Authorship Attribution

TxMM 4-AA — Group: 33

Marie-Sophie Simon s1023848
Thijs de Jong s1015438
Rik Jansen s1019150

November 14, 2023

Design Choices

We have chosen the following feature groups for our classifier, giving us a total of 244 features:

Character based features

This feature group consists of 5 features with relation to characters. the features in this group are: total character count, uppercase frequency, digit frequency, whitespace frequency, and alphabet frequency. The alphabet count is the number of letters used in the text.

This feature group will give us more information about the write style of an author, as well as the diversity of characters used in the text.

Individual letter frequencies

This feature group contains for each letter in the alphabet this letter's frequency, giving a total of 26. The frequency of a letter is calculated by counting the total number of occurrences per letter, and dividing this number by the total character count.

This feature group can capture unique writing styles, preferences for specific letters, and patterns used by an author. It can also indicate the use of difficult words.

Special character frequencies

This feature group is almost identical to the previous one, only it contains frequencies of 21 special characters (not including punctuation characters). The frequency is calculated by dividing the total special character count by the total character count.

Like last feature group, this group can also capture unique writing styles, or use of unique symbols.

Punctuation frequencies

Again, this feature is almost identical to the previous one, only then for all punctuation characters, containing 8 features.

This feature group is useful to see syntactic preferences, as well as emphasis in a text.

Word based features

This group contains features that are related to words, like: word count, average word length, average sentence length (with relation to total characters and total words), and the unique words frequency. This feature group contains 5 features.

Word based features are give information about average word and sentence lengths, and duplication of certain words. It can help establishing the author's vocabulary and writing complexity.

Function words

This feature group has with 150 features the most features. It contains the frequency of 150 function words[1]. Function words are words that express primarily a grammatical relationship. The frequency is calculated by for each word dividing the number of occurrences by the total word count.

These features will give us information about the writing style and preferences of an author.

Structural features

There are 2 features in this group, namely the sentence count, and the repeated punctuation frequency. The sentence count is calculated by splitting on a period. Repeated punctuations can occur for extra emphesis, like *This is rediculous!!!*. The number of repeated characters is then divided by the total number of punctuation characters.

The structural features can show information about the way an author organizes its text, as well as their writing style.

Part-Of-Speech tags

This feature group contains frequencies for each individual Part-Of-Speech tag. Parts-Of-Speech are also known as word classes or lexical categories. We exclude all POS tags for special characters, as we have other features for these already. The total number of occurrences per tag is divided by the total number of tags to get the frequency.

These features can provide insight to an author's syntactic style, preferences of word types, and sophistication of a text.

Evaluation

We got the following accuracies on the development dataset after applying ablation analysis:

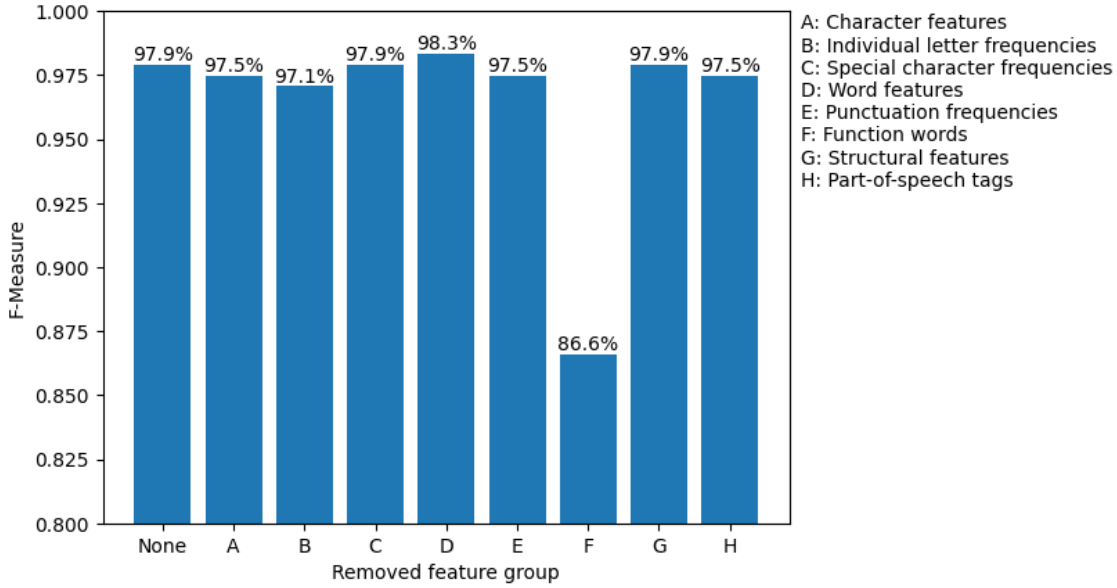

Figure 1: Ablation analysis on the feature groups

[1]Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques

As we can see in 1, most of the feature groups (A, C, E, H) make barely any or no difference for the classifier when removed. There are 2 feature groups howevever that have a negative impact on the classifier when removed, being *Individual letter frequencies* and *Function words*, where *Function words* has by far the most negative impact. The feature group *Word features* has a positive impact upon being removed.

The classifier itself is very effective, with an accuracy score of 97,9%, and 244 features. In our case, feature selection will probably not be useful, as the performance of the classifier is already pretty high, and adding feature selection would make the classifier slower.

Analysis

The feature groups *Character features*, *Special character frequencies*, *Punctuation frequencies* and *Part-of-Speech tags* have little to no impact on the classifier when being removed. This probably has to do with the fact that such a small number of features get removed, and the classifier will fill these gaps by using the other features.

The feature group *Individual letter frequencies* has a slightly bigger negative impact, albeit still not a lot. The reason for this is likely the same as with the previous groups; the gaps created by removing this feature group will be filled in by the other feature groups. The only difference here is that this group consists of 26 features, meaning it is bigger than the previous groups, which could explain the slightly bigger impact.

The feature group *Function words* has by far the biggest negative impact, but also contains the most features, with 150. Since this is more than half of our features, it comes as no surprise that this group has the biggest impact.

The feature group *Word features* has a positive impact upon being removed. An explanation for this could be that these features have very similar values, thus making it more difficult for the classifier to distinguish different authors, instead of making it easier. To further improve our classifier, we could remove this feature group.

The development set performed almost perfectly with an accuracy of 97,9%. The test set performed almkost identical; its accuracy was 97,5%.

Problems

We encountered a few problems during the development of our classifier. First of all, it was not completely clear to us how we should incorporate the development set in tuning the hyper-parameters. Next to that, it was somewhat tricky to store the features nicely, as each feature should be in a feature group as well. If we had 10 hours extra time, we could have spent time looking at the training data, to extract context-specific words. Next to that, we could also cluster feature groups and see the impact when we for instance cluster some of the groups that have barely any impact when this group alone is removed.

In Authorship Attribution, real challenges could be selecting the right features for content-specific Authorship Attribution, since it would matter whether all authors write poems or are amateur authors that post comments on internet. Next to that, authors could copy write styles from other authors, making it difficult to attribute authorship highly accurately. Lastly, authors themselves could also have different genres they write in, making it hard to map specific features to a specific author.