CO544: Machine Learning and Data Mining

Lab 05: Text Classification and Performance Analysis

E/20/197 Kawya A.H.D.

TASK1:

Find data preprocessing steps other than mentioned above.

1. Removing Stop Words

Stop words (like “is”, “the”, “on”, “at”) don’t carry much sentiment or meaning.

Removing them improves performance by focusing on meaningful words.

2. Stemming

Instead of converting words to dictionary form (lemmatization), stemming just removes suffixes. Faster, but less accurate than lemmatization.

3. Spelling Correction

We can use Text Blob or Sys Spell to correct common typos.

4. Removing HTML Tags

If the reviews contain HTML, this helps to remove them.

5. Removing Repeated Characters

TASK2:

Discuss advantages and disadvantages of the Bag of Words model.

Advantages

- Simple to understand
- Easy to implement
- Effective for basic text classification
- Fast computation
- Works well with traditional ML models

Disadvantages

- No semantic meaning
- No word order or grammar
- Sparse vectors - huge matrices mostly filled with zeros
- Vocabulary size grows fast
- Same word treated differently if spelling varies – This happens unless preprocessed properly, ex: “run” vs. “running”

TASK3:

Train a Random Forest model, a Support Vector Machine model and a Naïve 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.

```
--- Logistic Regression ---
Accuracy: 0.7945544554455446
Confusion Matrix:
 [[152  42]
 [ 41 169]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.78      0.79       194
           1       0.80      0.80      0.80       210

    accuracy                           0.79       404
   macro avg       0.79      0.79      0.79       404
weighted avg       0.79      0.79      0.79       404
```

```
--- Random Forest ---
Accuracy: 0.8193069306930693
Confusion Matrix:
 [[164  30]
 [ 43 167]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.85      0.82       194
           1       0.85      0.80      0.82       210

    accuracy                           0.82       404
   macro avg       0.82      0.82      0.82       404
weighted avg       0.82      0.82      0.82       404
```

```
--- Support Vector Machine ---
Accuracy: 0.754950495049505
Confusion Matrix:
 [[150  44]
 [ 55 155]]
Classification Report:
              precision    recall  f1-score   support

           0       0.73      0.77      0.75       194
           1       0.78      0.74      0.76       210

    accuracy                           0.75       404
   macro avg       0.76      0.76      0.75       404
weighted avg       0.76      0.75      0.76       404
```

```
--- Naive Bayes ---
Accuracy: 0.7970297029702971
Confusion Matrix:
 [[154  40]
 [ 42 168]]
Classification Report:
              precision    recall  f1-score   support

           0       0.79      0.79      0.79       194
           1       0.81      0.80      0.80       210

    accuracy                           0.80       404
   macro avg       0.80      0.80      0.80       404
weighted avg       0.80      0.80      0.80       404
```

Random forest model was the best model based on the accuracy (81.93%) AND F1-score. It had balanced precision and recall, indicating consistent performance across both classes.

While Naive Bayes and Logistic Regression also performed well, Random Forest slightly outperformed them.

Therefore, **Random Forest** is recommended for this classification task.