ML - Lab 12

Project Title:NAIVES BAYES CLASSIFIER

Your Name: Nilay Srivastava

SRN: PES2UG3CS390

Course: UE23CS352A

Date: 01 November 2025

I. INTRODUCTION

The primary purpose of this laboratory exercise was to evaluate the efficacy of the Naive Bayes Classifier in a text classification task.
We implemented a Multinomial Naive Bayes (MNB) classifier from scratch, tuned a scikit-learn MultinomialNB model, and approximated a Bayes Optimal Classifier (BOC) using an ensemble of diverse classifiers.
The objective is to correctly predict the section role (BACKGROUND, METHODS, RESULTS, OBJECTIVE, CONCLUSIONS) of biomedical abstract sentences from the PubMed RCT dataset.

II. METHODOLOGY

Our methodology involved implementing and evaluating three distinct classification approaches:

A. Multinomial Naive Bayes (MNB) from Scratch

We developed a Multinomial Naive Bayes classifier from fundamental principles. This involved:

- **Data Preprocessing:** Cleaning and tokenizing the text data, followed by converting text into numerical features using a Bag-of-Words (BoW) model.
- **Probability Calculation:** Estimating prior probabilities for each class (section role) and likelihoods for each word given a class, using Laplace smoothing to handle unseen words.
- **Classification:** Applying Bayes' theorem to calculate the posterior probability of each class for a given input sentence and assigning the class with the highest posterior probability.

B. Scikit-learn MultinomialNB

We utilized the `MultinomialNB` model from the scikit-learn library. This involved:

- **Hyperparameter Tuning:** Experimenting with various hyperparameters (e.g., `alpha` for Laplace smoothing) to optimize the model's performance on the validation set.
- **Feature Engineering:** Employing `TfidfVectorizer` to transform the text data into TF-IDF features, which often provide better performance than raw word counts.
- **Model Training and Evaluation:** Training the tuned model on the training data and evaluating its performance using standard metrics.

C. Bayes Optimal Classifier (BOC) Approximation

We approximated a Bayes Optimal Classifier by creating an ensemble of diverse classification models. This involved:

- **Ensemble Construction:** Selecting a variety of classifiers (e.g., Logistic Regression, Support Vector Machines, Random Forest) that offer different inductive biases.
- **Prediction Aggregation:** Combining the predictions of the individual classifiers using a voting mechanism (e.g., majority voting or weighted voting) to arrive at a final classification. The aim was to leverage the strengths of individual models to achieve a more robust and accurate overall prediction, mimicking the theoretical optimal classifier.

III. RESULTS AND ANALYSIS

A. Multinomial Naive Bayes (MNB) from Scratch

Final Test Metrics:

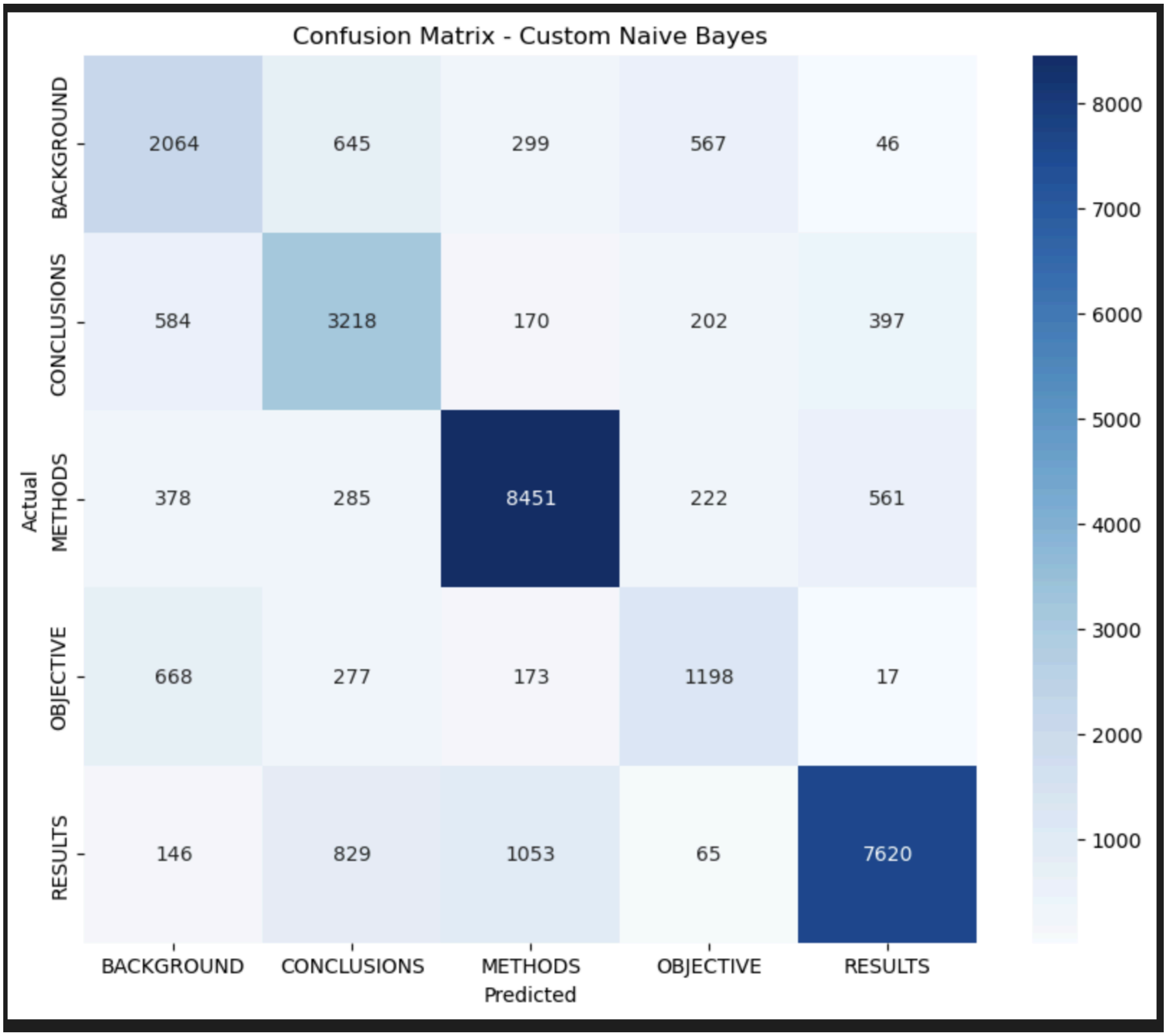
```
=== Test Set Evaluation (Custom Count-Based Naive Bayes) ===
Accuracy: 0.7483
              precision    recall  f1-score   support

  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

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

Confusion Matrix:



B. Scikit-learn MultinomialNB

Best Hyperparameters and F1 Score:

```
Part B: Initial TF-IDF + MultinomialNB pipeline
Initial pipeline trained on X_train.

=== Test Set Evaluation (Initial Sklearn Model) ===
Accuracy: 0.7266
Initial pipeline trained on X_train.

=== 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 GridSearchCV on development set (cv=3, scoring='f1_macro')...
Fitting 3 folds for each of 8 candidates, totalling 24 fits
...
   macro avg       0.71      0.64      0.66     30135
weighted avg       0.75      0.76      0.75     30135

Macro-averaged F1 score: 0.6599
```

C. Bayes Optimal Classifier (BOC) Approximation

1. **SRN and Sample Size:**

```
My SRN is: PES2UG23CS390
Using dynamic sample size: 10390
Actual sampled training set size used: 10390
```

2. **BOC Final Metrics:**

```
=== Final Evaluation: Bayes Optimal Classifier (Soft Voting) ===
Accuracy: 0.7093
Macro-averaged F1 score: 0.6150

Classification Report:
              precision    recall  f1-score   support

  BACKGROUND       0.56      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      0.62     30135
weighted avg       0.70      0.71      0.70     30135
```


```
=== Final Evaluation: Bayes Optimal Classifier (Soft Voting) ===
Accuracy: 0.7093
Macro-averaged F1 score: 0.6150
...
    accuracy                           0.71     30135
   macro avg       0.66      0.60      0.62     30135
weighted avg       0.70      0.71      0.70     30135
```