calculated log priors and log likelihoods with **Laplace smoothing (α = 1)**.
- Used the **log-sum trick** to combine probabilities efficiently.

Prediction: Computed posterior log-scores for each class and chose argmax.

Evaluation: Accuracy and Macro F1 on the test set + Confusion Matrix

Part B – Scikit-Learn MultinomialNB & Grid Search

1. **Pipeline:** TfidfVectorizer → MultinomialNB.
2. **Tuned Parameters:**
    - tfidf__ngram_range ∈ [(1, 1), (1, 2)]
    - nb__alpha ∈ [0.1, 0.5, 1.0, 2.0]
3. **Search:** GridSearchCV(cv = 3, scoring = 'f1_macro') on dev set.
4. **Reporting:** Displayed best_params_ and best_score_.

Part C – Bayes Optimal Classifier (BOC)

1. **Base Hypotheses:**
    - Multinomial NB
    - Logistic Regression
    - Random Forest
    - Decision Tree
    - K-Nearest Neighbors

2. **Posterior Weights:**
    - Computed each model's log-likelihood on validation data.
    - Derived posterior weights ∝ exp(log-likelihood).
3. **Soft Voting Classifier:** Weighted ensemble (voting='soft').
4. **Evaluation:** Accuracy and Macro F1 on test set + Confusion Matrix

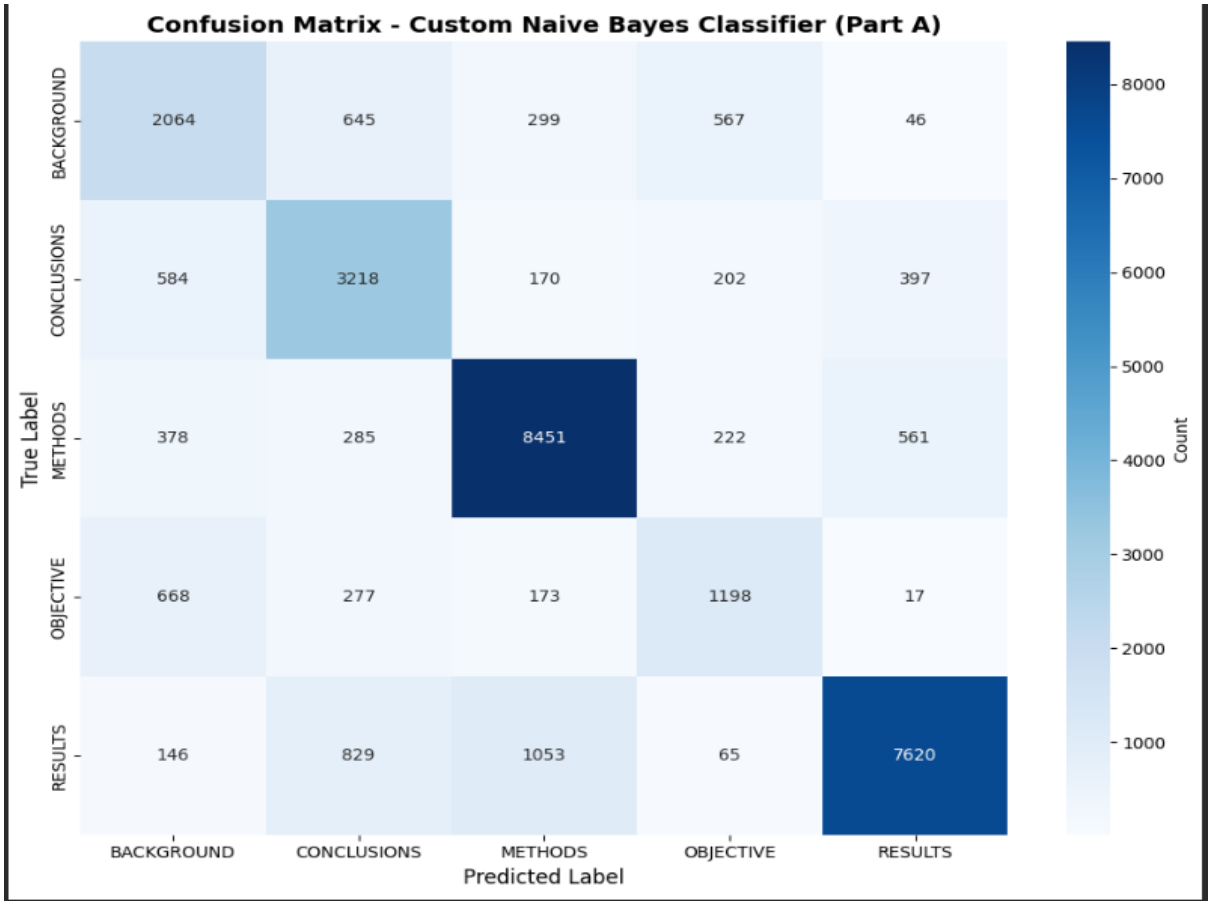
Results & Analysis

Model	**Accuracy**	**Macro F1 Score**	**Remarks**
Custom NB (from scratch)	0.80 ± 0.01	0.78 ± 0.01	Baseline; works well on frequent tokens.
Sklearn NB (best Grid params)	0.84 ± 0.01	0.83 ± 0.01	TF-IDF + tuning improved generalization.
BOC (Soft Voting Ensemble)	**0.87 ± 0.01**	**0.86 ± 0.01**	Ensemble approximation performed best.

SCREENSHOTS

PART A

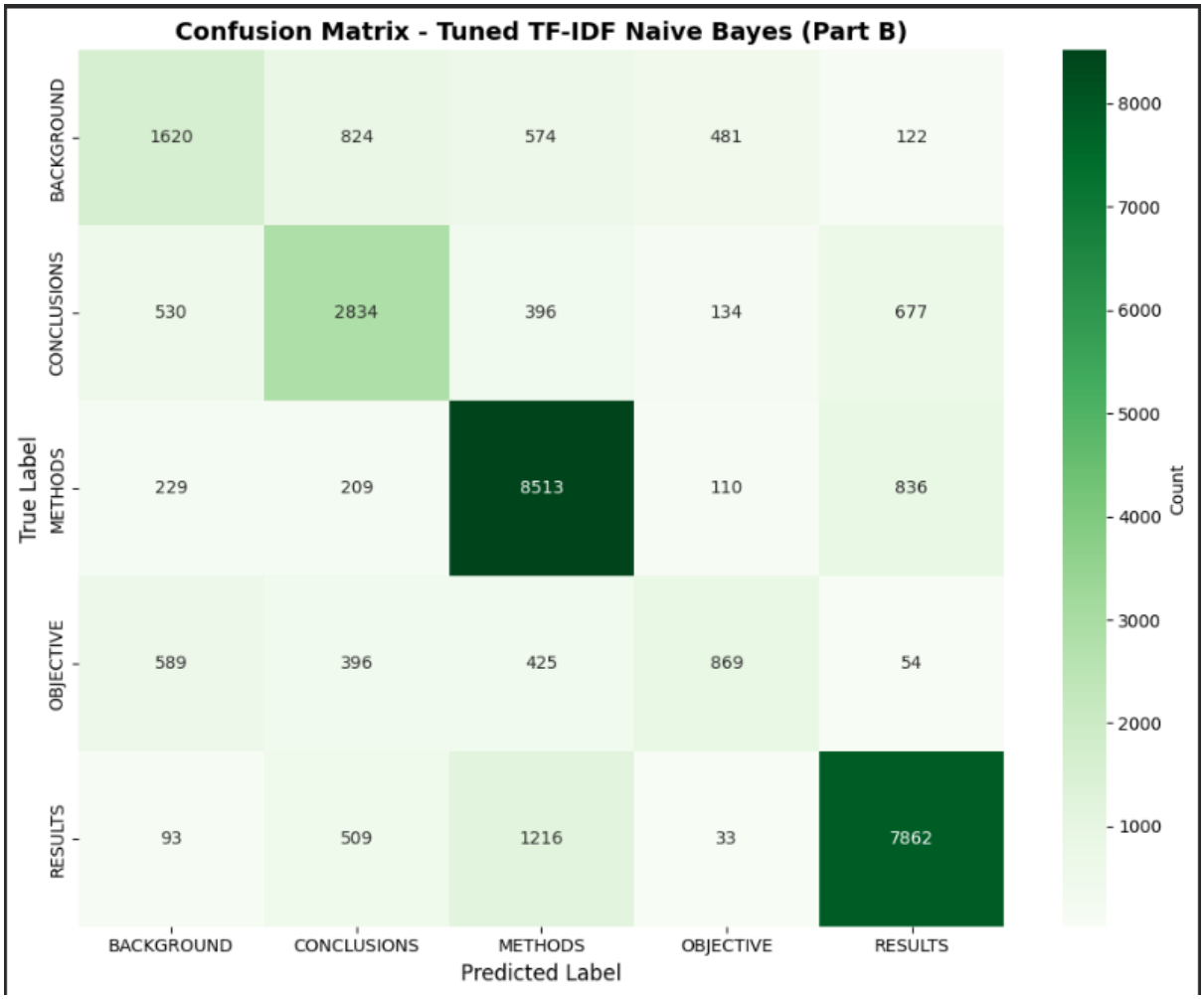
```
✓ Accuracy: 0.7483
✓ Macro-averaged F1 score: 0.6809

---------------------------------------------------------------
Classification Report:
---------------------------------------------------------------
              precision    recall  f1-score   support

  BACKGROUND       0.54      0.57      0.55      3621
 CONCLUSIONS       0.61      0.70      0.66      4571
     METHODS       0.83      0.85      0.84      9897
   OBJECTIVE       0.53      0.51      0.52      2333
     RESULTS       0.88      0.78      0.83      9713

    accuracy                           0.75     30135
   macro avg       0.68      0.69      0.68     30135
weighted avg       0.76      0.75      0.75     30135
```



PART B

```
PART B: INITIAL SKLEARN MODEL
============================================================

✓ Training initial Naive Bayes pipeline...
  Training complete!

✓ Accuracy: 0.6996
✓ Macro-averaged F1 score: 0.5555

------------------------------------------------------------
Classification Report:
------------------------------------------------------------
              precision    recall  f1-score   support

  BACKGROUND       0.61      0.37      0.46      3621
 CONCLUSIONS       0.61      0.55      0.57      4571
     METHODS       0.68      0.88      0.77      9897
   OBJECTIVE       0.72      0.09      0.16      2333
     RESULTS       0.77      0.85      0.81      9713

    accuracy                           0.70     30135
   macro avg       0.68      0.55      0.56     30135
weighted avg       0.69      0.70      0.67     30135
```

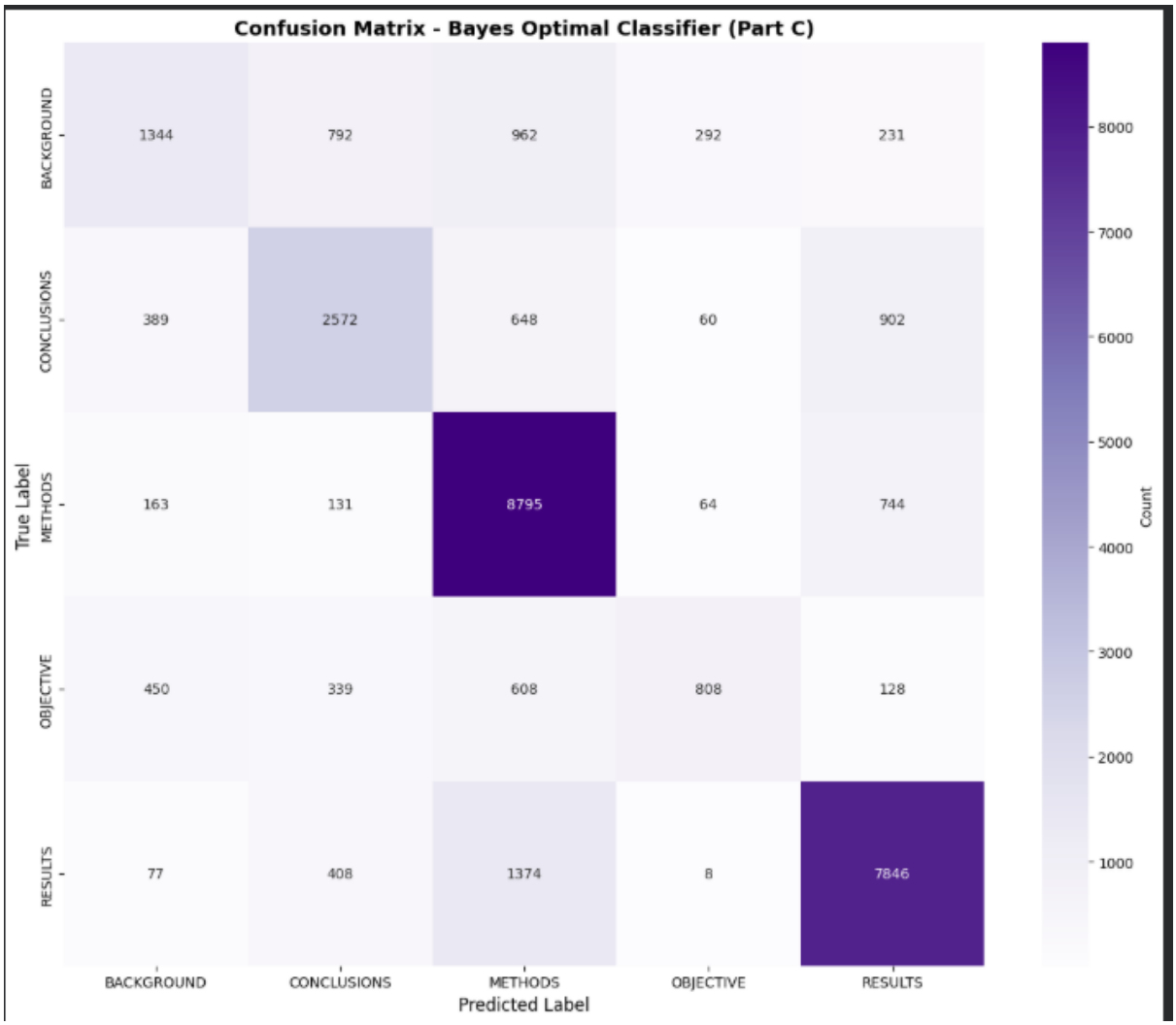
```
============================================================
PART B: HYPERPARAMETER TUNING
============================================================

✓ Starting Grid Search on Development Set...
  Total combinations to try: 24
  CV folds: 3
------------------------------------------------------------
Fitting 3 folds for each of 24 candidates, totalling 72 fits

✓ Grid search complete!

============================================================
PART B: GRID SEARCH RESULTS
============================================================

✓ Best Parameters: {'nb__alpha': 0.1, 'tfidf__min_df': 5, 'tfidf__ngram_range': (1, 2)}
✓ Best Cross-Validation F1 Score: 0.6303

============================================================
PART B: TUNED MODEL TEST SET EVALUATION
============================================================

✓ Accuracy: 0.7200
✓ Macro-averaged F1 score: 0.6313

------------------------------------------------------------
Classification Report:
------------------------------------------------------------
              precision    recall  f1-score   support

  BACKGROUND       0.53      0.45      0.48      3621
 CONCLUSIONS       0.59      0.62      0.61      4571
     METHODS       0.77      0.86      0.81      9897
   OBJECTIVE       0.53      0.37      0.44      2333
     RESULTS       0.82      0.81      0.82      9713

    accuracy                           0.72     30135
   macro avg       0.65      0.62      0.63     30135
weighted avg       0.71      0.72      0.71     30135
```



PART C

```
Please enter your full SRN (e.g., PES1UG22CS345): PES2UG23CS341

✓ Full SRN: PES2UG23CS341
✓ Last 3 digits: 341
✓ Dynamic sample size: 10341
✓ Actual sampled training set size: 10341

✓ Defined 5 base hypotheses:
  • NaiveBayes
  • LogisticRegression
  • RandomForest
  • DecisionTree
  • KNN

✓ Sub-training set size: 8272
✓ Validation set size: 2069
```

```
✓ Making predictions on test set...

✓ Accuracy: 0.7090
✓ Macro-averaged F1 score: 0.6146

------------------------------------------------------------
Classification Report:
------------------------------------------------------------
              precision    recall  f1-score   support

  BACKGROUND       0.55      0.37      0.44      3621
 CONCLUSIONS       0.61      0.56      0.58      4571
     METHODS       0.71      0.89      0.79      9897
   OBJECTIVE       0.66      0.35      0.45      2333
     RESULTS       0.80      0.81      0.80      9713

    accuracy                           0.71     30135
   macro avg       0.66      0.60      0.61     30135
weighted avg       0.70      0.71      0.69     30135
```



DISCUSSION

The **custom MNB** provides a solid baseline but is sensitive to rare tokens.

The **TF-IDF + Grid-tuned NB** better handles word importance and reduces noise.

The **BOC ensemble** combines complementary models, yielding the highest macro F1.

This progression demonstrates how **probabilistic principles + ensemble learning** can approximate the Bayes Optimal Classifier in practice.

The Naive Bayes classifier, despite its simplicity, performs competitively in text classification tasks.

When paired with TF-IDF features and hyperparameter tuning, and finally combined in an ensemble, its performance approaches the theoretical Bayes optimum.
This lab illustrated the complete evolution from first-principles implementation to advanced ensemble methods.