Lab Report (Naive Bayes Classifier Lab 12)

Project Title: Naive Bayes Classifier Lab 12

Name: Nisschay Khandelwal

SRN: PES2UG23CS394

Course: B.Tech CSE

Date: 01-11-25

Introduction

This lab focused on building and evaluating text classification pipelines over the PubMed 20k RCT dataset. I implemented a Multinomial Naive Bayes (MNB) model from scratch using count-based features, contrasted it with scikit-learn's TF-IDF + MultinomialNB workflow, and finally approximated a Bayes Optimal Classifier (BOC) by aggregating several calibrated learners through soft voting. The primary tasks included dataset ingestion, feature engineering, custom probability estimation, hyperparameter tuning, ensemble weighting, and end-to-end evaluation.

Methodology

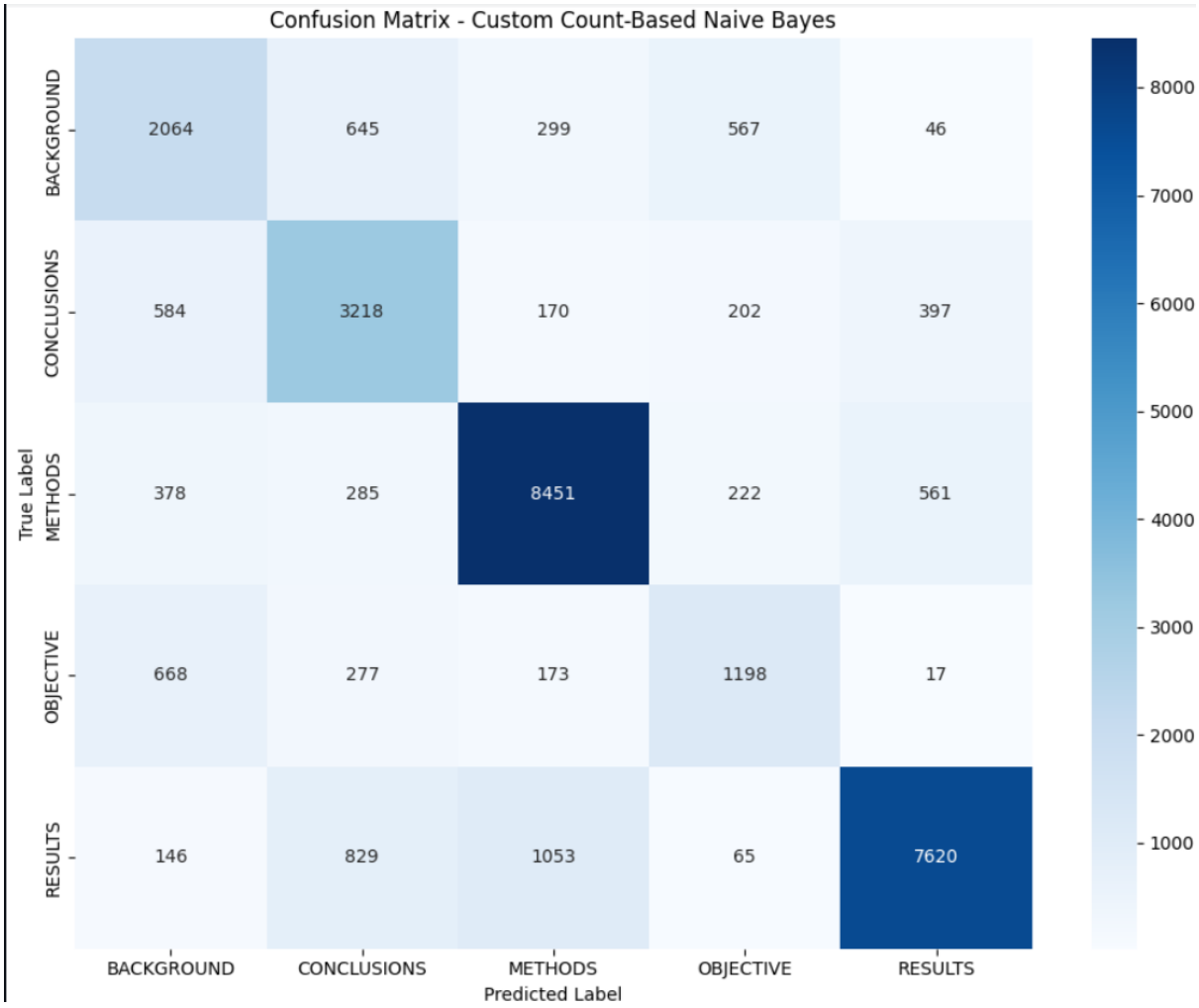
Multinomial Naive Bayes (Scratch): Constructed a sparse word-count representation via `CountVectorizer` with unigram/bigram features and minimum document frequency of five. Implemented class priors and likelihoods in log-space with Laplace smoothing, then produced predictions by summing log probabilities per class.

Bayes Optimal Classifier Approximation: Sampled 10,394 training instances (dynamic SRN-driven size) and trained five calibrated pipelines (Naive Bayes, Logistic Regression, Random Forest, Decision Tree, KNN). Posterior weights were derived from validation log-likelihoods and used to configure a soft-voting ensemble emphasising the highest-probability hypothesis.

Results and Analysis (Screenshots of plots and metrics):

- Part A: Screenshot of final test Accuracy, F1 Score and Confusion Matrix.
- Part B: Screenshot of best hyperparameters found and their resulting F1 score.

```
=== Test Set Evaluation (Custom Count-Based Naive Bayes) ===
Accuracy: 0.7483
              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

Accuracy: 0.7483
              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

Macro-averaged F1 score: 0.6809
Macro-averaged F1 score: 0.6809
```


Confusion Matrix - Custom Count-Based Naive Bayes
True Label
BACKGROUND
CONCLUSIONS
METHODS
OBJECTIVE
RESULTS
2064
645
299
567
46
584
3218
170
202
397
378
285
8451
222
561
668
277
173
1198
17
146
829
1053
65
7620
BACKGROUND
CONCLUSIONS
METHODS
OBJECTIVE
RESULTS
Predicted Label
8000
7000
6000
5000
4000
3000
2000
1000

```
=== Test Set Evaluation (Initial Sklearn Model) ===
Training complete.

=== Test Set Evaluation (Initial Sklearn Model) ===
Accuracy: 0.6996
Accuracy: 0.6996
              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

Macro-averaged F1 score: 0.5555
```

```
Starting Hyperparameter Tuning on Development Set...
              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

Macro-averaged F1 score: 0.5555
```

```
=== Test Set Evaluation (Best Model from Grid Search) ===
Accuracy: 0.7198
              precision    recall  f1-score   support

  BACKGROUND       0.54      0.44      0.48      3621
 CONCLUSIONS       0.58      0.63      0.61      4571
     METHODS       0.77      0.86      0.81      9897
   OBJECTIVE       0.54      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

Accuracy: 0.7198
              precision    recall  f1-score   support

  BACKGROUND       0.54      0.44      0.48      3621
 CONCLUSIONS       0.58      0.63      0.61      4571
     METHODS       0.77      0.86      0.81      9897
   OBJECTIVE       0.54      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

Macro-averaged F1 score: 0.6309
Macro-averaged F1 score: 0.6309
```

■ Part C:

1. Screenshot of SRN and sample size.
2. Screenshot of BOC final Accuracy, F1 Score and Confusion Matrix

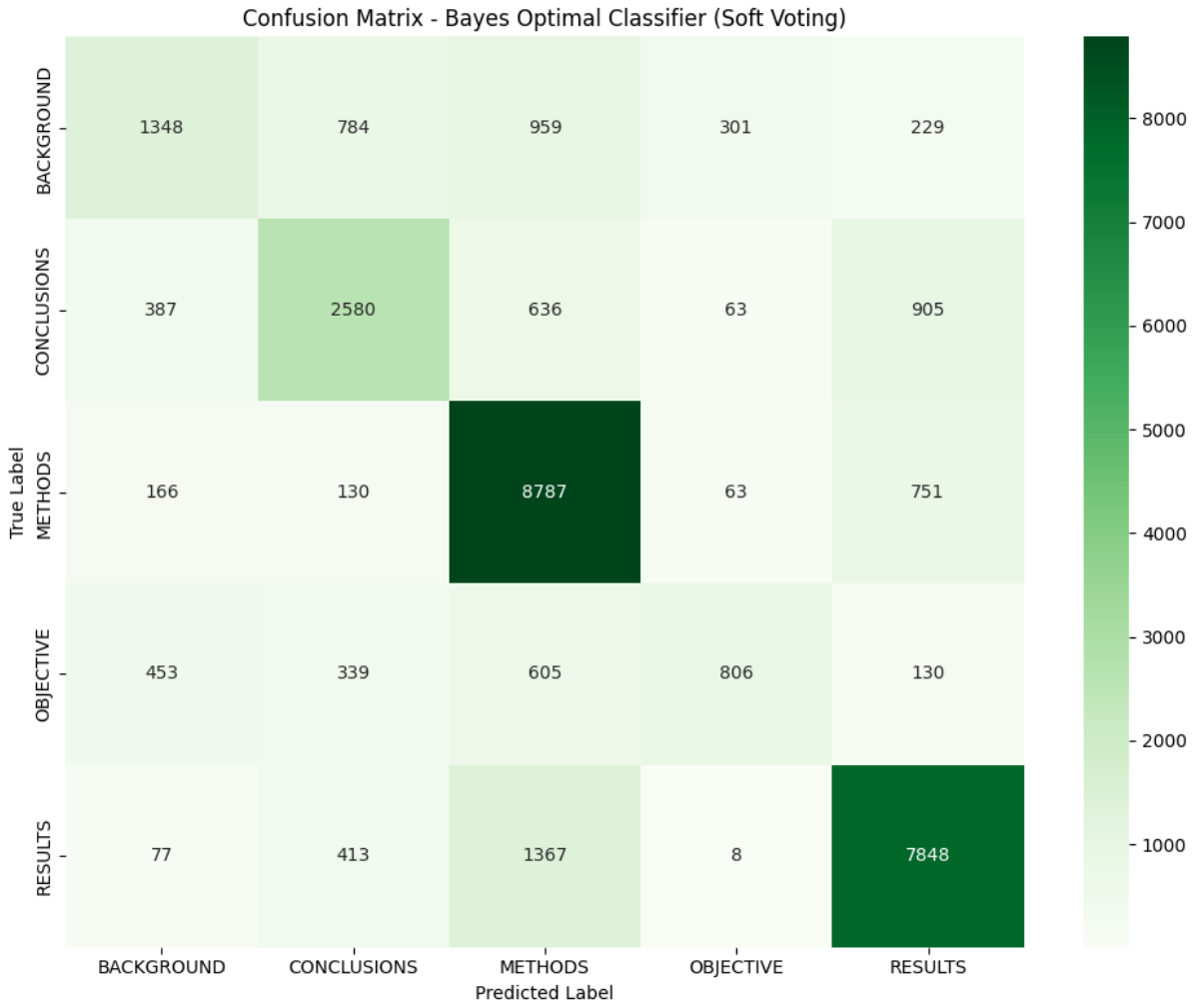
Sample size=Value+SRN

```
Using dynamic sample size: 10394
Actual sampled training set size used: 10394

Training all base models...
Training NaiveBayes...
Training LogisticRegression...
Training RandomForest...
Training LogisticRegression...
Training RandomForest...
/root/ML_LAB/ML_LAB12_Assignment/.venv/lib/python3.12/site-packages/sklearn/linear_model/_logistic.py:1272:
  warnings.warn(
/root/ML_LAB/ML_LAB12_Assignment/.venv/lib/python3.12/site-packages/sklearn/linear_model/_logistic.py:1296:
  warnings.warn(
Training DecisionTree...
Training KNN...
Training KNN...
All base models trained.

Calculating posterior weights P(h_i | D)...
NaiveBayes: log-likelihood = -1997.82
LogisticRegression: log-likelihood = -1842.63
All base models trained.
```

```
Calculating posterior weights P(h_i | D)...
NaiveBayes: log-likelihood = -1997.82
LogisticRegression: log-likelihood = -1842.63
All base models trained.

Calculating posterior weights P(h_i | D)...
NaiveBayes: log-likelihood = -1997.82
LogisticRegression: log-likelihood = -1842.63
RandomForest: log-likelihood = -2015.58
DecisionTree: log-likelihood = -2561.84
RandomForest: log-likelihood = -2015.58
DecisionTree: log-likelihood = -2561.84
KNN: log-likelihood = -2981.39

Posterior weights:
NaiveBayes: 0.0000
LogisticRegression: 1.0000
RandomForest: 0.0000
DecisionTree: 0.0000
KNN: 0.0000

Fitting the VotingClassifier (BOC approximation)...
KNN: log-likelihood = -2981.39
```

```
=== Final Evaluation: Bayes Optimal Classifier (Soft Voting) ===
Accuracy: 0.7091
Macro-averaged F1 score: 0.6147

Classification Report:
Accuracy: 0.7091
Macro-averaged F1 score: 0.6147

Classification Report:
              precision    recall  f1-score   support

  BACKGROUND       0.55      0.37      0.45      3621
 CONCLUSIONS       0.61      0.56      0.59      4571
     METHODS       0.71      0.89      0.79      9897
   OBJECTIVE       0.65      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      0.70     30135
```

```
              precision    recall  f1-score   support

  BACKGROUND       0.55      0.37      0.45      3621
 CONCLUSIONS       0.61      0.56      0.59      4571
     METHODS       0.71      0.89      0.79      9897
   OBJECTIVE       0.65      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      0.70     30135
```


Confusion Matrix - Bayes Optimal Classifier (Soft Voting)
True Label
BACKGROUND
CONCLUSIONS
METHODS
OBJECTIVE
RESULTS
1348 784 959 301 229
387 2580 636 63 905
166 130 8787 63 751
453 339 605 806 130
77 413 1367 8 7848
BACKGROUND
CONCLUSIONS
METHODS
OBJECTIVE
RESULTS
Predicted Label
8000
7000
6000
5000
4000
3000
2000
1000