# FOI Cross-Domain Authorship Attribution for Criminal Investigations

## Poster for PAN at CLEF 2019

Authorship attribution techniques have existed for a long time, but they are seldom evaluated in conditions similar to the real-world scenarios in which they have to work if they should be useful tools in criminal investigations involving digital communication. We have used a **SVM** classifier as a base, onto which we have added two sets of hand-crafted **stylometric features** and evaluated it using data from the PAN-CLEF 2019 cross-domain authorship attribution task. Results outperform the baseline systems to which our classifiers have been compared.

### Authors

Fredrik Johansson and Tim Isbister
FOI, Swedish Defence Research Agency

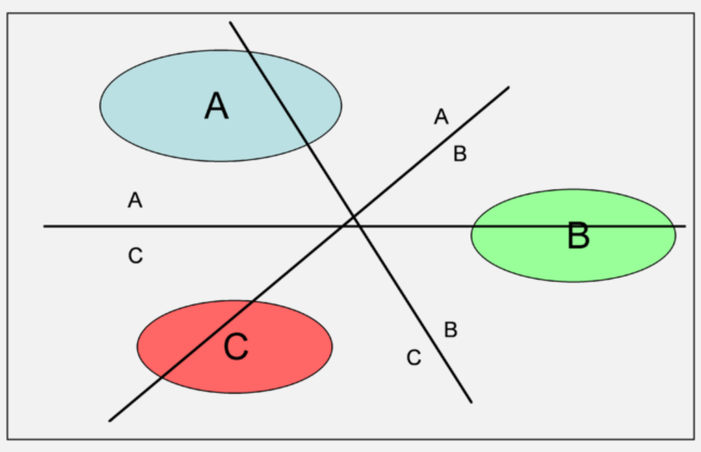
| isbister19 | johansson19 |
|---|---|
| Lix | Lix |
| CharUpperLowerRatio | CharUpperLowerRatio |
| CountWordCaps | CountWordCaps |
| avg_sen_len | avg_sen_len |
| std_sen_len | std_sen_len |
| lex_diversity | avg_word_len |
| avg_word_len | std_word_len |
| std_word_len | word_sizes |
| shannon_entropy | word_ngrams |
| word_sizes | char_ngrams |
| word_ngrams | POS_ngrams |
| char_ngrams | Masked ngrams |
| binary_char_ngrams | |

## Features

This section presents the full list of hand-crafted features that have been implemented in our submitted solutions. Overall, stylometric features are intended to reflect stylistic characteristics of the writing of individual authors. The idea is that they should be as independent of topic as possible, and instead capture the more general writingstyle of an author.

## Transformation

The whole corpus for each problem (including the texts from the unknown authors as well) were used to create the vocabulary for the data-driven n-gram representations in the isbister19 submission. A small increase of performance could be gained when using the vocabulary also from the **unknown authors**. Grid search has been used for both submitted systems in order to find good parameters for the n-grams. In both systems we have concatenated all used features into a single feature vector. This vector has been transformed by scaling each feature by its **maximum absolute value**. Some-what surprisingly, the choice of scaling method had a huge impact on the predictive performance, as other standard scaling methods performed much worse.




## Classifier

We have experimented with several types of standard classifiers, but the linear **SVM** classifier performed consistently better than standard alternatives such as random forest and boosting classifiers. However, it is important to note that the SVM classifier performed much better when using a **one-vs-all** regime, training as many binary classifiers as there are candidate authors.

## Conclusion

The final submitted systems can be seen as extensions of the PAN-CLEF 2019 baseline-SVM system, to which we have added a large amount of hand-crafted stylometric features. The submitted systems have achieved overall scores of approximately **0.62** macro F1 on the final TIRA test set.

As future work, we would like to contrast this type of models with more modern pretrained deep-learning architectures such as **language models**, which are fine-tuned on the specific problem instances at hand.