Overview of the Style Change Detection Task at PAN 2019

Eva Zangerle,[1] Michael Tschuggnall,[1] Günther Specht,[1]
Benno Stein,[2] and Martin Potthast[3]

[1]University of Innsbruck, Austria
[2]Bauhaus-Universität Weimar, Germany
[3]Leipzig University, Germany

pan@webis.de https://pan.webis.de

Abstract The task of style change detection aims at segmenting documents into stylistically homogeneous passages. These segments can subsequently be utilized for distinguishing different authors of a document. In this year's PAN style change detection task we asked the participants to answer the following questions for a given document: (1) Is the document written by one or more authors? and, (2) if the document is multi-authored, how many authors have collaborated? The task is performed and evaluated on a dataset compiled from an English Q&A platform, covering a set of diverse topics. The paper in hand introduces the style change detection task and the underlying dataset, surveys the participant's submissions, and analyzes their performance.

1 Introduction

The task of style change detection aims at detecting whether a given document was written by a single author or by multiple authors. Also previous PAN editions aimed to analyze multi-authored documents: In 2016, the task was to identify and group fragments of the document that correspond to individual authors [23]. In 2017, the task was to detect within a first step whether a given document is multi-authored. If this is indeed the case, the next step was to determine the borders at which authorship changes [28]. The results obtained showed that accurately identifying individual authors and their specific contributions within a single document is hard to achieve. In 2018 the task hence was substantially relaxed, asking participants to predict whether a given document is single- or multi-authored [12], recasting it as a binary classification task. Considering the promising results achieved by the submitted approaches, this year's competition continues this line of research and additionally asks participants to predict the number of involved authors.

The remainder of this paper is structured as follows. Section 2 presents previous approaches towards style change detection. Section 3 introduces the style change detection task as part of PAN 2019 along with the underlying dataset and the evaluation

procedure. Section 4 outlines the received submissions, and Section 5 analyzes and compares the achieved results.

2 Related Work

Besides the aforementioned shared tasks and the contributions of its participants, previous work on detecting style change and multi-author documents is still limited. Glover and Hirst [9] were the first to suggest the detection of inconsistencies in collaborative writing, mostly as a tool to aid writers to homogenize their style [11]. In the context of plagiarism detection, we introduced a class of algorithms for intrinsic plagiarism detection [15, 26], which also formed part of our corresponding shared tasks on plagiarism detection [21, 17, 18] and its underlying evaluation framework [20]. A number of participants tackled the problem of intrinsic plagiarism detection, most notably Stamatatos [25], who made use of n-gram profiles to quantify style variation. In their series of four subsequent works, Koppel et al. [13, 14] and Akiva and Koppel [1, 2] develop methods to identify distinct components of multi-author documents and decomposing them in an unsupervised manner. Similarly, we proposed an unsupervised decomposition of multi-author documents based on grammar features [27]. Bensalem et al. [4] follow on Stamatatos' earlier work and use n-grams to identify author style changes. More recent contributions include that of Gianella [8] who employs stochastic modeling to split a document by authorship, and that of Rexha et al. [22] who extend their previous work to predict the number of authors who wrote a text. Dauber et al. [7] propose an approach to tackle authorship attribution on multi-author documents. Aldebei et al. [3] as well as Sarwar et al. [24] use hidden Markov models and basic stylometric features to build a so-called co-authorship graph.

3 Style Change Detection Task

3.1 Task Definition

The PAN 2019 style change detection task is defined as follows. Given a document, participants have to apply intrinsic style analyses (i.e., exploit the given document as the only source) to answer, consecutively, the following two questions:

1. Do style changes exist? (which we interpret as the fact that the document is multi-authored)
2. If the document is multi-authored, how many authors did collaborate?

Figure 1 depicts three example documents and the respective answers to the above questions. The first document (left) is considered as a single-authored document since no style changes are detected. The two other documents (middle and right) contain style changes, with the second document having been authored by two authors and the third document by three authors respectively;note that an author may have worked on one or more passages.

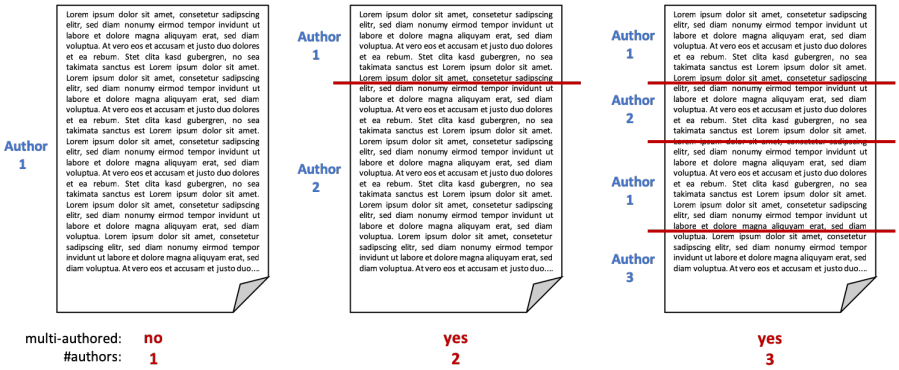


Figure 1. Exemplary documents that illustrate different style change situations.

3.2 Dataset Construction

The datasets provided for training, validation, and test were curated using the StackExchange Q&A platform,[1] which provides a network of Q&A-websites on a wide variety of topics. Based on a publicly available dump we extracted user questions and answers from all available sites such that a diverse set of topics is covered, ranging from cooking over philosophy to bitcoin technology. Each site (topic) contains several subtopics, identified by user-created tags. Within this year's task only English documents were considered.

Based on this set of questions and answers (referred to as documents in what follows), we remove the following documents (or parts thereof) to ensure that the dataset contains only valid, parsable documents of sufficient length:

- short documents consisting of only a few words or symbols
- documents edited by users other than the original author
- external URLs
- embedded images
- code snippets (particularly, for the StackOverflow topic)
- bullet lists (including content)
- block quotes
- texts containing Arabic characters (particularly, for the Islam topic)

If a document contains less than three sentences after having applied the above cleansing steps, we remove it from the dataset.

Based on the resulting set of documents, we compute the final dataset as follows. For each subtopic of a main topic, we assemble a document based on the questions and answers belonging to the same subtopic. This ensures that authors cannot be distinguished based on the topic or content discussed. Particularly, the datasets per topic

[1] https://stackexchange.com/

Table 1. Parameters for constructing the style change detection datasets.

Parameter	Configurations
Number of style changes	0 - 10
Number of collaborating authors	1 - 5
Document length	300 - 1 500 tokens
Change positions	at the end of paragraph, within paragraph, mixed
Segment length distribution	100 - 1 500

are assembled by varying the parameters shown in Table 1. Given that the first part of the task—detecting whether a document contains a style change or not—is a binary classification problem, we ensure that the two classes are balanced, i.e., 50% are single-authored and 50% are multi-authored documents. For single-author documents, we assemble one or more documents, resulting in a document containing 300-1 500 tokens. For multi-author documents, we vary not only the number of authors, but also the number of style changes within the final document between 1 and 10, which is done by combining multiple documents of multiple authors, varying the parameters stated in Table 1. The resulting dataset contains a total of 5 028 documents.

The compiled dataset was split into subsets for training, validation, and test. The employs approximate stratified random sampling, yielding a 50% training, a 25% validation, and a 25% test split. Both the training and the validation set were released to participants while the test set was withheld for the final evaluation. An overview of the datasets is given in Table 2, where we list the number of documents for the different number of authors (as absolute numbers and relative shares) and the average number of tokens per document for single- and multi-authored documents. The three datasets feature a similar distribution in terms of the number of authors per document and the average document length.

3.3 Evaluation Framework

For computing the desired evaluation measures (see Section 3.4) the participants had the choice either to use the provided evaluation script or to work directly on the TIRA

Table 2. Overview of the datasets. SA and MA stand for single-authored and multi-authored documents; text length is measured by the average number of tokens per document.

Dataset	Documents	Authors					Length	
		1	2	3	4	5	SA	MA
Training	2 546	1 273 50.00%	325 12.76%	313 12.29%	328 12.88%	307 12.06%	977	1 604
Validation	1 272	636 50.00%	179 14.07%	152 11.95%	160 12.58%	145 11.40%	957	1 582
Test	1 210	605 50.00%	147 12.15%	144 11.90%	159 13.15%	155 12.81%	950	1 627

platform [10, 19]. By deploying their system on the TIRA platform, the participants can see the obtained performance values on the training and the validation set. For the final evaluation (done with the test set), participants were asked to mark their favored run to be considered in the overall ranking (see Section 5).

3.4 Performance Measures

For the comparison of the submitted approaches we report both the achieved performances for the subtasks in isolation and their combination as a staged task.

For the first subtask, the binary classification of single- versus multi-authorship, the accuracy metric is used, i.e., we compute the fraction of correctly classified documents. For the second subtask, the prediction of the number of authors, we compute not only the number of correctly classified documents regarding the number of authors, but we also consider the extent to which the prediction differs from the true class. E.g., classifying a document authored by 5 authors as a 4-author document is here regarded as a better result than classifying it as a 2-author document. For this "distance" between the predicted and the true author number we employ the *Ordinal Classification Index (OCI)* [5], which is based on the confusion matrices resulting from the classification tasks. It yields a value between 0 and 1, with 0 being a perfect prediction.

Finally, we combine the two measures into a single rank measure, where we put equal weight on accuracy and OCI (recall that OCI is an error measure):

$$\text{rank} = \frac{\text{accuracy} + (1 - \text{OCI})}{2}$$

4 Survey of Submissions

The style change detection task at PAN 2019 received two submissions, which are outlined in the following.

4.1 Clustering-based Style Change Detection

The approach by Nath [16] is based on the combination of two clustering approaches. The authors propose to first divide documents into windows of similar size. For each of these windows, a representation based on the 50 most frequent tokens is computed. In this regard, the authors resort to the normalized frequency of the 50 most frequent tokens in each document to represent each window. Based on the pair-wise distances of documents, two clustering approaches are applied: Threshold-based Clustering (TBC) and Window Merge Clustering (WMC), where the assumption is that windows assigned to the same cluster have been written by the same author. Threshold-based clustering operates on the distance matrix of individual windows, and clusters are formed by iterating over the list of pairwise window distances (sorted by distance in ascending order). For each window pair, either these windows are joint to a new cluster, they extend an existing cluster (in which one of the two windows is already contained in), or two clusters are merged based on a given threshold. The Window Merge Clustering follows the

hierarchical clustering paradigm, where pairs of clusters (initially one window per cluster) are joint iteratively until a certain threshold in terms of cluster similarity is reached. The authors propose to combine the results of the two clustering approaches by taking their combined minimum. Furthermore, the authors employ statistics about the number of duplicated sentences contained per number of authors in the training set, and they add this information also to the clustering procedure. The authors report that utilizing the TBC-based method achieves the best results (cf. Section 5).

4.2 Feed-forward Network-based Style Change Detection

Zuo et al. [30] propose to split the given task into two subtasks, which are dealt with individually. For the first subtask, the binary classification of single- vs. multi-authored documents, the authors propose a multi-layer perceptron (MLP) with a single hidden layer. Each document is represented by its TFIDF-weighted word vector. For the second task, the authors compute a number of established features at the paragraph-level for each of the multi-authored documents, which is based on the winning submission of the PAN style change detection task 2018 [29]: lexical features (e.g., POS tags, number of sentences or token length), contracted word forms, British or American English, frequent words, and readability scores (e.g., Flesch-Kincaid grade level or Gunning Fog Index). Furthermore, they add the TFIDF-weighted words. Based on this representation, they propose to cluster segments by applying an ensemble of three clustering approaches to predict the number of authors for a given document: k-means clustering based on the TFIDF-representation of segments, hierarchical clustering with all features, and an MLP classifier.

5 Evaluation Results

The submissions were evaluated on the TIRA experimentation platform. For comparison, we also evaluated two baselines:

1. BASELINE-RND. An enhanced guessing baseline that exploits the data set statistics with respect to document lengths and style changes. In particular it is assumed that longer documents tend to contain more style changes shorter documents.
2. BASELINE-C99. A baseline that employs the C99 text segmentation algorithm to predict the number of segments [6].

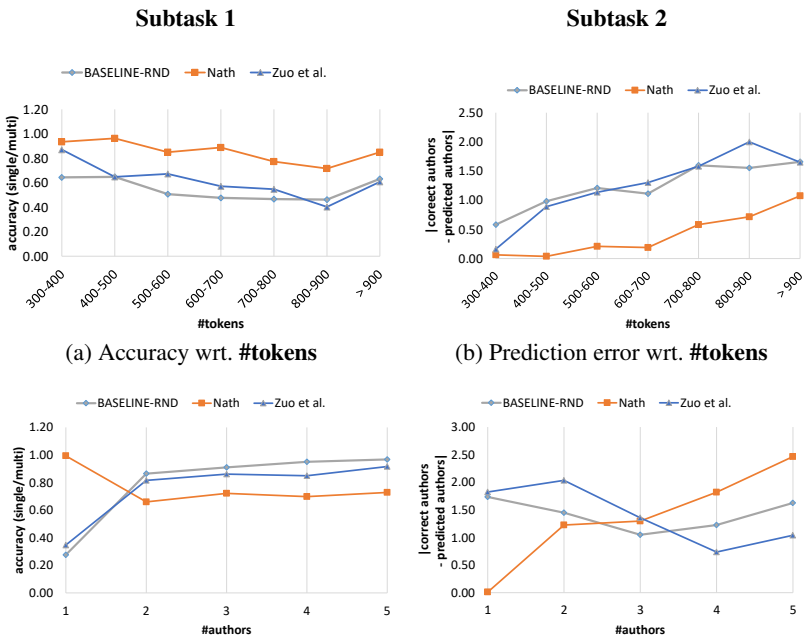
Table 3 shows the participants' evaluation results, where we list accuracy, the ordinal classification index, and the proposed overall rank measure. Both submission surpass the baselines with respect to every measure, where Nath achieved the highest rank due to the high accuracy in distinguishing single- from multi-author documents. Zuo et al. achieved a slightly better OCI value, i.e., they could better estimate the true number of authors in multi-author documents. In terms of runtime, the clustering-based approach by Nath significantly outperforms the deep learning approach by Zuo et al.

To get a deeper understanding of the performance of the submitted approaches, we also analyze the results with respect to certain document characteristics. For Subtask 1

Table 3. Overall results for the style change detection task.

Participant	Accuracy	OCI	Rank	Runtime
Nath	0.848	0.865	0.491	00:02:23
Zuo et al.	0.604	0.809	0.398	00:25:50
BASELINE-RND	0.600	0.856	0.372	-
BASELINE-C99	0.582	0.882	0.350	00:00:30

we compute the respective accuracy values, and for Subtask 2 we compute the absolute prediction errors. Figures 2 and 3 overview the results. Figures 2a and 2b show the average performance of the approaches and the random baseline grouped by document length in tokens. We observe that Nath is able to outperform the approach by Zuo et al. over all document lengths. In Figures 2c and 2d, by analyzing the results based on the number of authors, we observe that the approach by Nath is able to exploit documents with a lower number of authors better for Subtask 2, while Zuo et al. are able to exploit documents with more than two authors better and predict the number of authors more accurately. For Subtask 1, Nath achieves better results than both Zuo et al. and the random baseline for single-authored documents. Looking into the lengths of segments of individual authors (cf. Figures 3a and 3b), we observe quite diverging behaviors for the submitted approaches for Subtask 1, whereas for Subtask 2 the differences are more



(a) Accuracy wrt. **#tokens**

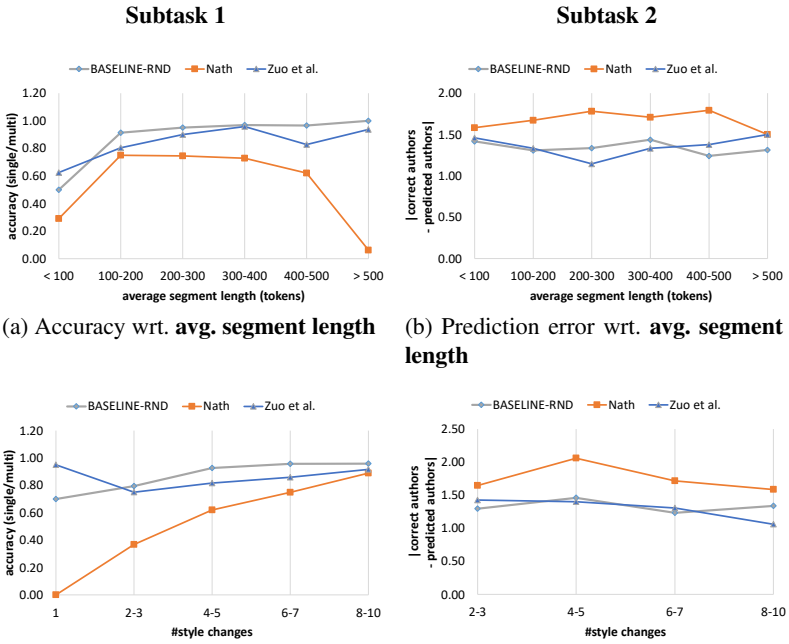
(b) Prediction error wrt. **#tokens**

(c) Accuracy wrt. **#authors**

(d) Prediction error wrt. **#authors**

Figure 2. Detailed evaluation results of the approaches with respect to the number of tokens and the number of authors for Subtask 1 (left column) and Subtask 2 (right column).

subtle: With regard to the number of style changes in a document (Figures 3c and 3d), for documents with a lower number of style changes the approach by Nath achieves a lower accuracy as well as a higher deviation in terms of the prediction of the author number than does the approach by Zuo et al.



(a) Accuracy wrt. **avg. segment length**

(b) Prediction error wrt. **avg. segment length**

(c) Accuracy wrt. **#style changes**

(d) Prediction error wrt. **#style changes**

Figure 3. Detailed evaluation results of the approaches with respect to the average segment length and the number of style changes for Subtask 1 (left column) and Subtask 2 (right column).

Bibliography

[1] Akiva, N., Koppel, M.: Identifying distinct components of a multi-author document. In: Memon, N., Zeng, D. (eds.) 2012 European Intelligence and Security Informatics Conference, EISIC 2012, Odense, Denmark, August 22-24, 2012. pp. 205–209. IEEE Computer Society (2012), https://doi.org/10.1109/EISIC.2012.16

[2] Akiva, N., Koppel, M.: A generic unsupervised method for decomposing multi-author documents. JASIST 64(11), 2256–2264 (2013), https://doi.org/10.1002/asi.22924

[3] Aldebei, K., He, X., Jia, W., Yeh, W.: SUDMAD: sequential and unsupervised decomposition of a multi-author document based on a hidden markov model. JASIST 69(2), 201–214 (2018), https://doi.org/10.1002/asi.23956

[4] Bensalem, I., Rosso, P., Chikhi, S.: Intrinsic plagiarism detection using n-gram classes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1459–1464. Association for Computational Linguistics, Doha, Qatar (Oct 2014), https://www.aclweb.org/anthology/D14-1153

[5] Cardoso, J., Sousa, R.: Measuring the performance of ordinal classification. International Journal of Pattern Recognition and Artificial Intelligence 25(08), 1173–1195 (2011)

[6] Choi, F.Y.Y.: Advances in domain independent linear text segmentation. In: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference. pp. 26–33. NAACL 2000, Association for Computational Linguistics, Stroudsburg, PA, USA (2000), http://dl.acm.org/citation.cfm?id=974305.974309

[7] Dauber, E., Overdorf, R., Greenstadt, R.: Stylometric authorship attribution of collaborative documents. In: Dolev, S., Lodha, S. (eds.) Cyber Security Cryptography and Machine Learning - First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, 2017, Proceedings. Lecture Notes in Computer Science, vol. 10332, pp. 115–135. Springer (2017), https://doi.org/10.1007/978-3-319-60080-2_9

[8] Giannella, C.: An improved algorithm for unsupervised decomposition of a multi-author document. JASIST 67(2), 400–411 (2016), https://doi.org/10.1002/asi.23375

[9] Glover, A., Hirst, G.: Detecting Stylistic Inconsistencies in Collaborative Writing, pp. 147–168. Springer London, London (1996), https://doi.org/10.1007/978-1-4471-1482-6_12

[10] Gollub, T., Stein, B., Burrows, S.: Ousting ivory tower research: towards a web framework for providing experiments as a service. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. pp. 1125–1126. ACM (2012)

[11] Graham, N., Hirst, G., Marthi, B.: Segmenting documents by stylistic character. Natural Language Engineering 11(4), 397–415 (2005), https://doi.org/10.1017/S1351324905003694

[12] Kestemont, M., Tschuggnall, 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: Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al. pp. 1–25 (2018)

[13] Koppel, M., Akiva, N., Dershowitz, I., Dershowitz, N.: Unsupervised decomposition of a document into authorial components. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 1356–1364. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), https://www.aclweb.org/anthology/P11-1136

[14] Koppel, M., Akiva, N., Dershowitz, I., Dershowitz, N.: Unsupervised decomposition of a document into authorial components. In: Lin, D., Matsumoto, Y., Mihalcea, R. (eds.) The

49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA. pp. 1356–1364. The Association for Computer Linguistics (2011), http://www.aclweb.org/anthology/P11-1136

[15] Meyer zu Eißen, S., Stein, B., Kulig, M.: Plagiarism Detection without Reference Collections. In: Decker, R., Lenz, H. (eds.) Advances in Data Analysis. Selected papers from the 30th Annual Conference of the German Classification Society (GFKL 2006). pp. 359–366. Studies in Classification, Data Analysis, and Knowledge Organization, Springer, Berlin Heidelberg New York (2007)

[16] Nath, S.: UniNE at PAN-CLEF 2019: Style Change Detection by Threshold Based and Window Merge Clustering Methods. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)

[17] Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: Braschler, M., Harman, D., Pianta, E. (eds.) Working Notes Papers of the CLEF 2010 Evaluation Labs (Sep 2010), http://www.clef-initiative.eu/publication/working-notes

[18] Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection. In: Petras, V., Forner, P., Clough, P. (eds.) Working Notes Papers of the CLEF 2011 Evaluation Labs (Sep 2011), http://www.clef-initiative.eu/publication/working-notes

[19] Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the reproducibility of PAN's shared tasks. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 268–299. Springer (2014)

[20] Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Huang, C.R., Jurafsky, D. (eds.) 23rd International Conference on Computational Linguistics (COLING 10). pp. 997–1005. Association for Computational Linguistics, Stroudsburg, Pennsylvania (Aug 2010)

[21] Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 2009). pp. 1–9. CEUR-WS.org (Sep 2009), http://ceur-ws.org/Vol-502

[22] Rexha, A., Klampfl, S., Kröll, M., Kern, R.: Towards a more fine grained analysis of scientific authorship: Predicting the number of authors using stylometric features. In: Mayr, P., Frommholz, I., Cabanac, G. (eds.) Proceedings of the Third Workshop on Bibliometric-enhanced Information Retrieval co-located with the 38th European Conference on Information Retrieval (ECIR 2016), Padova, Italy, March 20, 2016. CEUR Workshop Proceedings, vol. 1567, pp. 26–31. CEUR-WS.org (2016), http://ceur-ws.org/Vol-1567/paper3.pdf

[23] Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN'16—New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. In: Fuhr, N., Quaresma, P., Larsen, B., Gonçalves, T., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 16). Springer, Berlin Heidelberg New York (Sep 2016)

[24] Sarwar, R., Yu, C., Nutanong, S., Urailertprasert, N., Vannaboot, N., Rakthanmanon, T.: A scalable framework for stylometric analysis of multi-author documents. In: Pei, J., Manolopoulos, Y., Sadiq, S.W., Li, J. (eds.) Database Systems for Advanced Applications - 23rd International Conference, DASFAA 2018, Gold Coast, QLD, Australia, May 21-24,

2018, Proceedings, Part I. Lecture Notes in Computer Science, vol. 10827, pp. 813–829. Springer (2018), https://doi.org/10.1007/978-3-319-91452-7_52
[25] Stamatatos, E.: Intrinsic Plagiarism Detection Using Character n-gram Profiles. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 38–46. Universidad Politécnica de Valencia and CEUR-WS.org (Sep 2009), http://ceur-ws.org/Vol-502
[26] Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic Plagiarism Analysis. Language Resources and Evaluation (LRE) 45(1), 63–82 (Mar 2011)
[27] Tschuggnall, M., Specht, G.: Automatic decomposition of multi-author documents using grammar analysis. In: Klan, F., Specht, G., Gamper, H. (eds.) Proceedings of the 26th GI-Workshop Grundlagen von Datenbanken, Bozen-Bolzano, Italy, October 21st to 24th, 2014. CEUR Workshop Proceedings, vol. 1313, pp. 17–22. CEUR-WS.org (2014), http://ceur-ws.org/Vol-1313/paper_4.pdf
[28] Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the author identification task at pan-2017: style breach detection and author clustering. In: Working Notes Papers of the CLEF 2017 Evaluation Labs/Cappellato, Linda [edit.]; et al. pp. 1–22 (2017)
[29] Zlatkova, D., Kopev, D., Mitov, K., Atanasov, A., Hardalov, M., Koychev, I., Nakov, P.: An ensemble-rich multi-aspect approach for robust style change detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) CLEF 2018 Working Nots of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR-WS.org (Sep 2018)
[30] Zuo, C., Zhao, Y., Banerjee, R.: Style Change Detection with Feedforward Neural Networks Notebook for PAN at CLEF 2019 . In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers. CEUR-WS.org (Sep 2019)