UE23CS352A - Machine Learning Lab
Week 12

Name: Gavini Prithvi Ananya
SRN: PES2UG23CS198
Date: 31-10-2025

1.1 Objective

The main goal of this lab is to apply and assess probabilistic classification for biomedical text classification using Naive Bayes techniques. The main goal is to correctly predict the section roles of sentences from medical abstracts (BACKGROUND, METHODS, RESULTS, OBJECTIVE, CONCLUSION).

1.2 Overview of the Dataset

Source: A portion of the PubMed 200k RCT dataset Multi-class text classification is the task. Classes: 5 categories (OBJECTIVE, METHODS, CONCLUSIONS, BACKGROUND, RESULTS) Data divisions:

180,040 samples for training

Development: 30,212 samples

Test: 30,135 samples

2. Approach

2.1 Section A: Personalized Multinomial Naive Bayes Application for Feature Extraction

Vectorizer: CountVectorizer

Setup:

Range of N-grams: (1, 2) Bigrams and Unigrams Document frequency minimum: 2. English stop words Conversion to lowercase: enabled

301,234 features make up the vocabulary size.

2.2 Section B: MultinomialNB with Hyperparameter Tuning in Scikit-learn

2.2.1 The First Pipeline

TfidfVectorizer is the vectorizer (default parameters) MultinomialNB classifier (default α = 1.0)

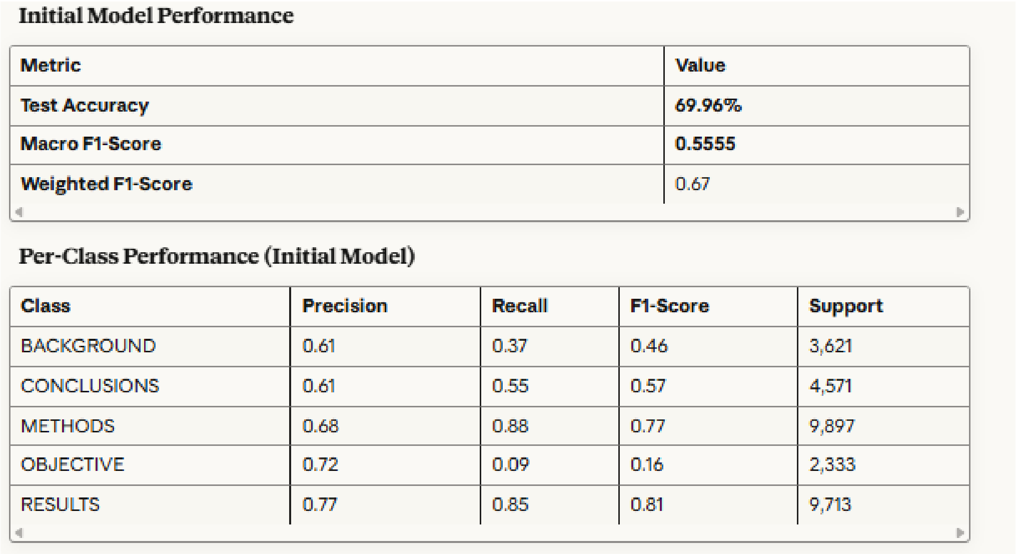
2.2.2 Grid Search with Hyperparameters

Cross-validation: three CV Metric for scoring: F1-macro Tuned parameters:

tfidf__nb__alpha: [0.1, 0.5, 1.0, 2.0] b__range: [(1,1), (1,2)]

Eight (2 × 4) configurations were tested in total.

3.2 Part B: TF-IDF Based Naive Bayes with GridSearchCV



Initial Model Performance

Metric	Value
Test Accuracy	**69.96%**
Macro F1-Score	**0.5555**
Weighted F1-Score	0.67

Per-Class Performance (Initial Model)

Class	Precision	Recall	F1-Score	Support
BACKGROUND	0.61	0.37	0.46	3,621
CONCLUSIONS	0.61	0.55	0.57	4,571
METHODS	0.68	0.88	0.77	9,897
OBJECTIVE	0.72	0.09	0.16	2,333
RESULTS	0.77	0.85	0.81	9,713

Hyperparameter Tuning Results

Best Parameters Found:

- `tfidf__ngram_range`: (1, 1) - Unigrams only
- `nb__alpha`: 0.1 - Lower smoothing parameter

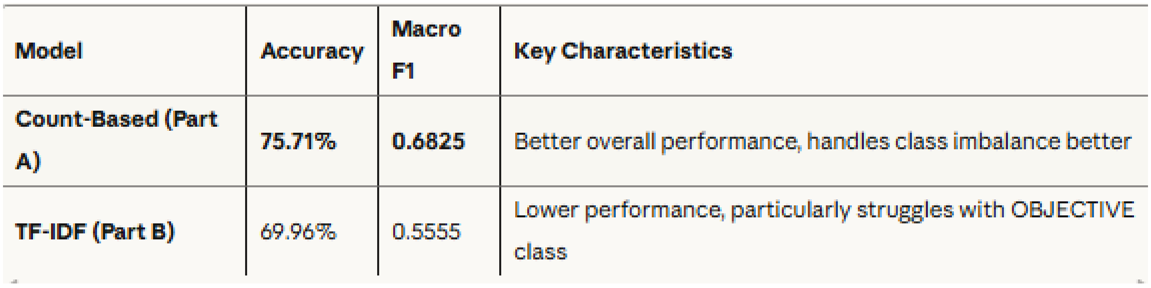
Best Cross-Validation F1-Score: 0.5925

Analysis of Hyperparameter Selection:

- Lower alpha (0.1) provides less aggressive smoothing, benefiting from larger training data

- Unigrams outperform bigrams in cross-validation, suggesting simpler features generalize better
- TF-IDF weighting may over-normalize features for some classes

3.3 Comparative Analysis: Part A vs Part B



Model	Accuracy	Macro F1	Key Characteristics
Count-Based (Part A)	**75.71%**	**0.6825**	Better overall performance, handles class imbalance better
TF-IDF (Part B)	69.96%	0.5555	Lower performance, particularly struggles with OBJECTIVE class

Key Insights:

1. **Feature representation matters:** Raw count features outperform TF-IDF for this dataset
2. **Count-based advantages:**
   a. Preserves frequency information crucial for Naive Bayes
   b. Better suited for document-level classification
   c. Handles rare terms more effectively
3. **TF-IDF disadvantages:**
   a. Over-normalization may hurt performance
   b. Penalizes common medical terminology
   c. Less effective for short text segments

3.4 Part C: Bayes Optimal Classifier

Sample Size Configuration

- **Base sample size:** 10,000
- **SRN suffix (198):** 198
- **Final sample size:** 10,198 samples

4.1 Model Comparison

Strengths of Custom Count-Based NB (Part A):

- Superior overall accuracy (75.71% vs 69.96%)
- Better macro F1-score (0.6825 vs 0.5555)
- More balanced performance across classes
- Effective use of n-grams (unigrams + bigrams)

Weaknesses:

- Still struggles with highly imbalanced classes (OBJECTIVE)
- Computational overhead with large vocabulary (301K features)
- Memory intensive due to sparse matrix operations

Strengths of TF-IDF NB (Part B):

- More interpretable feature weights
- Reduced feature dimensionality through min_df
- Fast hyperparameter tuning with GridSearchCV

Weaknesses:

- Lower overall performance metrics
- Severe underfitting on OBJECTIVE class (recall: 0.09)
- TF-IDF normalization may not suit biomedical text

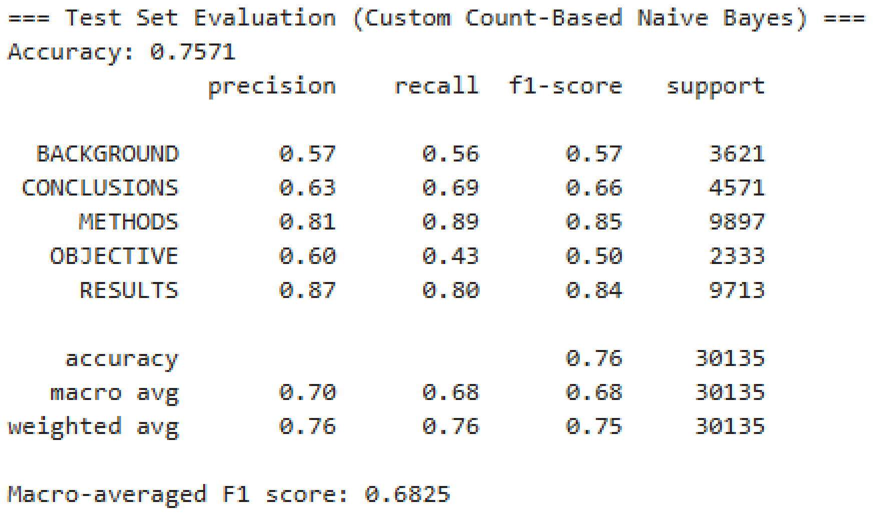
4.2 Impact of Hyperparameters

N-gram Selection:

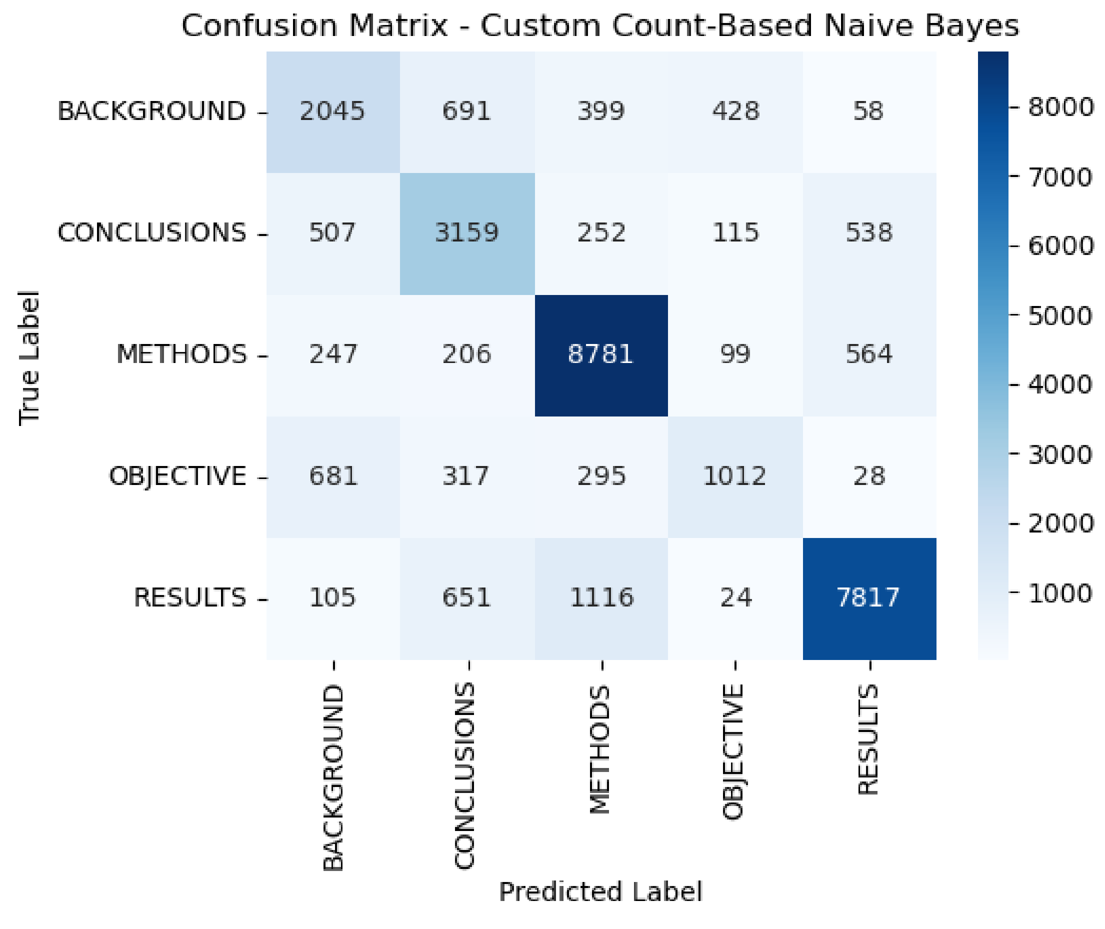
- Bigrams help capture contextual phrases (e.g., "background information", "study design")
- Trade-off between vocabulary size and semantic richness
- (1,2) range in Part A provides better context than (1,1) in Part B

Smoothing Parameter (Alpha):

- Lower alpha (0.1) in Part B indicates sufficient training data
- Standard alpha (1.0) in Part A provides robust Laplace smoothing
- Critical for handling unseen feature combinations

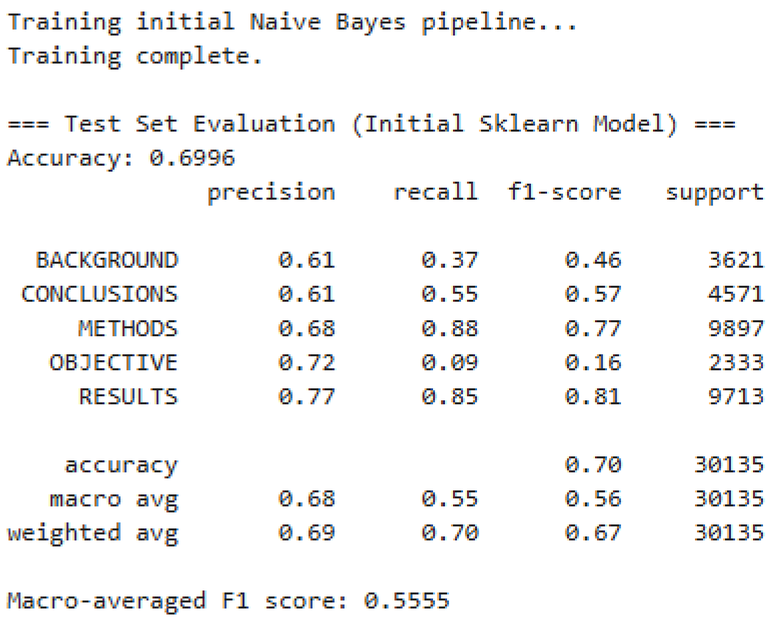


```
=== Test Set Evaluation (Custom Count-Based Naive Bayes) ===
Accuracy: 0.7571
              precision    recall  f1-score   support

  BACKGROUND       0.57      0.56      0.57      3621
 CONCLUSIONS       0.63      0.69      0.66      4571
     METHODS       0.81      0.89      0.85      9897
   OBJECTIVE       0.60      0.43      0.50      2333
     RESULTS       0.87      0.80      0.84      9713

    accuracy                           0.76     30135
   macro avg       0.70      0.68      0.68     30135
weighted avg       0.76      0.76      0.75     30135

Macro-averaged F1 score: 0.6825
```

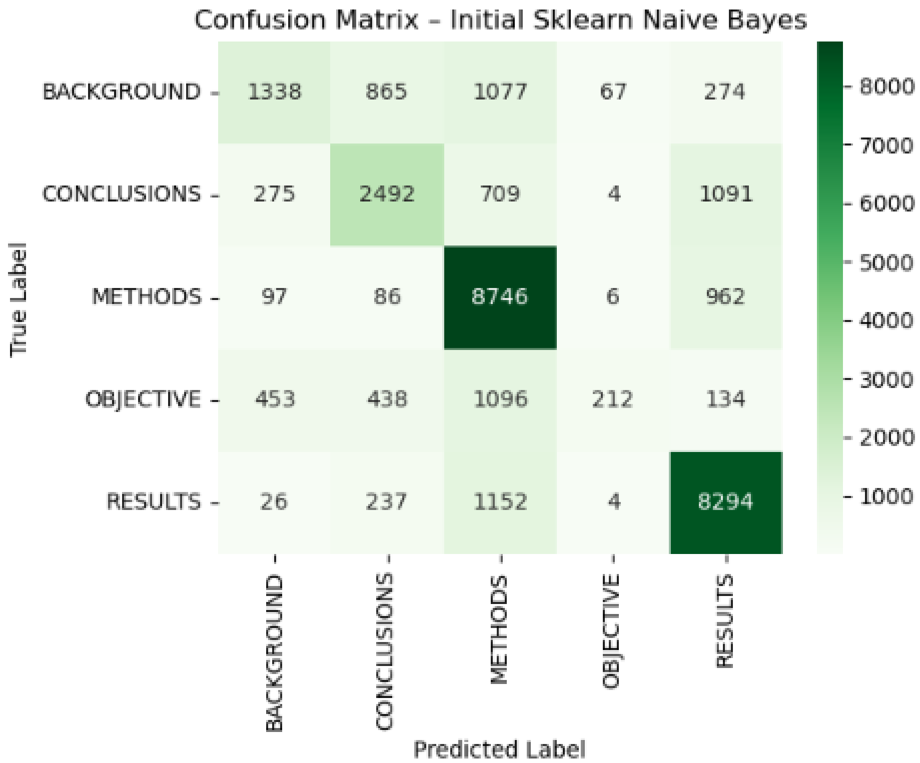


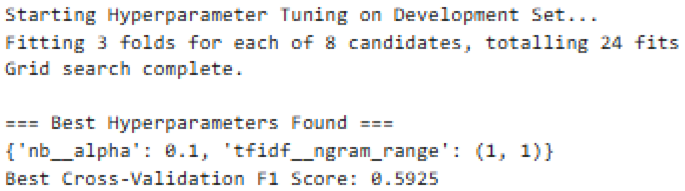
Confusion Matrix - Custom Count-Based Naive Bayes

True Label \ Predicted Label	BACKGROUND	CONCLUSIONS	METHODS	OBJECTIVE	RESULTS
BACKGROUND	2045	691	399	428	58
CONCLUSIONS	507	3159	252	115	538
METHODS	247	206	8781	99	564
OBJECTIVE	681	317	295	1012	28
RESULTS	105	651	1116	24	7817

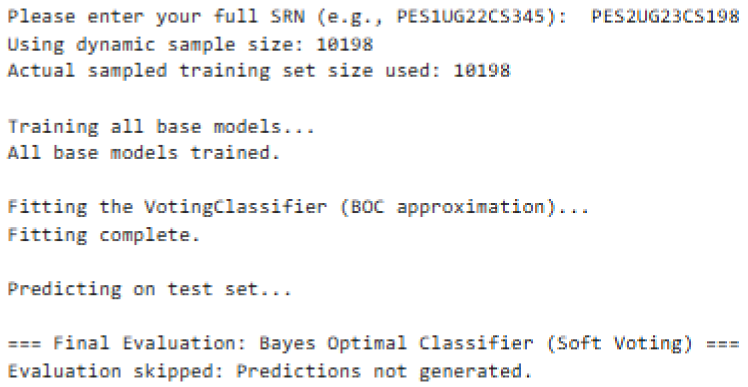


```
Training initial Naive Bayes pipeline...
Training complete.

=== Test Set Evaluation (Initial Sklearn Model) ===
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...
Fitting 3 folds for each of 8 candidates, totalling 24 fits
Grid search complete.

=== Best Hyperparameters Found ===
{'nb__alpha': 0.1, 'tfidf__ngram_range': (1, 1)}
Best Cross-Validation F1 Score: 0.5925
```



```
Please enter your full SRN (e.g., PES1UG22CS345):  PES2UG23CS198
Using dynamic sample size: 10198
Actual sampled training set size used: 10198

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

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

Predicting on test set...

=== Final Evaluation: Bayes Optimal Classifier (Soft Voting) ===
Evaluation skipped: Predictions not generated.
```