# Dynamic Parameter Search for Cross-Domain Authorship Attribution

## Notebook for PAN at CLEF 2018

Benjamin Murauer, Michael Tschuggnall, and Günther Specht

Universität Innsbruck
<firstname>.<lastname>@uibk.ac.at

**Abstract** In this paper, we present our solution to the PAN workshop challenge of authorship attribution. In multiple sub-problems, the original authors from given documents have to be chosen from a fixed training set. The core of our approach lies in traditional character n-gram analysis in combination with a linear SVM as a classifier. To find optimal values for our model parameters, we combined predefined parameters determined by a preliminary run on a training set run with dynamically determined parameters from an ad-hoc grid search approach.

## 1 Introduction

The 2018 PAN workshop [12] features an authorship attribution task [3] that consists of multiple sub-problems. The main challenge of the task lies in two characteristics:

First, the training documents for each author cover a different domain than the testing documents. All texts are chosen from different fan fiction domains, in which fans of a specific author, novel, movie, TV-show, etc. produce new content while adhering to the original work's environment and atmosphere. Typical examples of such domains, which will be referred to as *fandom* in the remainder of this paper, are Harry Potter or Star Wars. The testing part of each dataset contains at least one text from each author, but only covers one fandom. The training part covers multiple fandoms that are distinct from the evaluation fandom, whereby the same number of documents is provided for each author. Table 1 shows an exemplary schematic overview of the task. It can be seen that every sub-problem can have a varying amount of authors, and each author can have different fandoms (displayed as capital letters) available for training.

Secondly, the sub-problems cover a variety of different languages and sizes. Documents within a sub-problem are all in the same language, and the sub-problems themselves can be English, French, Italian, Polish or Spanish.

For developing a classification model, the organizers of the challenge have provided a development dataset. The characteristics of this dataset can be seen in Table 2. Each contestant is given a virtual machine on the *tira* web service [6], on which the model can run. To prevent cheating, evaluation runs must be initiated with a web-application, and any external network connections are cut off from the virtual machine once started.

For evaluation, the macro-F1-score is computed for each sub-problem, and the arithmetic mean of these scores represents the final score of a contestant's model.

Table 1: Exemplary task structure. Capital letters denote possible fanfic domains.

| **Problem 1: English** | | | **Problem 2: English** | | | | **Problem 10: Spanish** | | |
|---|---|---|---|---|---|---|---|---|---|
| author | train | test | author | train | test | | author | train | test |
| 1 | A, D | B | 1 | G, D | A | ... | 1 | A, D | C |
| 2 | E | B | 2 | B, E, F | A | | 2 | E | C |
| 3 | A, E | B | 3 | C, G | A | | | | |
| | | | 4 | B | A | | | | |

Table 2: Training dataset characteristics

| **Problem** | **Language** | **Training docs** | **Testing docs** | **Avg. words/doc** | **Authors** |
|---|---|---|---|---|---|
| p01 | EN | 140 | 105 | 777 | 20 |
| p02 | EN | 35 | 21 | 782 | 5 |
| p03 | FR | 140 | 49 | 774 | 20 |
| p04 | FR | 35 | 21 | 782 | 5 |
| p05 | IT | 140 | 80 | 787 | 20 |
| p06 | IT | 35 | 46 | 807 | 5 |
| p07 | PL | 140 | 103 | 807 | 20 |
| p08 | PL | 35 | 15 | 788 | 5 |
| p09 | ES | 140 | 117 | 829 | 20 |
| p10 | ES | 35 | 64 | 851 | 5 |

## 2 Related Work

In 2011 and 2012, similar authorship attribution tasks have been held an PAN [1,2]. One of the main differences was that the dataset used consisted of single-topic (or domain) documents.

Sapkota et al. [8] show that the cross-topic nature of datasets increases the difficulty for classification tasks for many traditional models. Furthermore, they demonstrate that by increasing the number of topics that are used for training (while still not having the testing topic(s) available), the accuracy of their model can be increased significantly.

The use of character n-grams has been proven to be an efficient feature for authorship attribution in multiple works [4,10,9]. Variations to this feature (e.g., by distorting the text [11] or distinguishing the relative position of each n-gram inside each word [7]) can improve a model even further. Markov et al. [5] showed that character-n-gram-based models can be used efficiently for cross-topic authorship attribution, and preprocessing the corpora can improve the performance of cross-topic models.

Along the line of these works, the main focus of this work is to build a general-purpose model based on character n-grams that is able to perform well on different languages and topics.

## 3 Methodology

To make the start of the challenge easier for the contestants, a baseline script was provided along with the training dataset. It uses the scikit-learn python machine learning library[1] and character 3-gram frequencies in combination with a linear SVM. Based on this approach, we implemented several improvements.

Originally, we wanted to implement a dynamic search for optimal parameters for each sub-problem individually. Therefore, we made use of the grid search tool shipped with scikit-learn, which tries all combinations of given parameter ranges in a brute-force fashion. The detailed values of the tested parameters are listed in Table 3. All runs were performed with 5-fold cross validation, while optimizing the `f1_macro` target.

This way, an important limitation of the model is its computational complexity. Even with a small set of possible parameter values, the number of possible combinations increases quickly. For example, given the parameter combinations in Table 3 and the 5-fold cross validation, each sub-problem is trained 10,800 times before one parameter combination is selected for predicting the testing data. This made this approach difficult for the evaluation run, where no information regarding the corpus size was given. Therefore, we limited the complete grid-search procedure to the development dataset and fixed the parameters that showed the most similar values for multiple problems. For example, the parameter *strip accents*, which if set, removes accents from letters, was chosen to be *true* for most of the sub-problems of the development dataset by the grid search. For the sub-problems with the parameter set to *false*, we were not able to determine which characteristic of the sub-problem caused the change of the parameter. Therefore, a majority vote was performed instead and the parameter was set to *true* for the entire evaluation set.

Two remaining parameters (i.e., the minimal document frequency of character n-grams and their size), that did not show a clear majority among the sub-problems, were left in the grid search to be determined dynamically for each part of the evaluation set, yielding $6 \times 3 = 18$ possible parameter combinations for each sub-problem. The fixed values are displayed in bold in Table 3, whereas the per-problem trained parameters are printed in italics.

## 4 Results

The results of the development dataset can be seen in Table 4a. As expected, it can be seen that the (even numbered) shorter sub-problems (containing less authors) are generally easier to classify than the bigger ones. No significant differences between the different languages can be detected. For overall evaluation and comparison between the

[1] `http://scikit-learn.org`

Table 3: Parameters tested with grid search. Parameters with results in bold face are fixed for each problem. Parameters written in italics are dynamically adapted.

| **Feature Parameter** | **Range tested** |
|---|---|
| *minimal document frequency* | 0 - 5 |
| *n-gram order* | 3, 4, 5 |
| lowercase | true, **false** |
| use idf normalization | **true**, false |
| strip accents | true, **false** |
| norm | **None**, L1, L2 |
| svm C | **0.01**, 0.1, 1, 10, 100 |

contestants, the arithmetic mean of all macro-F1-scores is used, which is displayed at the bottom of the table.

While the details of the evaluation dataset are not available at the time of writing this paper, the evaluation results are visible for the author. In Table 4b, the F1-scores for the evaluation dataset are displayed. The sparse information available suggests that problems 13,15 and 16 seem to be especially hard for our model and have lessened our total score notably. However, no implications can be made on the reasons for this performance at this point.

Table 5 shows the final ranking of the contestants. It can be seen that our solution reached the second rank, clearly beating the baseline provided by the task organizers.

Table 4: F1-scores of the final solution

(a) Development dataset

| **Problem** | **F1-score** |
|---|---|
| p01 | 0.565 |
| p02 | 0.774 |
| p03 | 0.658 |
| p04 | 0.777 |
| p05 | 0.690 |
| p06 | 0.588 |
| p07 | 0.549 |
| p08 | 0.867 |
| p09 | 0.791 |
| p10 | 0.827 |
| Average | 0.708 |

(b) Evaluation dataset

| **Problem** | **F1-score** | **Problem** | **F1-score** |
|---|---|---|---|
| p01 | 0.73 | p11 | 0.841 |
| p02 | 0.689 | p12 | 0.534 |
| p03 | 0.8 | p13 | 0.473 |
| p04 | 0.83 | p14 | 0.527 |
| p05 | 0.55 | p15 | 0.369 |
| p06 | 0.608 | p16 | 0.432 |
| p07 | 0.609 | p17 | 0.631 |
| p08 | 0.662 | p18 | 0.771 |
| p09 | 0.659 | p19 | 0.783 |
| p10 | 0.616 | p20 | 0.75 |
| Average | 0.643 | | |

Table 5: Final ranks

| Contestant | Mean F1-score |
|---|---|
| custodio18 | 0.685 |
| **murauer18** | **0.643** |
| halvani18 | 0.629 |
| ... | |
| baseline | 0.584 |

## 5 Conclusion

In this paper, a combination of precomputed parameters and a dynamic grid search is used for the task of cross-fandom authorship attribution. Our method relies on traditional character n-gram analysis, which uses specific parameters for each sub-problem, which are found by a standard grid-search approach. Although the methodology is simple in its approach, we were able to reach the second place on the PAN task leader board. Given more time and resources, more parameters could be optimized at runtime rather than pre-calculating a sensible default value.

## References


1. Argamon, S., Juola, P.: Overview of the international authorship identification competition at PAN-2011. In: CLEF (Notebook Papers/Labs/Workshop) (2011)
2. Juola, P.: An Overview of the Traditional Authorship Attribution Subtask. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
3. Kestemont, M., Tschugnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
4. Luyckx, K., Daelemans, W.: The effect of author set size and data size in authorship attribution. Literary and Linguistic Computing 26(1), 35–55 (2011)
5. Markov, I., Stamatatos, E., Sidorov, G.: Improving Cross-Topic Authorship Attribution: The Role of Pre-Processing. Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'2017). (2017)
6. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
7. Sapkota, U., Bethard, S., Montes, M., Solorio, T.: Not All Character N-grams Are Created Equal: A Study In Authorship Attribution. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human language Technologies. pp. 93–102 (Jun 2015)

8. Sapkota, U., Solorio, T., y Gómez, M.M., Bethard, S., Rosso, P.: Cross-Topic Authorship Attribution: Will Out-Of-Topic Data Help? In: Proceedings of the 25th International Conference on Computational Linguistics (COLING'2014). pp. 1228–1237 (Aug 2014)
9. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology 60(3), 538–556 (March 2009)
10. Stamatatos, E.: On the Robustness of Authorship Attribution Based on Character N-Gram Features. Journal of Law & Policy pp. 421–439 (2013)
11. Stamatatos, E.: Authorship Attribution Using Text Distortion. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'2017). pp. 1138–1149. Association for Computational Linguistics (Apr 2017)
12. Stamatatos, E., Rangel, F., Tschuggnall, M., Kestemont, M., Rosso, P., Stein, B., Potthast, M.: Overview of PAN-2018: Author Identification, Author Profiling, and Author Obfuscation. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 9th International Conference of the CLEF Initiative (CLEF 18). Springer, Berlin Heidelberg New York (Sep 2018)