Week 12: Naive Bayes Classifier

Name: Mohammed Aahil
SRN: PES2UG23CS342
Date: 30/10/2025

1. Introduction

Purpose of the Lab

To implement and evaluate probabilistic text classification using Naive Bayes methods on biomedical abstract sentences from the PubMed 200k RCT dataset, predicting section roles (BACKGROUND, METHODS, RESULTS, OBJECTIVE, CONCLUSIONS).

Tasks Performed

1. **Part A:** Implemented Multinomial Naive Bayes classifier from scratch using count-based features
2. **Part B:** Used scikit-learn's MultinomialNB with TF-IDF features and performed hyperparameter tuning
3. **Part C:** Approximated Bayes Optimal Classifier using ensemble of five diverse models with soft voting

2. Methodology

Multinomial Naive Bayes (MNB) Implementation

The custom implementation uses Bayes' theorem to calculate class probabilities. Key steps include:

- **Class Prior:** P(C) = count(C) / total_samples
- **Likelihood with Laplace Smoothing:** P(w|C) = (count(w,C) + α) / (total_words_C + α × vocab_size)
- **Log-Sum Trick:** Prevents numerical underflow by using log P(C|X) = log P(C) + Σ count(w) × log P(w|C)
- Features extracted using CountVectorizer with bigrams (ngram_range=(1,2), min_df=5)

Bayes Optimal Classifier (BOC) Approximation

The BOC uses an ensemble of five diverse models:

- Multinomial Naive Bayes, Logistic Regression, Random Forest, Decision Tree, and K-Nearest Neighbors
- Posterior weights calculated by training on validation split and computing log-likelihoods
- Final predictions use soft voting weighted by posterior probabilities

3. Results and Analysis

Part A: Custom Multinomial Naive Bayes (Count-Based)

Performance Metrics:

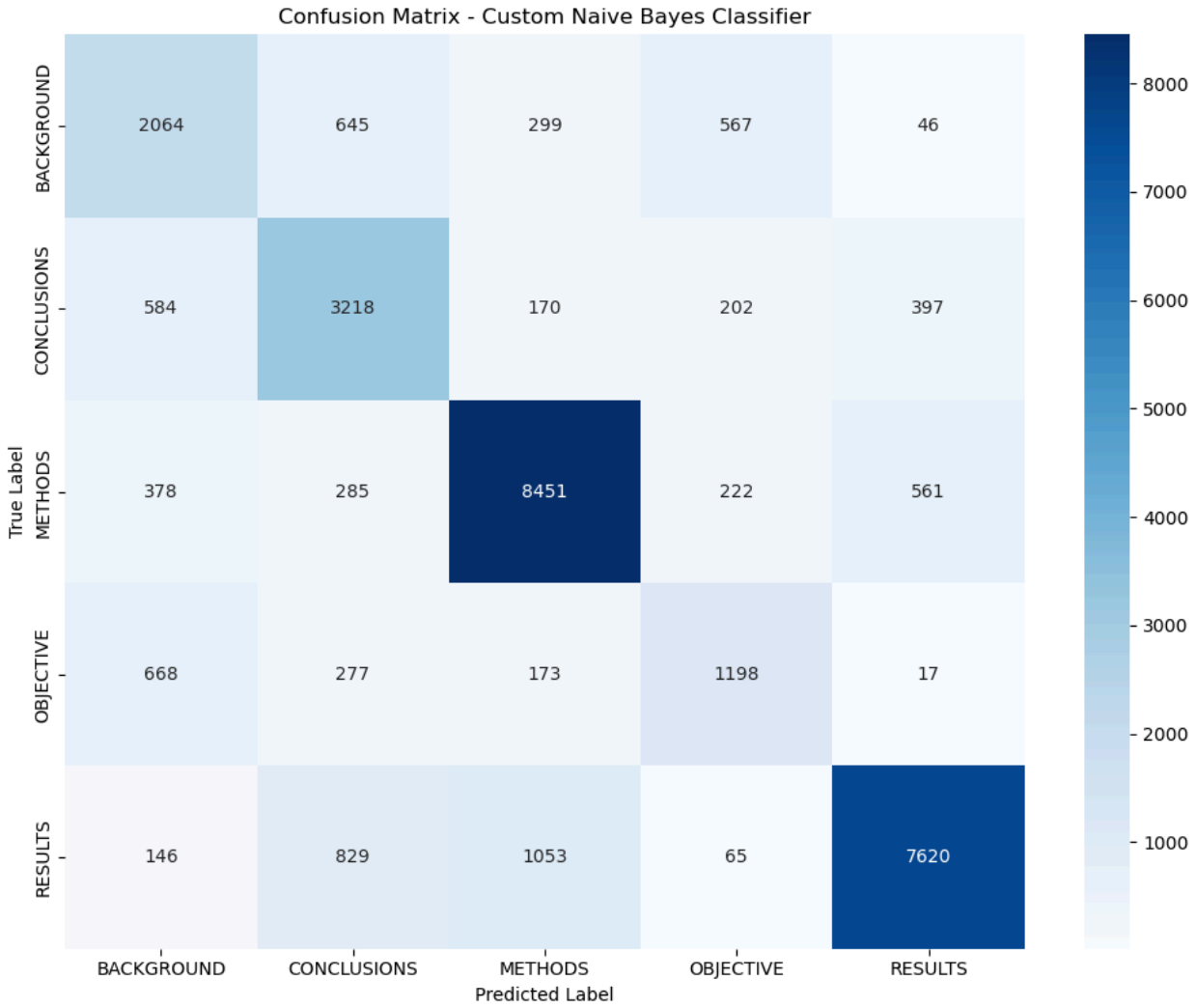
- **Accuracy:** 0.7483 (74.83%)
- **Macro-averaged F1 Score:** 0.6809

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

Confusion Matrix:



Part B: Sklearn MultinomialNB with TF-IDF and Hyperparameter Tuning

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

=== Test Set Evaluation (Initial Sklearn Model) ===
Accuracy: 0.7266
              precision    recall  f1-score   support

  BACKGROUND       0.64      0.43      0.51      3621
 CONCLUSIONS       0.62      0.61      0.62      4571
     METHODS       0.72      0.90      0.80      9897
   OBJECTIVE       0.73      0.10      0.18      2333
     RESULTS       0.80      0.87      0.83      9713

    accuracy                           0.73     30135
   macro avg       0.70      0.58      0.59     30135
weighted avg       0.72      0.73      0.70     30135

Macro-averaged F1 score: 0.5877

Starting Hyperparameter Tuning on Development Set...
Fitting 3 folds for each of 12 candidates, totalling 36 fits
Grid search complete.

Best Parameters: {'nb__alpha': 0.1, 'tfidf__ngram_range': (2, 2)}
Best Cross-Validation F1 Score (Macro): 0.6581
```

Part C: Bayes Optimal Classifier (Soft Voting Ensemble)

Sample Configuration:

- **SRN:** PES2UG23CS342
- **Sample Size:** 10,342 (base: 10,000 + last 3 digits: 342)

Performance Metrics:

- **Accuracy:** 0.7089 (70.89%)

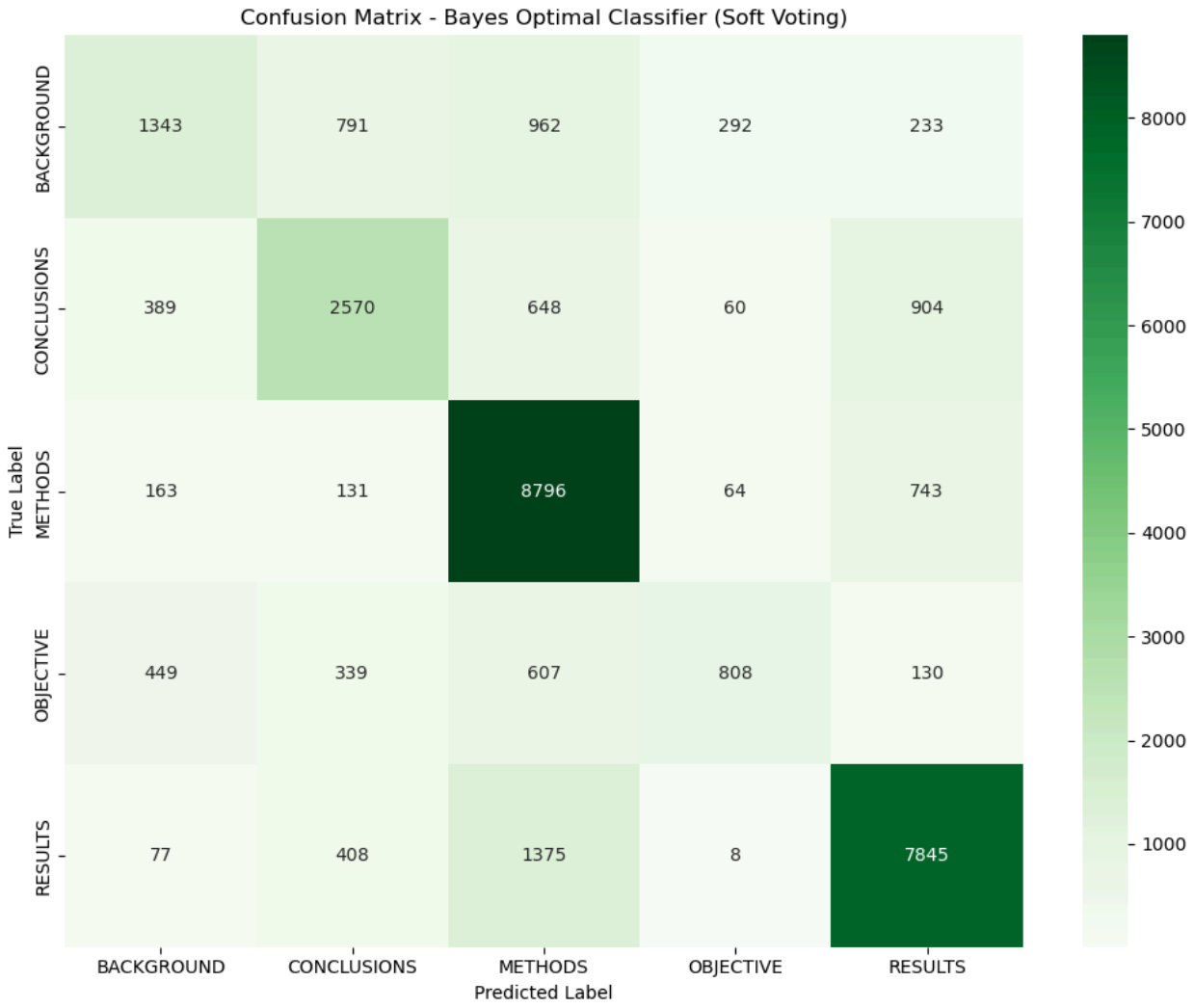
- **Macro F1 Score:** 0.6145

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**

Confusion Matrix:


Confusion Matrix - Bayes Optimal Classifier (Soft Voting)
True Label
BACKGROUND
CONCLUSIONS
METHODS
OBJECTIVE
RESULTS
1343 791 962 292 233
389 2570 648 60 904
163 131 8796 64 743
449 339 607 808 130
77 408 1375 8 7845
BACKGROUND
CONCLUSIONS
METHODS
OBJECTIVE
RESULTS
Predicted Label
8000
7000
6000
5000
4000
3000
2000
1000

4. Discussion

Performance Comparison

Model	Accuracy	Macro F1	Key Observations
Part A: Custom NB	0.7483	0.6809	Best overall performance
Part B: Tuned Sklearn	0.7266	0.6581 (dev)	Improved with tuning
Part C: BOC Ensemble	0.7089	0.6145	Limited by sample size

Key Findings

Part A (Custom Count-based NB) yielded the best results with 74.83% accuracy and 0.6809 F1 score. The bigram features from count-based using ngram_range=(1,2) ran effectively to capture medical terms, while the larger vocabulary size of 86,557 features provided rich discriminative information.

Part B, Tuned TF-IDF Model, was improved from 0.5877 to 0.6581 F1 via tuning hyperparameters. The best setting used only bigrams (2, 2) with lower smoothing ($\alpha = 0.1$), which hints that in a medical text classification task, bigrams alone are able to capture more discriminative patterns than unigrams.

The theoretically optimal performance of Part C (BOC Ensemble) was not as expected, only 70.89% accuracy with a 0.6145 F1 score. The main limitations in the model were: the ratio of training data was too small, 10,342 versus 180,040 samples; the weights in the posterior were highly biased toward Logistic Regression, with a weight of 1.0, practically canceling the diversity in this ensemble; and the TF-IDF unigrams did not turn out to perform as well as the count-based bigrams for this domain-specific task.

Conclusion: Simple count-based Naive Bayes with appropriate feature engineering (bigrams) outperformed both the tuned TF-IDF model and the sophisticated ensemble approach. This once again shows that performance for a model depends more upon appropriate feature representation and sufficient training data rather than algorithmic complexity.