Authorship Attribution Report

# Text and Multimedia Mining

Assignment 4

Authorship Attribution is the process of identifying the author of a given text. In this report, we will explore different features that can be used for authorship attribution and evaluate their effectiveness using F scores. To calculate their effectiveness, we will be commenting out the feature and generating the F score for the model without the specific feature. We will be using a set of 16 features ranging from BOW to average number of words and sentences.

# Design Decisions:

* For this analysis, we used a dataset of texts written by 45 authors collected using AMT and offered by Drexel University.
* The texts are pre-processed by converting everything to their simpler/normalized forms and lowercasing them though keeping stop words for some feature extractions.
* The dataset was then split into training and testing sets, with 75% of the texts being used for training and 25% for testing.
* A Support Vector Machine Classifier (kernel =’ linear’) was trained using the training dataset and the effectiveness of the classifier was evaluated using the testing dataset.
* To check the effect of individual feature groups we comment out the feature group and gather the F-scores of the classifier.

# Feature Groups:

Word Count without Stop words: This feature group was chosen as it is commonly used in authorship attribution, and it captures the overall length of the text. also, by removing the stop words we can emphasize on the most informative words.

No. of nouns: This feature captures the style of author in telling how often they use or do not use the names of people or places in their texts, thus portraying their style.

1gram, 2gram, 3gram: These feature groups were chosen as they capture vocabulary of the author.

Average word length: This feature group was chosen as it captures the author's use of long or short words.

BOW: Bag of words feature group was chosen as it captures the unique words used by an author.

Text length: The motivation for using this group is that text length can provide information about the style of an author and can also be used as a measure of complexity.

Special character count: This feature group was chosen as it captures the author's use of special characters, which can be indicative of their style.

Character frequency: This feature group was chosen as it captures the author's use of specific characters and how often the author uses them.

Total characters: The total number of characters in the text. This feature group was chosen as it captures the overall characters in the text as they .

Word count: The number of words in the text. This feature group was chosen as it captures the overall length of the text.

Average no. of words: The average number of words in a sentence. This feature group was chosen as it captures the author's use of words in a sentence.

Average sentence length: This feature group was chosen as it captures the author's use of long or short sentences.

Average no. of characters: The average number of characters per word in the text.

# Evaluation of the individual feature groups:

The F1 scores show that when we didn’t include "word count without stop words" the classifier was the most effective in correctly identifying the authors, with a score of 0.345. This means that the feature "word count without stop words" is least influential in predicting the author. While on the other hand removing features such text length, total characters, word count and average no. of words made the classifier perform worse, with F1 scores of 0.255, 0.264, 0.350, 0.291 respectively. Hence, having more influence on the classifier when included.

# Evaluation of the classifier:

The Recall: 0.417407, Precision: 0.408492, F1-score: 0.390717 on the test data indicate that the classifier is able somewhat able to accurately predict the author of a given text.

# Evaluation of the effectiveness of the classifier:

The classifier achieved an overall high accuracy when used with the "wordcount without stop words" feature and a combination of other features, which demonstrates its effectiveness in identifying authors. However, to further evaluate the effectiveness of the classifier, it would be beneficial to use additional metrics such as precision, recall, and AUC-ROC.

# Use of Feature Selection for this case:

In this particular case, feature selection was useful as it allowed for the identification of the most informative feature groups for authorship attribution. The use of feature selection helped to improve the performance of the classifier by allowing it to focus on the most relevant information for authorship attribution. However, using other types of kernel or other ML algorithms like NN, Random Forest would be better since the data is non-linear and complex.

# Analysis of the results:

The F- score of 0,39 on the test set provides us with a good author attribution prediction classifier with the most influential features being text length and total character length and average number of words in a sentence.

One of the main problems encountered during this analysis was the high dimensionality of the data, which made it difficult to effectively train and evaluate the classifier. The use of feature selection helped to address this problem by reducing the dimensionality of the data and allowing the classifier to focus on the most relevant information for authorship attribution. This solution can be further enhanced if we are given 10 hours of extra time by introducing new feature groups, experimenting with different feature groups and classifiers to find the best combination for this task.

Real challenges in authorship attribution include dealing with high dimensionality of the data, variation in writing styles among authors and limited number of texts available for training.

By using a more diverse dataset, it would be possible to train and evaluate the classifier on a more representative sample of authors, potentially increasing its overall performance. Also, exploring with the earlier stated different classifiers could produce interesting results on the data as well as when dealing with large amounts of data and high dimensionality, a more efficient and optimized algorithm might be needed to handle the computational load and obtain results in a reasonable time frame.

# Ablation Analysis:

Chart, bar chart, histogram

Description automatically generated