Overview of the Style Change Detection Task at PAN 2020

Eva Zangerle,[1] Maximilian Mayerl,[1] Günther Specht,[1]
Martin Potthast,[2] and Benno Stein[3]

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

pan@webis.de http://pan.webis.de

Abstract The goal of style change detection is to identify text positions within a multi-author document at which the author switches. Detecting these positions is a crucial part of processing multi-author documents for purposes of authorship identification. In this year's PAN style change detection task, we asked the participants to answer the following questions for a given document: (1) Given a document, was it written by multiple authors? (2) For each pair of consecutive paragraphs in a given document, is there a style change between these paragraphs? The task is performed and evaluated on two datasets compiled from an English Q&A platform, which differ in their topical breadth (i.e., the number of different topics that are covered in the documents contained). The paper in hand introduces style change detection as a task and its underlying dataset, surveys the participants' submissions, and analyzes their performance.

1 Introduction

The task of style change detection aims at detecting positions of author changes within a collaboratively written text. Previous PAN editions paved the way for PAN'20 by analyzing multi-authored documents for style changes. This includes the identification and clustering of text segments by author in 2016 [25]. In 2017, participants were asked to detect whether a given document has been authored by multiple authors, and in that case, to determine the boundaries at which authorship changes [34]. The results showed that accurately determining such boundaries is still beyond current capabilities. Hence, in 2018, the task was relaxed by formulating it as a binary classification problem, where the goal was to predict whether a given document is written by a single author or multiple authors [15]. At PAN 2019, this classification task was extended to also predict the number of authors for multi-author documents [35]. In 2020, the task was steered back into its original direction: Participants were asked to detect whether a document was authored by one or multiple authors, and the positions of style changes at the paragraph-level.

The remainder of this paper is structured as follows. Section 2 presents previous style change detection approaches. Section 3 introduces the style change detection task as part of PAN 2020, along with the datasets and evaluation measures employed. Section 4 summarizes the received submissions, and Section 5 analyzes and compares the achieved results and Section 6 concludes the paper.

2 Related Work

Style change detection is closely related to the fields of stylometry, plagiarism detection, and text segmentation. All of them have in common that they rely on intrinsic stylometric analyses of documents, without referring to external documents or corpora for comparison. Hence, stylistic profiles are created that are based on lexical features like character n-grams (e.g., [29, 20]), word frequencies (e.g., [11]) and average word/sentence lengths (e.g., [36]), syntactic features like part-of-speech tag frequencies/structures (e.g., [32]), and structural features such as indentation usage (e.g., [36]).

One of the earliest works on style change detection by analyzing stylometric features to detect author boundaries is by Glover and Hirst [9], which aims at identifying inconsistencies of writing style in collaborative documents. Meyer zu Eißen and Stein [21, 31, 30] were the first to investigate intrinsic plagiarism detection based on style change detection using word frequency classes. Koppel et al. [18, 19] and Akiva and Koppel [1, 2] propose an unsupervised method to decompose multi-author documents into authorial threads by applying clustering methods on lexical features. Tschuggnall et al. [33] proposed an unsupervised decomposition approach based on grammar tree representations, whereas Rexha et al. [24] use stylistic features to predict the number of authors who wrote a text. Bensalem et al. [3] rely on n-grams to identify author style changes. Gianella [8] employs Bayesian modeling to split a document by authorship, Further approaches include that of Graham et al. [10], who utilize neural networks with several stylometric features.

At PAN 2017 [34], the goal was to find the exact positions of authorship changes. This task was mostly tackled by using stylometric features to characterize sentences and paragraphs and detecting boundaries by computing similarities [14, 16], or by applying outlier detection [26]. For the binary classification task whether a document is single-author or multi-author at PAN 2018 [15], the best performing system is a stacking ensemble classifier based on lexical and syntactical features extracted via multiple sliding window approaches [37]. Alternatively, deep learning approaches such as convolutional neural networks that operate on a character input [28] and recurrent neural networks operating on parse tree features [12] have been proposed. Other participants used stylometric features to compute the similarity of sentences and paragraphs to find homogeneous text segments that correspond to an individual author [17], or as input to a binary ensemble classifier [27]. In addition, to predict the number of authors of a multi-author document, at PAN 2019, Nath [22] uses two clustering approaches based on token frequencies, whereas Zuo et al. [38] use a classification ensemble based on lexical, syntactic and word frequency features.

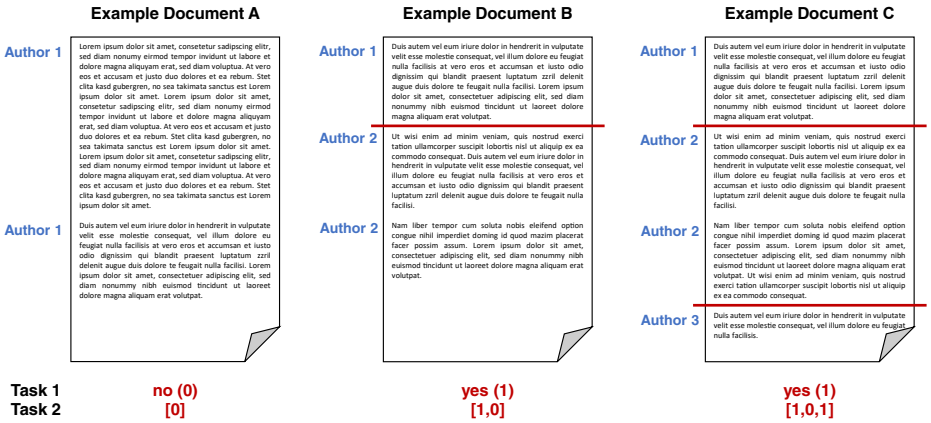


Figure 1. Exemplary documents that illustrate different style change situations and the expected solution for Task 1 (single- or multi-authored document) and Task 2 (position of style changes).

3 Style Change Detection Task

This section introduces the style change detection task, the dataset constructed for it, the performance measures employed, and our evaluation platform on which the task has been deployed.

3.1 Task Definition

The goal of style change detection is to segment documents into stylistically homogeneous passages, which can subsequently be utilized for authorship identification and attribution. Hence, style change detection aims at identifying text positions within a given multi-author document at which the author changes. Beforehand, it must be determined whether such a segmentation is necessary at all by checking whether the document in question is indeed a multi-author document. We there study the following two tasks:

- *Task 1.* Given a document, is the document written by multiple authors?
- *Task 2.* Given a sequence of paragraphs of a (supposedly) multi-author document, is there a style change between any of the paragraphs?

Figure 1 illustrates the possible scenarios and the expected output for the two tasks. Document A does not contain any style changes and hence, was authored by a single author; Document B contains a single style change between Paragraphs 1 and 2, and Document C contains two style changes. In order to render the task more feasible, we ensure that all documents comprised in our evaluation dataset are written in the same language (English), and that style changes occur only between paragraphs, not within them (i.e., a single paragraph is always authored by a single author and does not contain any style changes). Moreover, the documents may contain between zero and ten style changes, resulting from at most three different authors.

3.2 Dataset Construction

As with last year's task [35], the datasets are based on data taken from StackExchange. StackExchange is a popular network of Q&A sites, covering a wide range of topics. We used a dump of the questions and answers on the various StackExchange sites[1] as the basis for our datasets. As a first step, we cleaned this data as follows:

- Removal of questions and answers that contain fewer than 30 characters.
- Removal of questions and answers that were edited by a user different from the one who originally wrote them.
- Removal of the following items within all questions and answers: images, URLs, code snippets, block quotes, and bullet lists along with their contents.

After cleaning the data, we constructed two datasets differing in the number of topics covered to also enable an investigation into how well approaches are able to deal with topical diversity. For the first dataset, called `dataset-narrow`, we used only questions and answers belonging to a subset of the StackExchange sites that deal with topics related to computer technology.[2] For the second dataset, called `dataset-wide`, we used a subset of sites covering a wide range of different topics,[3] including technology, economics, literature, philosophy, and mathematics.

For every site in the subset of sites that was used for the creation of a given dataset, we grouped together all questions and answers written by the same user and split them into paragraphs, removing paragraphs with fewer than 100 characters. This yielded a list of paragraphs for every user on a particular site. In a next step, we constructed the documents making up the dataset. Each dataset contains an equal number of single-author and multi-author documents. For single-author documents, we selected a random user and drew paragraphs from the paragraph list of that user until the document had a sufficient length (randomly chosen to be between 1,000 and 3,000 words). For multi-author documents, we first randomly chose whether the document should have two or three authors. Then, we randomly constructed a structure for the document (i.e., a sequence of author changes for a set of paragraphs). Based on that, we randomly chose distinct authors from our list of users and drew paragraphs from their paragraph lists until the document had the predetermined structure and the chosen length (again, randomly chosen to be between 1,000 and 3,000 words).

The resulting documents were then split into training, validation, and test sets with approximately 50% of the documents being assigned to the training set, and 25% each being assigned to the validation and test sets. The procedure we used for splitting ensured that every subset contains approximately the same number of single-author and multi-author documents. Finally, we filtered all documents based on their language. As

[1] See https://archive.org/details/stackexchange

[2] `dataset-narrow` contains questions and answers from the following sites: Code Review, Computer Graphics, CS Educators, CS Theory, Data Science, DBA, DevOps, Game Dev, Network Engineering, Raspberry Pi, Serverfault.com, Superuser.com.

[3] `dataset-wide` contains questions and answers from the following sites: Academia, Astronomy, Bicycles, Biology, Buddhism, Code Review, Coffee, DBA, Earth Science, Economics, Engineering, Fitness, History, Interpersonal, Linguistics, Literature, Mathoverflow.net, Outdoors, Philosophy, Serverfault.com, Skeptics, Sports, Travel, Workplace, Worldbuilding.

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

Parameter	Configurations
Number of collaborating authors	1-3
Number of style changes	0-10
Document length	1,000-3,000
Change positions	between paragraphs
Document language	English

Table 2. Dataset overview. Text length is measured as average number of tokens per document.

Dataset	Documents	Documents / #Authors			Length / #Authors		
		1	2	3	1	2	3
Narrow-Train	3,418	1,709 50.00%	854 24.99%	855 25.01%	11,872	11,659	11,717
Narrow-Valid.	1,713	855 49.91%	415 24.23%	443 25.86%	11,931	11,996	11,605
Narrow-Test	1,701	852 50.09%	426 25.04%	423 24.87%	11,715	11,637	11,708
Wide-Train	8,030	4,025 50.12%	1,990 24.78%	2,015 25.09%	11,751	12,191	12,095
Wide-Valid.	4,019	2018 50.21%	969 24.11%	1,032 25.68%	12,113	12,113	12,069
Wide-Test	3,995	2,004 50.16%	987 24.70%	1,004 25.13%	12,242	12,015	11,729

we want our datasets to consist only of English documents, we removed all documents where at least one paragraph was identified as being written in a language other than English. For this, we used the Python library langdetect.[4] A summary of the parameters for both datasets is given in Table 1. Table 2 shows an overview of the created datasets, including the number of contained documents as well as the average document lengths, partitioned by the number of authors.

For development, participants are provided with the documents and ground truth information. For each training and validation document, we provided the number of authors, the StackExchange site the texts were gathered from, the order of the authors within the document, the positions of the style changes, and whether the document was indeed multi-authored.

3.3 Performance Measures

To evaluate and compare the submitted approaches, we report both, the achieved performance for the individual subtasks, and their combination as a staged task. Furthermore,

[4]https://pypi.org/project/langdetect/

we evaluate the approaches on both datasets individually. Submissions are evaluated using the F_α-Measure for each document, where $\alpha = 1$ equally weighs the harmonic mean between precision and recall. For Task 1, we compute the average F_1 measure across all documents, and for Task 2, we use the micro-averaged F_1 measure across all documents. The submissions for the two datasets are evaluated independently and the resulting F_1 measures for the two tasks are averaged across datasets.

3.4 Evaluation Framework

To ensure the reproducibility of the submitted solutions, participants were asked to deploy their software on our TIRA platform [23]. Each participant was assigned a virtual machine on TIRA, where the software had to be setup with the only constraint of being executable via a POSIX command. The web frontend of TIRA allows for configuring pieces of software that are deployed within an participant's virtual machine, and to remotely execute them via an appropriate command. This enabled participants both, to test their software on the freely available training and validation datasets, as well as to self-evaluate their software on the test dataset, which is not freely accessible. TIRA prevents direct access by participants by moving the virtual machine into a secure sandbox before enabling the a deployed software to process a test dataset. This way, TIRA enables blind evaluation, thus foreclosing optimization against the test data. Runs resulting from processing the training, validation, or test data can be evaluated using the aforementioned evaluation measure at the click of a button.

4 Survey of Submissions

For this year's edition of the style change detection task, we received three submissions. However, only two participating teams submitted a working notes paper. In the following, we describe the approaches used in those submissions.

4.1 Mixed Style Feature Representation and B-maximal Clustering

The approach developed by Castro-Castro et al. [6] makes use of a variant of B_0-maximal clustering to solve the style change detection task. First, they formulate a representation for a paragraph as a set of 185 stylometric features, consisting of character-based, lexical, and syntactic features, but excluding features which explicitly capture the semantics of the given text. The features are divided into three different categories: boolean features, features consisting of a single floating-point number, and features consisting of vectors of numbers. For each of these categories, a comparison criterion is defined which expresses whether a given feature of two different paragraphs is "similar" or not. Then, the similarity between two paragraphs is defined to be the number of similar features between them.

Based on this, B_0-maximal clustering is performed to group the paragraphs in a document into clusters, where every cluster is regarded to be one author. This clustering approach assigns all paragraphs with a similarity larger than a defined threshold to the same cluster. Since this makes it possible for a paragraph to be assigned to multiple

Table 3. Overall results for the style change detection task, ranked by average F_1.

Participant	Task1 F_1	Task2 F_1	Avg. F_1
Iyer and Vosoughi	0.6401	0.8567	0.7484
Castro-Castro et al.	0.5399	0.7579	0.6489
Nath	0.5204	0.7526	0.6365
Baseline (random)	0.5007	0.5001	0.5004

clusters, and hence to multiple authors, a basic approach for deciding to which cluster a paragraph will be assigned is proposed. In such a case, from all possible candidate clusters, the one which contains the paragraph that occurs earliest in the document is chosen. Thus, all the paragraphs in a document are assigned to authors. From this, the tasks posed in this year's style change detection task are solved as follows: For the first task, it is simply checked whether the clustering has produced more than one cluster. For the second task, the positions in the document are identified, where consecutive paragraphs were assigned to different clusters.

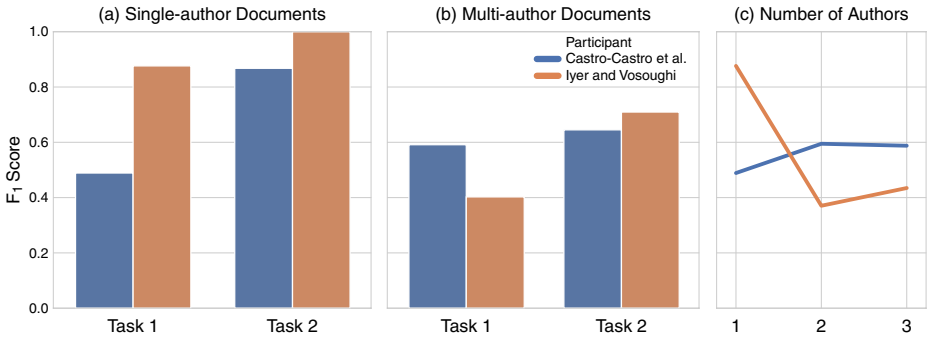
4.2 Style Change Detection Using BERT

The approach of Iyer and Vosoughi [13] is based on using Google's BERT language model [7] as a feature extractor, and random forests as a classifier. First, the documents contained in the dataset are split into sentences, and every sentence is fed to BERT, taking the outputs of the last four BERT layers to represent a given sentence. Since the size of the feature matrix produced by this depends on the number of tokens in a sentence, the values along the length dimension are summed to obtain a feature matrix of a fixed length. After this, representations are formulated for consecutive pairs of paragraphs (to solve the second task), and the whole document (to solve the first task), based on the representations of sentences, by summing (paragraphs) or averaging (whole documents) the feature values of the sentences that make up the paragraph or document. These feature representations are then used to trin random forest models for both tasks.

5 Evaluation Results

The results for the participants submissions as well as a random baseline are given in Table 3. The table shows the F_1 score for both tasks, as well as the overall average scores. Both participants' submissions significantly outperformed the baseline with respect to individual and overall score. The best-performing submission is the one by Iyer and Vosoughi, which achieved the best scores in both tasks as well as in the overall average score. The approach developed by Castro-Castro et al. performs significantly better than the random baseline, but also significantly worse than the approach by Iyer and Vosoughi, forming a middle ground. The approach of Nath[5] performs only slightly worse than that of Castro-Castro et al.

[5]The participant did not submit their working notes and was hence omitted from further analysis.



(d) Topic Diversity

Participant	**Task 1 Narrow**	**Task 1 Wide**	**Task 2 Narrow**	**Task 2 Wide**
Iyer and Vosoughi	0.7042	0.5760	0.8823	0.8310
Castro-Castro et al.	0.5379	0.5419	0.8242	0.6915

Figure 2. Overall performance of the submitted approaches regarding (a) single-author documents, (b) multi-author documents, and (c) dependent on the number of authors per document. (d) The table shows the F_1 scores achieved dependent on topic diversity.

We further analyzed the performance of both approaches with regard to the specific properties of the documents in our datasets. First, compared both approaches with respect to single-author versus multi-author documents. The results for this analysis are shown in Figures 2a and b. The approach by Iyer and Vosoughi reaches an F_1 score of almost 1.0 for Task 2 on single-author documents. This suggests that it may be beneficial for them to reduce their approach to one model predicting style changes between paragraphs, and then calculating predictions for Task 1 based on the output of that model (i.e., predicting a document to be multi-author if and only if there was at least one style change predicted between the paragraphs of that document). Another point to note is that the approach by Castro-Castro et al. performed best for Task 1 on multi-author documents. This suggests that their model is especially well-suited for detecting documents that have been written by more than one author. Moreover, we analyzed how the performance for both submitted approaches changes depending on the number of authors. The results for this analysis are shown in Figure 2c, confirming that the approach of Castro-Castro et al. performs better for multi-author documents, regardless of whether the number of authors is two or three. It is interesting that Castro-Castro et al.'s approach improves for multi-author documents, whereas that of Iyer and Vosoughi performs best for single-author documents, exerting a sharp drop in performance when a document is written by multiple authors.

Finally, we analyzed the performance of the participants' approaches dependent on topic diversity (see Figure 2d). In most cases, we found a significant difference in performance between both datasets. The exception to this is the approach by Castro-Castro et al. on Task 1, where the performance on the narrow and wide datasets are almost identical. In all other cases, the performance differs significantly, with performance on the narrow dataset being higher than on the wide dataset, implying that dealing with documents of a diverse topical variety renders the task more difficult.

6 Conclusion

In the 2020 edition of the PAN style change detection task, we asked participants to answer the following questions for a given document: (1) Given a document, was it written by multiple authors? (2) For each pair of consecutive paragraphs in a given document, is there a style change between these paragraphs? Three participants submitted their systems and two participants submitted a working notes paper. The two approaches differed fundamentally, the best-performing system relying on semantic features (i.e., BERT embeddings), while the second-best approach focused on syntactic ones. Future challenges include finding the exact position of authorship changes beyond the paragraph level, and assigning paragraphs to individual authors.

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, URL 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, URL https://doi.org/10.1002/asi.22924

[3] 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://doi.org/10.3115/v1/D14-1153, URL https://www.aclweb.org/anthology/D14-1153

[4] Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.): CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland, CEUR Workshop Proceedings, CEUR-WS.org (2017), ISSN 1613-0073, URL http://ceur-ws.org/Vol-1866/

[5] Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.): CLEF 2018 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Avignon, France, CEUR Workshop Proceedings, CEUR-WS.org (2018), ISSN 1613-0073

[6] Castro-Castro, D., Rodríguez-Lozada, C.A., noz, R.M.: Mixed Style Feature Representation and B-maximal Clustering for Style Change Detection. In: Cappellato, L., Ferro, N., Névéol, A., Eickhoff, C. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org (Sep 2020)

[7] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[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, URL https://doi.org/10.1002/asi.23375

[9] Glover, A., Hirst, G.: Detecting stylistic inconsistencies in collaborative writing. In: The New Writing Environment, pp. 147–168, Springer (1996)

[10] Graham, N., Hirst, G., Marthi, B.: Segmenting documents by stylistic character. Natural Language Engineering **11**(4), 397–416 (2005)

[11] Holmes, D.I.: The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing **13**(3), 111–117 (1998)

[12] Hosseinia, M., Mukherjee, A.: A Parallel Hierarchical Attention Network for Style Change Detection—Notebook for PAN at CLEF 2018. In: [5]
[13] Iyer, A., Vosoughi, S.: Style Change Detection Using BERT. In: Cappellato, L., Ferro, N., Névéol, A., Eickhoff, C. (eds.) CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org (Sep 2020)
[14] Karaś, D., Śpiewak, M., Sobecki, P.: OPI-JSA at CLEF 2017: Author Clustering and Style Breach Detection—Notebook for PAN at CLEF 2017. In: [4], URL http://ceur-ws.org/Vol-1866/
[15] 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)
[16] Khan, J.: Style Breach Detection: An Unsupervised Detection Model—Notebook for PAN at CLEF 2017. In: [4], URL http://ceur-ws.org/Vol-1866/
[17] Khan, J.: A Model for Style Change Detection at a Glance—Notebook for PAN at CLEF 2018. In: [5]
[18] 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), URL https://www.aclweb.org/anthology/P11-1136
[19] 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), URL http://www.aclweb.org/anthology/P11-1136
[20] Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology **60**(1), 9–26 (2009)
[21] Meyer zu Eißen, S., Stein, B.: Intrinsic Plagiarism Detection. In: Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsikrika, T., Yavlinsky, A. (eds.) Advances in Information Retrieval. 28th European Conference on IR Research (ECIR 2006), Lecture Notes in Computer Science, vol. 3936, pp. 565–569, Springer, Berlin Heidelberg New York (2006), ISBN 3-540-33347-9, ISSN 0302-9743, https://doi.org/10.1007/11735106_66
[22] 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)
[23] Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York (Sep 2019), ISBN 978-3-030-22948-1, https://doi.org/10.1007/978-3-030-22948-1_5
[24] 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), URL http://ceur-ws.org/Vol-1567/paper3.pdf

[25] 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), ISBN 978-3-319-44564-9, https://doi.org/10.1007/978-3-319-44564-9_28

[26] Safin, K., Kuznetsova, R.: Style Breach Detection with Neural Sentence Embeddings—Notebook for PAN at CLEF 2017. In: [4], URL http://ceur-ws.org/Vol-1866/

[27] Safin, K., Ogaltsov, A.: Detecting a Change of Style Using Text Statistics—Notebook for PAN at CLEF 2018. In: [5]

[28] Schaetti, N.: Character-based Convolutional Neural Network for Style Change Detection—Notebook for PAN at CLEF 2018. In: [5], URL http://ceur-ws.org/Vol-2125/

[29] Stamatatos, E.: Intrinsic Plagiarism Detection Using Character n-gram Profiles. In: Notebook Papers of the 5th Evaluation Lab on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN), Amsterdam, The Netherlands (Sep 2011)

[30] Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic Plagiarism Analysis. Language Resources and Evaluation (LRE) **45**(1), 63–82 (Mar 2011), ISSN 1574-020X, https://doi.org/10.1007/s10579-010-9115-y

[31] Stein, B., Meyer zu Eißen, S.: Intrinsic Plagiarism Analysis with Meta Learning. In: Stein, B., Koppel, M., Stamatatos, E. (eds.) 1st Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 2007) at SIGIR, pp. 45–50 (Jul 2007), ISSN 1613-0073, URL http://ceur-ws.org/Vol-276

[32] Tschuggnall, M., Specht, G.: Countering Plagiarism by Exposing Irregularities in Authors' Grammar. In: Proceedings of the European Intelligence and Security Informatics Conference (EISIC), pp. 15–22, IEEE, Uppsala, Sweden (Aug 2013)

[33] 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), URL http://ceur-ws.org/Vol-1313/paper_4.pdf

[34] 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)

[35] Zangerle, E., Tschuggnall, M., Specht, G., Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2019. In: Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.) CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org (Sep 2019), URL http://ceur-ws.org/Vol-2380/

[36] Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology **57**(3), 378–393 (2006)

[37] 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—Notebook for PAN at CLEF 2018. In: [5], URL http://ceur-ws.org/Vol-2125/

[38] 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)