Department of Computer Engineering
University of Peradeniya
CO544: Machine Learning and Data Mining

Lab 05: Text Classification and Performance Analysis

E/20/420 : WANASINGHE J.K.

TASK 1: Find data preprocessing steps other than mentioned above.

1. HTML Tag Removal
   - Purpose: Remove non-semantic HTML tags (e.g., <br>, <p>) from web-scraped reviews
   - Impact: Eliminates formatting noise while preserving text content
2. Contraction Expansion
   - Purpose: Manually expand common contractions to full forms
   - Impact: Unifies "don't" and "do not" as identical features
   - Key Contractions Handled: 12+ common combinations (e.g., "can't" → "cannot")
3. Negation Handling
   - Purpose: Add NOT_ prefix to words following negation (e.g., "not good" → "NOT_good")
   - Impact: Critical for sentiment analysis - clearly distinguishes negated terms
   - Negation Words: "not", "no", "never", "none", "nobody", "nothing", etc.
4. Emotional Punctuation Retention
   - Purpose: Preserve exclamation/question marks while removing other punctuation
   - Impact: Maintains sentiment intensity indicators (e.g., "Great movie!" vs "Great movie")
5. Word Length Filtering
   - Purpose: Remove very short (1 char) and very long (>15 chars) words
   - Impact: Eliminates noise from single-character fragments and likely typos
6. Stop Word Removal (with Negation Preservation)
   - Purpose: Remove common words ("the", "and") while preserving negation words
   - Impact: Reduces feature space while maintaining sentiment-critical words
7. Robust Lemmatization
   - Purpose: Normalize words to root forms with error handling
   - Impact: Unifies variants (e.g., "running" → "run") while preventing crashes

TASK 2: Discuss advantages and disadvantages of the Bag of Words model.

Advantages:

- Simplicity: Easy to understand and implement
- Computational efficiency: Fast processing for large datasets
- Interpretability: Clear feature-to-word mapping
- Baseline performance: Good starting point for text classification

Disadvantages:

- Loss of word order: "good movie" vs "movie good" treated identically
- Sparse representation: Most features are zero in high-dimensional space
- No semantic understanding: "excellent" and "great" treated as completely different
- Context ignorance: Cannot handle sarcasm or context-dependent meanings

TASK 3: Train a Random Forest model, a Support Vector Machine model and a Naive Bayesian classifier. Compare the accuracies and other performance measures (precision, recall, F1-score, confusion matrix) of all four models including the Logistic Regression model. What is the best model? Justify your answer based on these measures.

When the training was done with the preprocessing techniques mentioned in the lab sheet;

```
==================================================
Training Logistic Regression...

Logistic Regression Results:
Accuracy: 0.8200

Confusion Matrix:
[[164  44]
 [ 28 164]]

Classification Report:
              precision    recall  f1-score   support

         neg       0.85      0.79      0.82       208
         pos       0.79      0.85      0.82       192

    accuracy                           0.82       400
   macro avg       0.82      0.82      0.82       400
weighted avg       0.82      0.82      0.82       400


==================================================
Training Random Forest...

Random Forest Results:
Accuracy: 0.8275

Confusion Matrix:
[[168  40]
 [ 29 163]]

Classification Report:
              precision    recall  f1-score   support

         neg       0.85      0.81      0.83       208
         pos       0.80      0.85      0.83       192

    accuracy                           0.83       400
   macro avg       0.83      0.83      0.83       400
weighted avg       0.83      0.83      0.83       400


==================================================
Training Support Vector Machine...
```

```
Support Vector Machine Results:
Accuracy: 0.8100

Confusion Matrix:
[[165  43]
 [ 33 159]]

Classification Report:
              precision    recall  f1-score   support

         neg       0.83      0.79      0.81       208
         pos       0.79      0.83      0.81       192

    accuracy                           0.81       400
   macro avg       0.81      0.81      0.81       400
weighted avg       0.81      0.81      0.81       400


==================================================
Training Naive Bayes...

Naive Bayes Results:
Accuracy: 0.8150

Confusion Matrix:
[[166  42]
 [ 32 160]]

Classification Report:
              precision    recall  f1-score   support

         neg       0.84      0.80      0.82       208
         pos       0.79      0.83      0.81       192

    accuracy                           0.81       400
   macro avg       0.82      0.82      0.81       400
weighted avg       0.82      0.81      0.82       400

============================================================
MODEL PERFORMANCE COMPARISON
============================================================
                 Model Accuracy Precision Recall F1-Score
   Logistic Regression   0.8200    0.8226 0.8200   0.8200
         Random Forest   0.8275    0.8289 0.8275   0.8276
Support Vector Machine   0.8100    0.8112 0.8100   0.8101
           Naive Bayes   0.8150    0.8162 0.8150   0.8151

Best Model: Random Forest with accuracy: 0.8275
```

Based on the results, it was observed that the Random forest was high in Accuracy, Precision, Recall and also in F1-score, making it the best model for the data.

When the training was done with the preprocessing techniques mentioned in the lab sheet as well as other introduced techniques.

```
==================================================
Training Logistic Regression...

Logistic Regression Results:
Accuracy: 0.8100

Confusion Matrix:
[[160  48]
 [ 28 164]]

Classification Report:
              precision    recall  f1-score   support

         neg       0.85      0.77      0.81       208
         pos       0.77      0.85      0.81       192

    accuracy                           0.81       400
   macro avg       0.81      0.81      0.81       400
weighted avg       0.81      0.81      0.81       400


==================================================
Training Random Forest...

Random Forest Results:
Accuracy: 0.8200

Confusion Matrix:
[[160  48]
 [ 24 168]]

Classification Report:
              precision    recall  f1-score   support

         neg       0.87      0.77      0.82       208
         pos       0.78      0.88      0.82       192

    accuracy                           0.82       400
   macro avg       0.82      0.82      0.82       400
weighted avg       0.83      0.82      0.82       400


==================================================
Training Support Vector Machine...

Support Vector Machine Results:
Accuracy: 0.7925

Confusion Matrix:
[[162  46]
 [ 37 155]]

Classification Report:
              precision    recall  f1-score   support

         neg       0.81      0.78      0.80       208
         pos       0.77      0.81      0.79       192
    accuracy                           0.79       400
   macro avg       0.79      0.79      0.79       400
weighted avg       0.79      0.79      0.79       400


==================================================
Training Naive Bayes...

Naive Bayes Results:
Accuracy: 0.8200

Confusion Matrix:
[[167  41]
 [ 31 161]]

Classification Report:
              precision    recall  f1-score   support

         neg       0.84      0.80      0.82       208
         pos       0.80      0.84      0.82       192

    accuracy                           0.82       400
   macro avg       0.82      0.82      0.82       400
weighted avg       0.82      0.82      0.82       400


============================================================
MODEL PERFORMANCE COMPARISON
============================================================
                 Model Accuracy Precision Recall F1-Score
   Logistic Regression   0.8100    0.8139 0.8100   0.8099
         Random Forest   0.8200    0.8255 0.8200   0.8198
Support Vector Machine   0.7925    0.7935 0.7925   0.7926
           Naive Bayes   0.8200    0.8212 0.8200   0.8201

Best Model: Random Forest with accuracy: 0.8200
```

Here, both Random Forest and Naïve Bayes has equal accuracy and recall, however, when it comes to precision Random forest leads with around 0.043 and falls behind in F1 score from about 0.003. Therefore, the best model for this would be Random Forest.