Overview of PAN 2021: Authorship Verification, Profiling Hate Speech Spreaders on Twitter, and Style Change Detection

Janek Bevendorff,[1] BERTa Chulvi,[2] Gretel Liz De La Peña Sarracén,[2] Mike Kestemont,[3] Enrique Manjavacas,[3] Ilia Markov,[3] Maximilian Mayerl,[4] Martin Potthast,[5] Francisco Rangel,[6] Paolo Rosso,[2] Efstathios Stamatatos,[7] Benno Stein,[1] Matti Wiegmann,[1] Magdalena Wolska,[1] and Eva Zangerle[4]

[1]Bauhaus-Universität Weimar, Germany
[2]Universitat Politècnica de València, Spain
[3]University of Antwerp, Belgium
[4]University of Innsbruck, Austria
[5]Leipzig University, Germany
[6]Symanto Research, Germany
[7]University of the Aegean, Greece

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

Abstract The paper gives a brief overview of the three shared tasks organized at the PAN 2021 lab on digital text forensics and stylometry hosted at the CLEF conference. The tasks include authorship verification across domains, author profiling for hate speech spreaders, and style change detection for multi-author documents. In part the tasks are new and in part they continue and advance past shared tasks, with the overall goal of advancing the state of the art, providing for an objective evaluation on newly developed benchmark datasets.

1 Introduction

The PAN workshop series has been organized since 2007 and has included shared tasks on specific computational challenges related to authorship analysis, computational ethics, and determining the originality of a piece of writing. Over the years, the respective organizing committees of the 51 shared tasks have assembled evaluation resources for the aforementioned research disciplines that amount to 48 datasets plus nine datasets contributed by the community.[1] Each new dataset introduced new variants of author identification, profiling, and author obfuscation tasks as well as multi-author analysis and determining the morality, quality, or originality of a text. The 2021 edition of PAN continues in the same vein, introducing new resources and previously unconsidered problems to the community. As in earlier editions, PAN is committed to reproducible research in IR and NLP and all shared tasks will ask for software submissions on our TIRA platform [22]. The following sections outline the task definitions and summarize the participants' results.

[1]https://pan.webis.de/data.html

2 Author Profiling

Author profiling is the problem of distinguishing between classes of authors by studying how language is shared by people. This helps in identifying authors' individual characteristics, such as age, gender, and language variety, among others. During the years 2013-2020 we addressed several of these aspects in the shared tasks organised at PAN.[2] In 2013 the aim was to identify gender and age in social media texts for English and Spanish [31]. In 2014 we addressed age identification from a continuous perspective (without gaps between age classes) in the context of several genres, such as blogs, Twitter, and reviews (in Trip Advisor), both in English and Spanish [28]. In 2015, apart from age and gender identification, we addressed also personality recognition on Twitter in English, Spanish, Dutch and Italian [33]. In 2016, we addressed the problem of cross-genre gender and age identification (training on Twitter data and testing on blogs and social media data) in English, Spanish, and Dutch [34]. In 2017, we addressed gender and language variety identification in Twitter in English, Spanish, Portuguese, and Arabic [32]. In 2018, we investigated gender identification in Twitter from a multimodal perspective, considering also the images linked within tweets; the dataset was composed of English, Spanish, and Arabic tweets [30]. In 2019 the focus was on profiling bots and discriminating bots from humans on the basis of textual data only [27]. We used Twitter data both in English and Spanish. Bots play a key role in spreading inflammatory content and also fake news. Advanced bots that generated human-like language, also with metaphors, were the most difficult to profile. It is interesting to note that when bots were profiled as humans, they were mostly confused with males. In 2020 we focused on profiling fake news spreaders [25]. The easiness of publishing content in social media has led to an increase in the amount of disinformation that is published and shared. The goal was to profile those authors who have shared some fake news in the past. Early identification of possible fake news spreaders on Twitter should be the first step towards preventing fake news from further dissemination.

Author profiling at PAN'21: Hate speech spreaders on Twitter Having previously profiled bots and fake news spreaders, at PAN'21 we have focused on PROFILING HATE SPEECH SPREADERS in social media, more specifically on Twitter, addressing the problem both in English and Spanish, as we did in the previous author profiling tasks. The goal has been to identify those Twitter users that can be considered haters, depending on the number of tweets with hateful content that they had spread.

Hate speech (HS) is commonly defined as any communication that disparages a person or a group on the basis of some characteristic, such as race, colour, ethnicity, gender, sexual orientation, nationality, religion, or others [20]. Given the huge amount of user-generated content on the Web and, in particular, on social media, the problem of detecting and, if possible, contrasting the HS diffusion, is becoming fundamental, for instance, in the fight against misogyny and xenophobia [1]. While most of the approaches focus on detecting whether a text is hateful or not, few works focus on the user account level detection. In [17] the authors studied the flow of posts generated

[2] To generate the datasets, we have followed a methodology that complies with the EU General Data Protection Regulation [26].

by users on Gab, analysing the profiles and network of hateful and non-hateful users, focusing on the diffusion dynamics of hateful users. The observations suggested that hateful content propagates farther, wider and faster. Unlike this work, where the analysis was carried out statically, in [18] dynamic graphs were employed to investigate the temporal effects of hate speech. In [4] the authors presented a comparative study of hate speech users on Twitter. They investigated the distinctive characteristics of hateful users and targeted users in terms of their profile, activities, and online visibility. They found that hateful users can be more popular and that participating in hate speech can result in a greater online visibility. In [35] the focus was also on users for hate speech detection on Twitter. This study used a methodology to obtain a graph given the entire profile of users, and investigated the difference between hateful users and normal ones in terms of activity patterns, word usage and network structure. The authors observed that hateful users are densely connected, thus they focused on exploiting the network of connections. In [23] the authors proposed a model that considers intra-user and inter-user representation learning for hate speech detection. In [6] the focus was on studying the use of emojis in white nationalist conversation on Twitter. A difference between the 'pro' and 'anti' nationalist was observed.

Dataset and task As an evaluation setup, we have created a collection that contains Spanish and English tweets posted by users on Twitter. To build the PAN-AP-2021 corpus[3] we have proceeded as follows. Firstly, we have looked for users considered potential haters. To do so, we have followed two approaches: (1) a keyword-based one (e.g. searching for hateful words towards women or immigrants); and (2) a user-based one, by inspecting users known as haters (e.g. users appearing in reports and/or press) and following their networks (followers and followees). Secondly, for the identified users, we have collected their timelines and manually annotated those tweets conveying hate. Thirdly, we have labelled as "keen to spread hate speech" those users with more than ten hateful tweets. Finally, we have collected two hundred tweets per Twitter user to build up the final dataset. This dataset consists of three hundred users per language, with two hundred tweets per user. Two hundred users per language have been provided for training purposes, keeping the remaining one hundred for testing purposes. The dataset is completely balanced per class (hater vs. not hater) as well as by the number of tweets per user.

The goal in the task is to classify the user as hater or not hater (binary classification). Given that we have a balanced dataset (even though this is not a realistic scenario,[4]

[3]We should highlight that we are aware of the legal and ethical issues related to collecting, analysing and profiling social media data [26] and that we are committed to legal and ethical compliance in our scientific research and its outcomes. For instance, we have anonymised the user name, masked all the user mentions and also the class has been changed in order to avoid any explicit mention.

[4]In a realistic scenario, we would need to know a priori the distribution of haters vs non-haters; depending on the study, the number of hatred messages in Twitter ranges from 1% [21] to 10%-15% [39], although when the target are communities such as the LGBT, up to 78% of respondents had experienced online anti-LGBT and hate speech in the last 5 years (https://www.report-it.org.uk/files/online-crime-2020_0.pdf). Furthermore, one of the aims of this

Table 1. Baselines performance in terms of accuracy on the PAN-AP-2021 dataset on Hate Speech Spreaders identification.

Baseline	English	Spanish	Average
LDSE	70.0	82.0	76.0
SVM + char n-grams	69.0	83.0	76.0
NN + word n-grams	65.0	83.0	74.0
USE-LSTM	56.0	79.0	67.5
XLMR-LSTM	62.0	73.0	67.5
MBERT-LSTM	59.0	75.0	67.0
TFIDF-LSTM	61.0	51.0	56.0

we balance the dataset to prevent machine/deep learning models from being skewed towards the majority class) we use accuracy as the evaluation metric for the binary classification. Then, we average both accuracies for English and Spanish to come up with the final ranking.

Evaluation and results We have had a total number of 66 participants. The best performing team has used a 100-dimension word embedding representation to feed a Convolutional Neural Network. We have also run seven baselines covering the different technologies our participants usually use:

- LDSE [29]: This method represents documents based on the probability distribution of the occurrence of their words in the different classes. The key concept of LDSE is a weight representing the probability of a term to belong to one of the two categories: hate speech spreader / non hate speech spreader. The distribution of weights for a given document should be closer to the weights of its corresponding category;
- Character n-grams with n ranging from 2 to 6 and a SVM;
- Word n-grams with n ranging from 1 to 3 and a Neural Network (NN);
- Universal Sentence Encoder (USE) feeding up a BiLSTM;
- XLM-Roberta (XLMR) transformer feeding up a BiLSTM;
- Multilingual BERT (MBERT) transformer to feed up a BiLSTM;
- TFIDF vectors representing each user's text to feed up a BiLSTM.

The baseline results are shown in Table 1. Out of the 66 participants only 7 outperformed LDSE and SVM with char n-grams baselines, further 7 participants also outperformed the NN with word n-grams baseline, and only one was worse than the TFIDF+LSTM baseline. Only 4 teams participated just in English. More details will be available in the overview paper [24].

3 Authorship Verification

Author identification is concerned with the automated identification of the individual(s) who authored an anonymous document on the basis of text-internal properties related

shared task is to foster research on profiling haters in order to address this problem automatically.

to language and writing style [37, 9, 16]. Computational author identification has been a long-running subtask at PAN with a reasonably steady number of participants over the years. While authorship has been studied via quantitative means for several decades by now, the academic and industrial interest in this task shows no signs of abating. The history of this field is characterized by a number of interesting developments and the seminal application of machine learning to the problem has been a clear landmark near the end of the previous century.[5] Today, machine learning can be considered the dominant paradigm in the field, though certain otherwise ubiquitous methods have been slow to gain a foothold. Deep learning via neural networks, for example, has become the dominant form of machine learning in many fields, yet has remained relatively uncommon in recent editions of the authorship track at PAN and in the field of computational authorship studies in general. In the past, we tentatively ascribed this absence to (1) the lack of large-scale training resources in this field and (2) the increased infrastructural challenges that come with the hardware requirements of large neural networks [14]. This problem is exacerbated by our requirement for participants to submit fully-fledged software systems to the TIRA platform [22] instead of only their finished runs. This has been a clear incentive for us to try scaling up the training resources that we can make available to participants

Scaling up resources for authorship verification at PAN'21 With the view to benchmarking authorship systems at a much larger scale, our tasks in recent years [14, 12] have focused on transformative literature, so-called "fanfiction" [8], a text variety that is nowadays abundantly available on the internet [5] with rich metadata and in many languages. Additionally, fanfiction is an excellent source of material for studies of cross-domain scenarios, since users often publish "fics" ranging over multiple topical domains ("fandoms"), such as Harry Potter, Twilight, or Marvel comics. The datasets we provided for our tasks at PAN'20 and PAN'21 were crawled from the long-established fanfiction community `fanfiction.net`. Access to the data can be requested on Zenodo.[6]

Dataset and task The 2021 edition of the authorship verification task built upon last year's edition [10] with the same task layout and training data, yet with a conceptually different test set. The basic task remained authorship verification, the most fundamental and generally more demanding setup in the field, where one is to approximate the target function $\phi : (D_k, d_u) \rightarrow \{T, F\}$, D_k being a set of documensets of known authorship by the same author and d_u being a document of unknown or disputed authorship. If $\phi(D_k, d_u) = T$, then the author of D_k is also the author of d_u and if $\phi(D_k, d_u) = F$, then the author of D_k is not the same as the author of d_u. In our case, D_k contains only a single document, since our datasets consist of document *pairs*. For the 2021 edition, we adopted a cross-domain setting in which D_k and d_u do not share the topic or genre, which was accomplished by sampling the texts from different fandoms.

[5] Machine learning emerged as a methodology in authorship attribution in the 1990s. The first paper to apply a text classification approach in this domain is [19] to the best of our knowledge.
[6] `https://zenodo.org/record/3716403`

The training resources were identical to those from last year and came in the form of a "small" and "large" dataset. The large dataset contains 148,000 same-author and 128,000 different-authors pairs across 1,600 fandoms. Each single author has written in at least two, but not more than six fandoms. The small training set is a subset of the large training set with 28,000 same-author and 25,000 different-author pairs from the same 1,600 fandoms. The test set, however (19,999 text pairs in total) is conceptually different. While the overall sampling strategy remained the same, we shifted to an "open-set" verification scenario. Whereas last year's "closed-set" test problems included only texts from fandoms and authors that were already present in the training data, this year's test set included only fresh and previously unseen authors and fandoms. This setup forces participants into a "true" verification problem, while the previous "closed-set" task (in principle) could have also been re-cast as an attribution task (although this was not known to the participants beforehand). The pure verification task is generally considered more difficult than attribution because of the stylistic idiosyncrasies of human authors which often require bespoke ad-hoc models.

Evaluation and results For each of the 19,999 problems (or text pairs) in the test set, the systems had to produce a scalar score a_i (in the $[0, 1]$ range) indicating the (scaled) probability that the pair was written by the same author ($a_i > 0.5$) or different authors ($a_i < 0.5$). Systems could choose to leave problems too difficult to answer undecided by submitting a score of precisely $a_i = 0.5$ which is rewarded by some metrics. For this year's evaluation, we used the same four evaluations metrics as last year (AUC-ROC, F_1, $\mathrm{C@1}$ and $\mathrm{F}_{0.5u}$), to allow for a diverse assessment of the submitted systems. As a result of discussions at last year's workshop, we also included the complement of the BRIER score [3] as an additional metric.[7] The submitted systems are ranked by their mean performance across all 5 metrics. Two baseline systems were made available to the participants: a compression-based approach [7] and a naive distance-based, first-order bag-of-words model [13]. We use a short-text variant of Koppel and Schler's unmasking [2, 15] as a third baseline whose source code is also freely available, but which was not given explicitly to the participants. The overall results can be found in Table 2. As in previous years, we also carried out pair-wise significance tests (based on approximate randomization, with the score as a reference metric) to be able to assess whether the answers between systems were considered significantly different according to conventional statistics. The outcome of this procedure is summarized in Table 3.

As can be seen, most of the submitted systems reach an excellent performance (many scoring > 0.9 for multiple metrics) in spite of the anticipated difficulty of the test set in comparison to last year. Last year's best performing team again tops the list, though interestingly, the runner-up is a first-time participant. Most systems produced significantly differing set of answers, with the exception of the dense cohort following the system in first place. Like last year, it is striking that systems calibrated on the large dataset invariably and significantly outperform their counterparts trained on the smaller dataset indicating that these systems are capable of harnessing the increased size of the calibration resources well. Most systems outperform the three baselines, which encour-

[7] Thanks to Fabrizio Sebastiani (Consiglio Nazionale delle Ricerche, Italy) for this suggestion.

Table 2. Final results for the cross-domain, open-set authorship verification task at PAN'21. Submitted systems are ranked by their mean performance across five evaluation metrics. Best result per column is shown in bold. Participants were allowed to make one submission for both the small and the large calibration datasets.

System	Dataset	AUC-ROC	C@1	F_1	$F_{0.5u}$	BRIER	Overall
boenninghoff21	large	**0.9869**	**0.9502**	**0.9524**	**0.9378**	**0.9452**	**0.9545**
embarcaderoruiz21	large	0.9697	0.9306	0.9342	0.9147	0.9305	0.9359
weerasinghe21	large	0.9719	0.9172	0.9159	0.9245	0.9340	0.9327
weerasinghe21	small	0.9666	0.9103	0.9071	0.9270	0.9290	0.9280
menta21	large	0.9635	0.9024	0.8990	0.9186	0.9155	0.9198
peng21	small	0.9172	0.9172	0.9167	0.9200	0.9172	0.9177
embarcaderoruiz21	small	0.9470	0.8982	0.9040	0.8785	0.9072	0.9070
menta21	small	0.9385	0.8662	0.8620	0.8787	0.8762	0.8843
rabinovits21	small	0.8129	0.8129	0.8094	0.8186	0.8129	0.8133
ikae21	small	0.9041	0.7586	0.8145	0.7233	0.8247	0.8050
unmasking21	small	0.8298	0.7707	0.7803	0.7466	0.7904	0.7836
tyo21	large	0.8275	0.7594	0.7911	0.7257	0.8123	0.7832
naive21	small	0.7956	0.7320	0.7856	0.6998	0.7867	0.7600
compressor21	small	0.7896	0.7282	0.7609	0.7027	0.8094	0.7581
futrzynski21	large	0.7982	0.6632	0.8324	0.6682	0.7957	0.7516
liaozhihao21	small	0.4962	0.4962	0.0067	0.0161	0.4962	0.3023

agingly demonstrates how the field is making progress. More details on the results will be available in the overview paper [11].

4 Multi-Author Writing Style Analysis

The goal of the style change detection task is to identify – based on an intrinsic style analysis – the text positions at which the author switches within a given multi-author document. Detecting these positions is a crucial part of the authorship identification process and multi-author document analysis, but multi-author documents have been largely understudied in general.

This task has been part of PAN since 2016 with varying task definitions, datasets, and evaluation procedures. In 2016, participants were asked to identify and group fragments of a given document that correspond to individual authors [36]. In 2017, we asked participants to detect whether a given document is multi-authored and, if this is indeed the case, to determine the positions at which authorship changes [38]. Since this task was deemed as highly complex, its complexity was reduced in 2018 to asking participants only to predict whether a given document is single- or multi-authored [14]. Following the promising results, participants were asked in the 2019 task installment to first detect whether a document was single- or multi-authored and then, if it was indeed written by multiple authors, to predict the number of authors [42]. In 2020, based on the advances made over the previous years, we decided to go back towards the original definition of the task, i.e., finding the positions in a text where authorship changes. Participants first had to determine whether a document was written by one or by multiple authors and – in the case of a multi-author document – to detect at which paragraphs the author changes [41].

Table 3. Pairwise significance tests for approximate randomization with 10,000 bootstrap iterations, using F_1 as reference metric. Symbols: '=' (not significantly different with $p > 0.5$), '*', '**', '***' (significantly different with $p < 0.05$, $p < 0.01$, $p < 0.001$). Only the top-performing systems are shown here: a full comparison will be offered in the detailed overview paper.

	embarcaderoruiz21-large	weerasinghe21-large	weerasinghe21-small	menta21-large	peng21-small
boenninghoff21-large	***	***	***	***	***
embarcaderoruiz21-large		*	=	***	**
weerasinghe21-large			***	***	=
weerasinghe21-small				**	***
menta21-large					***

Style change detection at PAN'21 For style change detection, a fundamental question is the following: If multiple authors wrote a text together, can we find evidence of this fact, e.g., do we have a means to detect variations in the writing style? Answering this question is one of the most difficult and most interesting challenges in author identification and represents the only means to detecting plagiarism in a document if no other texts are given for comparison. Likewise, it can help to uncover "gifted authorship", to verify a claimed authorship, or to develop new technologies for writing assistance. We tackle this challenge by providing three style change detection tasks in increasing difficulty: (1) Single vs. Multiple Authors: given a text, find out whether the text was written by a single author or by multiple authors, (2) Style Change Basic: given a text written by two authors that contains only a single style change, find the position of this change, i.e., cut the text into two based on stylometric information (note that this task corresponds to authorship verification where the two authors are responsible only for the first and the remaining part of a text, respectively), (3) Style Change "Real-World": given a text written by two or more authors, find all positions of writing style changes, i.e., assign all paragraphs of a text uniquely to exactly one of all the authors you deem responsible for the multi-author document.

Dataset and evaluation As in previous years, a novel dataset was created from posts from the popular StackExchange network of Q&A sites. To generate the documents for the task, we used a dump of questions and answers from the StackExchange network as our data source, of which we used a subset of communities[8]. We cleaned the data by removing questions and answers that were edited after they were originally posted and by removing images, URLs, code snippets, block quotes and bullet lists from all ques-

[8]The following StackExchange sites were used: Code Review, Computer Graphics, CS Educators, CS Theory, Data Science, DBA, DevOps, GameDev, Network Engineering, Raspberry Pi, Superuser, and Server Fault.

Table 4. Overall results for the style change detection task, ranked by average performance across all three tasks.

Participant	Task1 F_1	Task2 F_1	Task3 F_1
Zhang et al.	0.753	**0.751**	**0.501**
Strøm	**0.795**	0.707	0.424
Singh et al.	0.634	0.657	0.432
Deibel et al.	0.621	0.669	0.263
Nath	0.704	0.647	—
Baseline	0.457	0.470	0.329

tions and answers. Subsequently, we split all questions and answers into paragraphs, dropping all paragraphs with fewer than 100 characters. To reduce the potential impact of topic changes, each document was generated from a single question thread this year. Hence, for each document, we pick a question thread to draw paragraphs from. Then, we decided randomly how many authors the document should have, settling a number between one and four authors per case. Following that, we randomly chose a corresponding number of authors from the authors who contributed to the question thread we were drawing paragraphs from. We then took all the paragraphs written by those authors and shuffled them to create the final documents. If a document created in this way had fewer than two paragraphs, or was fewer than 1,000 or more than 10,000 characters long, we discarded it. Applying this procedure, we created a total of 16,000 documents. We split the resulting set of documents into a training, a test and a validation set; the training set consisted of 70% of all generated documents whereas the test and validation set each consisted of 15% of all documents. Submissions were evaluated using the F_α measure for each task and for each document, with α set to 1.

Results The style change detection task received five software submissions. Table 4 presents the individual results achieved by the participants. We list the F_1 measures for all three tasks. The approach by Strøm achieved the highest score for Task 1, whereas Zhang et al. achieved the highest score for Tasks 2 and 3. All of the submitted approaches outperformed the random baseline. Further details on the approaches taken can be found in the overview paper [40].

Acknowledgments

The work of the researchers from Universitat Politècnica de València was partially funded by the Spanish MICINN under the project MISMIS-FAKEnHATE on MISinformation and MIScommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31), and by the Generalitat Valenciana under the project DeepPattern (PROMETEO/2019/121). This article is also based upon work from the DigForAsp COST Action 17124 on Digital Forensics: evidence analysis via intelligent systems and practices, supported by European Cooperation in Science and Technology.

Bibliography

[1] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P., Sanguinetti, M.: SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In: Proc. of the 13th Int. Workshop on Semantic Evaluation (SemEval-2019), co-located with the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019) (2019)

[2] Bevendorff, J., Stein, B., Hagen, M., Potthast, M.: Generalizing Unmasking for Short Texts. In: Burstein, J., Doran, C., Solorio, T. (eds.) 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), pp. 654–659, Association for Computational Linguistics (Jun 2019), URL https://www.aclweb.org/anthology/N19-1068

[3] BRIER, G.W.: Verification of forecasts expressed in terms of probability. Monthly Weather Review **78**(1), 1 – 3 (1950), https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2, URL https://journals.ametsoc.org/view/journals/mwre/78/1/1520-0493_1950_078_0001_vofeit_2_0_co_2.xml

[4] ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., Belding, E.: Peer to Peer Hate: Hate Speech Instigators and Their Targets. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018)

[5] Fathallah, J.: Fanfiction and the Author. How FanFic Changes Popular Cultural Texts. Amsterdam University Press (2017)

[6] Hagen, L., Falling, M., Lisnichenko, O., Elmadany, A.A., Mehta, P., Abdul-Mageed, M., Costakis, J., Keller, T.E.: Emoji use in Twitter White Nationalism Communication. In: Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing, pp. 201–205 (2019)

[7] Halvani, O., Graner, L.: Cross-domain authorship attribution based on compression: Notebook for PAN at CLEF 2018. In: Cappellato, L., Ferro, N., Nie, J., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018, CEUR Workshop Proceedings, vol. 2125, CEUR-WS.org (2018), URL http://ceur-ws.org/Vol-2125/paper_90.pdf

[8] Hellekson, K., Busse, K. (eds.): The Fan Fiction Studies Reader. University of Iowa Press (2014)

[9] Juola, P.: Authorship attribution. Foundations and Trends in Information Retrieval **1**(3), 233–334 (2006)

[10] Kestemont, M., Manjavacas, E., Markov, I., Bevendorff, J., Wiegmann, M., Stamatatos, E., Potthast, M., Stein, B.: Overview of the cross-domain authorship verification task at PAN 2020. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, CEUR Workshop Proceedings, vol. 2696, CEUR-WS.org (2020), URL http://ceur-ws.org/Vol-2696/paper_264.pdf

[11] Kestemont, M., Markov, I., Stamatatos, E., Manjavacas, E., Bevendorff, J., Potthast, M., Stein, B.: Overview of the Authorship Verification Task at PAN 2021. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org (2021)

[12] Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., Stein, B.: Overview of the Cross-domain Authorship Attribution Task at PAN 2019. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019)

[13] Kestemont, M., Stover, J.A., Koppel, M., Karsdorp, F., Daelemans, W.: Authenticating the writings of julius caesar. Expert Systems with Applications **63**, 86–96 (2016), https://doi.org/10.1016/j.eswa.2016.06.029, URL https://doi.org/10.1016/j.eswa.2016.06.029
[14] 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: CLEF 2018 Labs and Workshops, Notebook Papers (2018)
[15] Koppel, M., Schler, J.: Authorship verification as a one-class classification problem. In: Brodley, C.E. (ed.) Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004, ACM International Conference Proceeding Series, vol. 69, ACM (2004), https://doi.org/10.1145/1015330.1015448
[16] 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)
[17] Mathew, B., Dutt, R., Goyal, P., Mukherjee, A.: Spread of Hate Speech in Online Social Media. In: Proceedings of the 10th ACM conference on web science, pp. 173–182 (2019)
[18] Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., Mukherjee, A.: Hate Begets Hate: A Temporal Study of Hate Speech. Proceedings of the ACM on Human-Computer Interaction **4**(CSCW2), 1–24 (2020)
[19] MATTHEWS, R.A.J., MERRIAM, T.V.N.: Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher. Literary and Linguistic Computing **8**(4), 203–209 (01 1993), ISSN 0268-1145, https://doi.org/10.1093/llc/8.4.203
[20] Nockleby, J.T.: Hate speech. In: Encyclopedia of the American Constitution (2nd ed., edited by Leonard W. Levy, Kenneth L. Karst et al., New York: Macmillan), pp. 1277–1279 (2000)
[21] Pereira-Kohatsu, J.C., Quijano-Sánchez, L., Liberatore, F., Camacho-Collados, M.: Detecting and monitoring hate speech in twitter. Sensors **19**(21), 4654 (2019)
[22] 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, Springer (2019), https://doi.org/10.1007/978-3-030-22948-1_5
[23] Qian, J., ElSherief, M., Belding, E.M., Wang, W.Y.: Leveraging Intra-user and Inter-user Representation Learning for Automated Hate Speech Detection. arXiv preprint arXiv:1804.03124 (2018)
[24] Rangel, F., De-La-Peña-Sarracén, G.L., Chulvi, B., Fersini, E., Rosso, P.: Profiling Hate Speech Spreaders on Twitter Task at PAN 2021. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org (2021)
[25] Rangel, F., Giachanou, A., Ghanem, B., Rosso, P.: Overview of the 8th Author Profiling Task at PAN 2019: Profiling Fake News Spreaders on Twitter. In: CLEF 2020 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings (2020)
[26] Rangel, F., Rosso, P.: On the implications of the general data protection regulation on the organisation of evaluation tasks. Language and Law / Linguagem e Direito **5**(2), 95–117 (2019)
[27] Rangel, F., Rosso, P.: Overview of the 7th author profiling task at pan 2019: Bots and gender profiling. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019)
[28] Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd author profiling task at PAN 2014. In: CLEF 2014 Labs and Workshops, Notebook Papers (2014)

[29] Rangel, F., Rosso, P., Franco-Salvador, M.: A low dimensionality representation for language variety identification. In: In 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing'16. Springer-Verlag, LNCS(9624), pp. 156–169 (2018)
[30] Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th Author Profiling Task at PAN 2018: Multimodal Gender Identification in Twitter. In: CLEF 2019 Labs and Workshops, Notebook Papers (2018)
[31] Rangel, F., Rosso, P., Moshe Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: CLEF 2013 Labs and Workshops, Notebook Papers (2013)
[32] Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. Working Notes Papers of the CLEF (2017)
[33] Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at PAN 2015. In: CLEF 2015 Labs and Workshops, Notebook Papers (2015)
[34] Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. In: CLEF 2016 Labs and Workshops, Notebook Papers (Sep 2016), ISSN 1613-0073
[35] Ribeiro, M., Calais, P., Santos, Y., Almeida, V., Meira Jr, W.: Characterizing and Detecting Hateful Users on Twitter. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 12 (2018)
[36] 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: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 16) (2016)
[37] Stamatatos, E.: A survey of modern authorship attribution methods. JASIST **60**(3), 538–556 (2009), https://doi.org/10.1002/asi.21001, URL https://doi.org/10.1002/asi.21001
[38] 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: CLEF 2017 Labs and Workshops, Notebook Papers (2017)
[39] Waseem, Z.: Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In: Proceedings of the First Workshop on NLP and Computational Social Science, pp. 138–142, Association for Computational Linguistics, Austin, Texas (Nov 2016), https://doi.org/10.18653/v1/W16-5618, URL https://www.aclweb.org/anthology/W16-5618
[40] Zangerle, E., Mayerl, M., , Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2021. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org (2021)
[41] Zangerle, E., Mayerl, M., Specht, G., Potthast, M., Stein, B.: Overview of the Style Change Detection Task at PAN 2020. In: CLEF 2020 Labs and Workshops, Notebook Papers (2020)
[42] Zangerle, E., Tschuggnall, M., Specht, G., Stein, B., Potthast, M.: Overview of the Style Change Detection Task at PAN 2019. In: CLEF 2019 Labs and Workshops, Notebook Papers (2019)