UE23CS352A: Machine Learning Lab

Week 12: Naive Bayes Classifier

Name:- Deepthi J Kumbar

SRN:- PES2UG23CS164

Section:-5C

Date:- 31/10/25

Objective:-

The goal of this lab is to understand and implement the Naive Bayes classification algorithm for text classification. We work with biomedical abstract sentences from the PubMed dataset to predict which section (Background, Methods, Results, Objective, or Conclusion) each sentence belongs to.

Tasks Performed:-

● Implement the Multinomial Naive Bayes classifier from scratch

. ● Utilize scikit-learn's tools for text vectorization (CountVectorizer, TfidfVectorizer) and modeling (MultinomialNB).

● Perform hyperparameter tuning using GridSearchCV to find the optimal model settings.

● Approximate the Bayes Optimal Classifier (BOC) using an ensemble method built using diverse base models (hypothesis) and a Soft Voting Classifier using calculated posterior weights.

Methodology:-

1. Multinomial Naive Bayes (MNB):

- Converted text data into numerical features using CountVectorizer and TF-IDF techniques.
- Implemented Multinomial Naive Bayes from scratch, calculating:
    - Log priors → probability of each class.
    - Log likelihoods → probability of words given each class (with Laplace smoothing).
- Predicted the class with the highest combined log probability for each sentence.
- Evaluated model performance using accuracy, F1-score, and a confusion matrix.

2. Bayes Optimal Classifier (BOC):

- Trained five diverse models — MultinomialNB, Logistic Regression, Random Forest, Decision Tree, and KNN — on a sampled subset of the training data.
- Computed posterior weights for each model based on their log-likelihood performance on a validation set.
- Combined all models using a Soft Voting Classifier, applying the calculated posterior weights to approximate the theoretical Bayes Optimal Classifier.
- Evaluated final ensemble accuracy and macro F1-score on the test data.

Results and Analysis:-

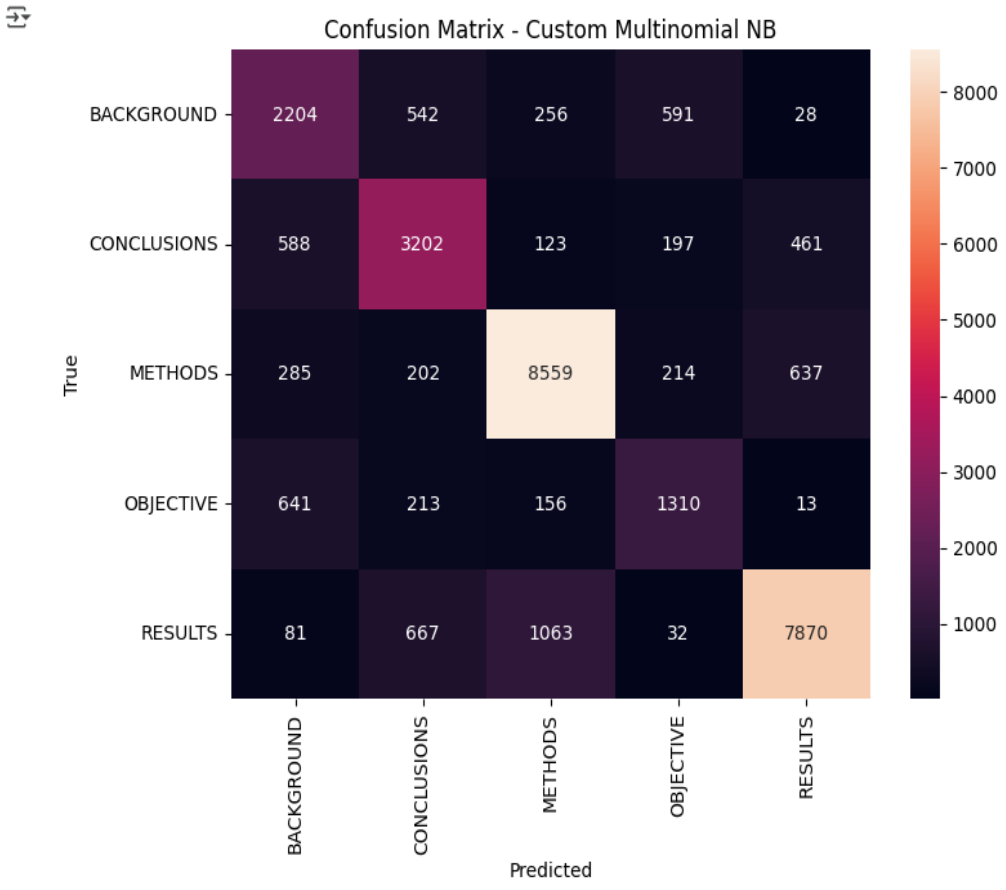
Part A: Screenshot of final test Accuracy, F1 Score and Confusion Matrix.

```
Accuracy (custom NB): 0.7680438028870085
Macro F1   (custom NB): 0.7060139534604033

Classification report:
              precision    recall  f1-score   support

  BACKGROUND       0.58      0.61      0.59      3621
 CONCLUSIONS       0.66      0.70      0.68      4571
     METHODS       0.84      0.86      0.85      9897
   OBJECTIVE       0.56      0.56      0.56      2333
     RESULTS       0.87      0.81      0.84      9713

    accuracy                           0.77     30135
   macro avg       0.70      0.71      0.71     30135
weighted avg       0.77      0.77      0.77     30135
```



Part B: Screenshot of best hyperparameters found and their resulting F1 score.

```
Fitting 3 folds for each of 8 candidates, totalling 24 fits
Best params: {'nb__alpha': 0.1, 'tfidf__ngram_range': (1, 2)}
Best score (f1_macro on dev folds): 0.6567136968391226
```

Part C: 1. Screenshot of SRN and sample size.

2. Screenshot of BOC final Accuracy, F1 Score and Confusion Matrix.

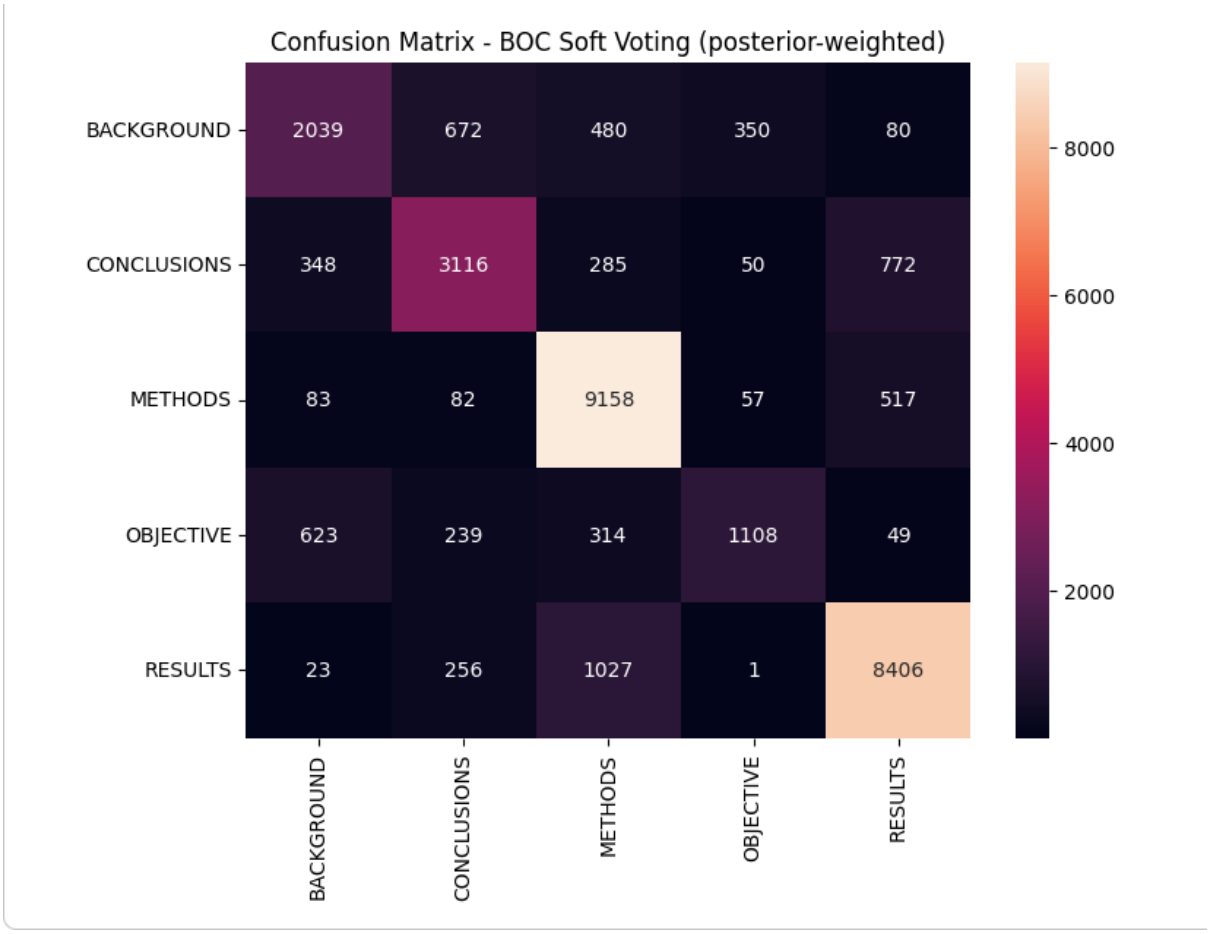
```
SRN: PES2UG23CS164
Sampled size: 36008
```

```
BOC Accuracy: 0.7906752945080471
BOC Macro F1: 0.7194288355466355

Classification report:
              precision    recall  f1-score   support

  BACKGROUND       0.65      0.56      0.61      3621
 CONCLUSIONS       0.71      0.68      0.70      4571
     METHODS       0.81      0.93      0.87      9897
   OBJECTIVE       0.71      0.47      0.57      2333
     RESULTS       0.86      0.87      0.86      9713

    accuracy                           0.79     30135
   macro avg       0.75      0.70      0.72     30135
weighted avg       0.78      0.79      0.78     30135
```



```
Test Accuracy (best pipeline): 0.7604446656711465
Test Macro F1  (best pipeline): 0.659870171277922

Classification report:
              precision    recall  f1-score   support

  BACKGROUND       0.64      0.46      0.54      3621
 CONCLUSIONS       0.66      0.66      0.66      4571
     METHODS       0.77      0.92      0.84      9897
   OBJECTIVE       0.67      0.30      0.42      2333
     RESULTS       0.84      0.87      0.85      9713

    accuracy                           0.76     30135
   macro avg       0.71      0.64      0.66     30135
weighted avg       0.75      0.76      0.75     30135
```