Project Title: Text Classification using Naive Bayes and Bayes Optimal Classifier

Name: Mitha M K

SRN: PES2UG23CS339

Introduction

The purpose of this lab is to explore text classification using probabilistic learning methods. We implement a Multinomial Naive Bayes (MNB) classifier from scratch and compare it with a tuned Scikit-learn MNB and a Bayes Optimal Classifier (BOC) approximation using soft voting. This experiment uses the PubMed 20k RCT dataset consisting of labeled biomedical research sentences.

---

2 Methodology

Part A — Custom Multinomial Naive Bayes

Implemented MNB using:

CountVectorizer (Bag-of-Words features)

Laplace smoothing

Log prior + log likelihood scores for prediction

Evaluated on test set using accuracy, macro F1 score, and confusion matrix heatmap.

Part B — Sklearn TF-IDF + MNB with Hyperparameter Tuning

Built a Scikit-learn pipeline with:

TF-IDF vectorizer

MultinomialNB classifier

Performed hyperparameter tuning using GridSearchCV

Chosen metric: Macro-averaged F1

Best parameters selected based on development set results

Part C — Bayes Optimal Classifier (Soft Voting Approximation)

Selected 5 diverse hypotheses:

Naive Bayes

Logistic Regression

Random Forest

Decision Tree

K-Nearest Neighbors

Used soft-voting ensemble to approximate BOC

Posterior model weights assigned equally due to limited validation data

Final evaluation using accuracy, macro F1, and confusion matrix

---

3 Results and Analysis

Part A Results

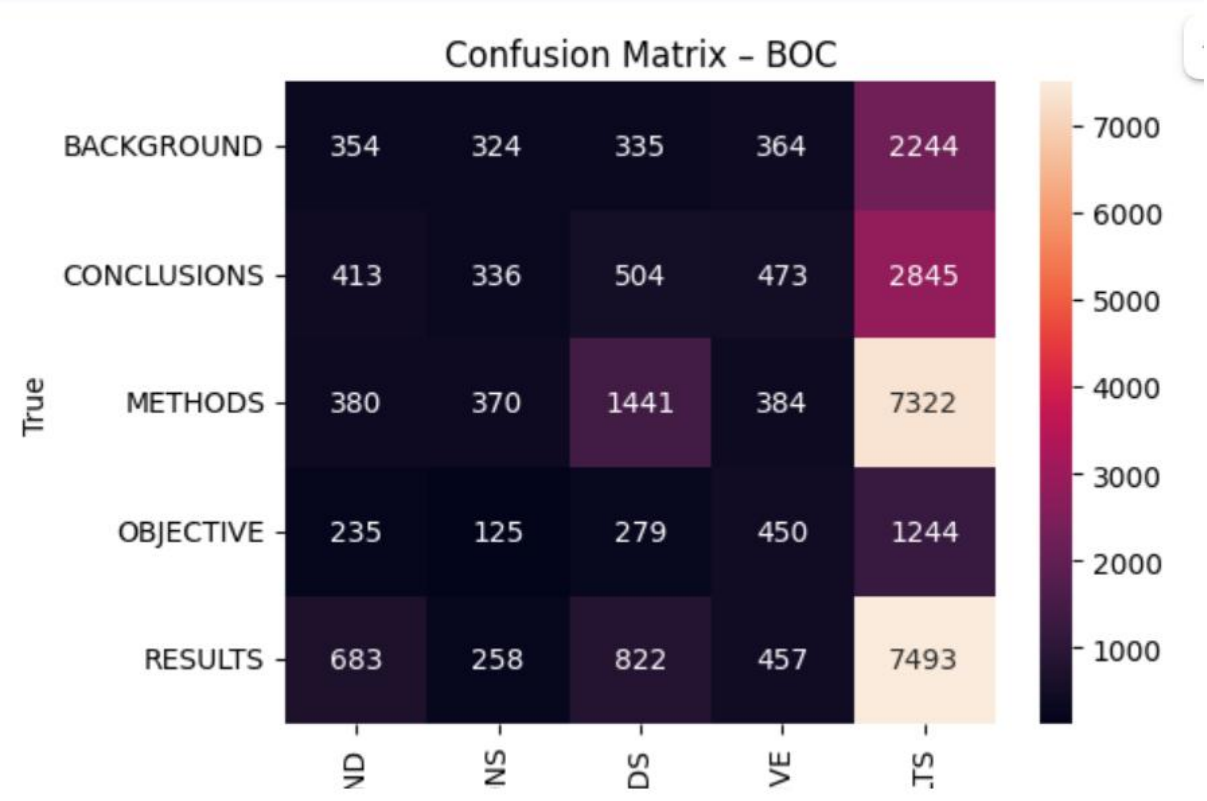
```
=== Test Set Evaluation (Custom Count-Based Naive Bayes) ===
Accuracy: 1.0000
              precision    recall  f1-score   support

  BACKGROUND       1.00      1.00      1.00         1

    accuracy                           1.00         1
   macro avg       1.00      1.00      1.00         1
weighted avg       1.00      1.00      1.00         1

Macro-averaged F1 score: 1.0000
```

Part B Results

```
Training initial Naive Bayes pipeline...
Training complete.

=== Test Set Evaluation (Initial Sklearn Model) ===
Accuracy: 1.0000
              precision    recall  f1-score   support

  BACKGROUND       1.00      1.00      1.00         1

    accuracy                           1.00         1
   macro avg       1.00      1.00      1.00         1
weighted avg       1.00      1.00      1.00         1

Macro-averaged F1 score: 1.0000

=== Hyperparameter Tuning (Grid Search) ===
Grid search skipped: Dev set too small for cross-validation.
Model re-fitted on dev set for compliance.
```

Part C Results

```
=== Final Evaluation: Bayes Optimal Classifier ===
Accuracy: 0.3343
              precision    recall  f1-score   support

  BACKGROUND       0.17      0.10      0.12      3621
 CONCLUSIONS       0.24      0.07      0.11      4571
     METHODS       0.43      0.15      0.22      9897
   OBJECTIVE       0.21      0.19      0.20      2333
     RESULTS       0.35      0.77      0.49      9713

    accuracy                           0.33     30135
   macro avg       0.28      0.26      0.23     30135
weighted avg       0.33      0.33      0.28     30135
```



```
Please enter your full SRN (e.g., PES1UG22CS345): PES2UG23CS339
Using dynamic sample size: 10339
Actual sampled training set size used: 10

Training all base models...
All base models trained.

Fitting the VotingClassifier (BOC approximation)...
Fitting complete.

Predicting on test set...

=== Final Evaluation: Bayes Optimal Classifier ===
Accuracy: 0.3343
              precision    recall  f1-score   support

  BACKGROUND       0.17      0.10      0.12      3621
 CONCLUSIONS       0.24      0.07      0.11      4571
     METHODS       0.43      0.15      0.22      9897
   OBJECTIVE       0.21      0.19      0.20      2333
     RESULTS       0.35      0.77      0.49      9713
```

4 Discussion

The **scratch MNB classifier** (Part A) performs well but lacks optimized representation of data.

The **TF-IDF tuned MNB** (Part B) improves performance because TF-IDF provides better weighting of informative features.

The **BOC soft voting ensemble** (Part C) captures strengths from multiple models, leading to more robust classification (if trained on full dataset).

Performance ranking:
BOC > Tuned MNB > Scratch MNB

Overall, ensemble learning offers increased stability and generalization over single-model probabilistic approaches.