ASSIGNMENT_3_AAI_520_Text Classification using TF-IDF

KWABENA MENSAH

University of San Diego

BBC News Summary Dataset

Content

The dataset contains news articles from the BBC News website, categorized into five topics: business, entertainment, politics, sport, and tech. Each text document is labeled with its corresponding category.

Approach/Requirements

1. Import the necessary libraries: pandas, scikit-learn, numpy.
2. Load the "BBC News Summary" dataset into a pandas DataFrame.
3. Preprocess the text data (remove stopwords, punctuation, lowercase, etc.).
4. Split the data into training and testing sets.
5. Apply TF-IDF Vectorization on the training data.
6. Build a text classification model using a classifier of your choice (e.g., Naive Bayes, Support VectorMachine, Random Forest, etc.).
7. Train the model using the TF-IDF transformed training data.
8. Predict the categories for the testing data.
9. Evaluate the model's performance using the following classification metrics:
   - Accuracy
   - Precision (weighted)
   - Recall (weighted)
   - F1 Score (weighted)
   - Confusion Matrix
   - Area Under the Receiver Operating Characteristic curve (AUC-ROC)

Summary of Results

1. The metric (AUC-ROC) was not used in evaluation because this is a multiclass exercise. Using it was possible but will have required a significant effort.
   The ROC curve and its associated AUC metric are primarily tailored for binary classification tasks. When extended to multiclass problems, challenges arise.
   A common method is the one-vs-all approach, where each class is treated as positive against all others, resulting in multiple ROC curves that can be hard to interpret. Aggregating these AUC values into a single metric, like a macro-average, can obscure class-specific performance nuances. Furthermore, multiclass problems often have imbalanced distributions, making the one-vs-all approach misleading for minor classes. By converting a multiclass problem to multiple binary ones, relationships between classes can be lost. For comprehensive evaluation of multiclass classifiers, it's crucial to consider other metrics alongside AUC-ROC.

2. There are three notebooks:
   In Notebook 1, I evaluated four models: MultinomialNB, Logistic Regression, Random Forest Classifier and Support Vector Machine (SVM).
   Based on the results of the models and several iterations we applied Grid Search to Logistic Regression and saw an improvement in Recall (See Notebook 2)

   In Notebook 2, I demonstrated that Gridsearch can improve metrics but it is computationally expensive and time consuming.

   In Notebook 3 I focused on Random Forest and had exceptional results on the Test Set.

```
               precision    recall  f1-score   support

     business       0.97      0.97      0.97        73
entertainment       0.98      0.94      0.96        63
     politics       0.98      0.97      0.97        60
        sport       0.95      0.99      0.97        73
         tech       0.97      0.98      0.98        65

     accuracy                           0.97       334
    macro avg       0.97      0.97      0.97       334
 weighted avg       0.97      0.97      0.97       334
```