Overview of PAN 2019: Bots and Gender Profiling, Celebrity Profiling, Cross-Domain Authorship Attribution and Style Change Detection

Walter Daelemans[1], Mike Kestemont[1], Enrique Manjavacas[1], Martin Potthast[2(✉)], Francisco Rangel[3], Paolo Rosso[4], Günther Specht[5], Efstathios Stamatatos[6], Benno Stein[7], Michael Tschuggnall[5], Matti Wiegmann[7], and Eva Zangerle[5]

[1] University of Antwerp, Antwerp, Belgium
[2] Leipzig University, Leipzig, Germany
martin.potthast@uni-leipzig.de
[3] Autoritas Consulting, Valencia, Spain
[4] Universitat Politècnica de València, Valencia, Spain
[5] University of Innsbruck, Innsbruck, Austria
[6] University of the Aegean, Samos, Greece
[7] Bauhaus-Universität Weimar, Weimar, Germany
http://pan.webis.de

Abstract. We briefly report on the four shared tasks organized as part of the PAN 2019 evaluation lab on digital text forensics and authorship analysis. Each task is introduced, motivated, and the results obtained are presented. Altogether, the four tasks attracted 373 registrations, yielding 72 successful submissions. This, and the fact that we continue to invite the submission of software rather than its run output using the TIRA experimentation platform, demarcates a good start into the second decade of PAN evaluations labs.

1 Introduction

The PAN 2019 evaluation lab organized four shared tasks related to authorship analysis, i.e., the analysis of authors based on their writing style. Two of the tasks addressed the profiling of authors with respect to traditional demographics as well as new ones from two perspectives: (1) whether they are bots or humans, and, (2) studying the public personas of celebrities in particular. Another task tackled the most traditional task of authorship analysis, authorship attribution, but from the new angle of attributing authors across different writing domains (i.e., topics). The fourth task addressed the important, yet exceedingly difficult task of handling multi-author documents and the detection of style changes within a given text written by more than one author.

Authors are listed in alphabetical order.

The four tasks continue the series of shared tasks, which has been organized for more than a decade now starting with PAN 2009 [19], preceded only by two PAN workshops at ECAI 2008 and SIGIR 2007, which laid the foundation for what was to come. Focusing on tasks from the areas digital text forensics, text reuse, and judging the trustworthiness and ethicality of texts, we have assembled new benchmarks for more than a dozen different tasks now, many of which continue to be used for evaluations throughout the research community. In this paper, each of the following sections gives a brief, condensed overview of the four aforementioned tasks, including their motivation and the results obtained.

2 Bots and Gender Profiling

Author profiling aims at classifying authors depending on how language is shared by groups of people. This may allow to identify demographics such as age and gender, and it can be of high interest from a marketing, security and forensics perspective. The research community has shown an increasing interest in the author profiling shared task throughout the past years, as evidenced by the growing number of participants.[1] Having addressed several aspects of author profiling in social media from 2013 to 2018, the author profiling shared task of 2019 aims at investigating whether the author of a Twitter feed is a bot or a human. Furthermore, in case of a human it was asked to profile the gender of the author. As in previous years, we have proposed the task from a multilingual perspective, covering English and Spanish languages. One of our main objectives was to demonstrate the feasibility of automatically identifying bots as well as demonstrating the difficulty of identifying more elaborate bots than basic information spreaders.

2.1 Evaluation Framework

To build the PAN-AP-2019 corpus,[2] we have combined Twitter accounts identified as bots in existing datasets with newly discovered ones on the basis of specific search queries. In both cases, a minimum of three annotators agreed with the annotation, or else the Twitter user was discarded. To annotate gender, we followed the same methodology as in previous editions of the shared task. In Table 1, some corpus statistics are shown. The corpus is balanced per type (bot/human), and in case of human, it is also balanced per gender. Each author is composed of exactly 100 tweets.

[1] In the past seven editions of the author profiling shared task at PAN, we have had 21 (2013 [26]), 10 (2014 [23]), 22 (2015 [20]), 22 (2016 [28]), 22 (2017 [27]), 23 (2018 [25]), and 55 (2019 [22]) participating teams, respectively.

[2] We should highlight that we are aware of the legal and ethical issues related to collecting, analyzing, and profiling social media data [21], and that we are committed to legal and ethical compliance in our scientific research and its outcomes.

Table 1. Number of authors per language. The corpus is balanced regarding bots vs. humans, and regarding gender in case of humans, and it contains 100 tweets per author.

Dataset	English (EN)	Spanish (ES)
Training	4,120	3,000
Test	2,640	1,800

The participants were asked to send two predictions per author: (1) whether the author is a bot or a human, and in case of a human (2) whether the author is male or female. The participants were allowed to approach the task also in one instead of all of the languages, and to address only one subproblems (bots or gender). Classification accuracy has been employed for evaluation. For each language, we obtain the accuracy for both problems in both languages separately and average them to obtain the final ranking.

2.2 Results

This year, 55 teams participated in the shared task. In Table 2, the overall performance per language and participant's ranking are shown. The best results have been obtained for both identification (95.95% in English vs. 93.33% in Spanish) and gender profiling (84.17% in English vs. 81.72% in Spanish). As can be seen, results for bot identification are higher than 90% in some cases, revealing the relative ease of this task. A more in-depth analysis is presented in the overview paper [22] where we show that certain types of bots are not as easy to detect as others, and the risks this entails.

In Table 2, the best results per language and problem are highlighted in bold font. The overall best result (88.05%) has been obtained by the author in [16]. They have approached the task with a Support Vector Machine with character and word n-grams as features. It is worth mentioning the high performance obtained by the word and character n-grams baselines, even greater than that of word embeddings [12,13] and Low Dimensionality Statistical Embedding (LDSE) [24].

3 Celebrity Profiling

Celebrities are a highly prolific population of Twitter users. They influence public opinion, are role models to their fans and follower, and sometimes they are the voices of the disenfranchised. For these reasons, the "rich and famous" have been studied in the social sciences and economics as a matter of course, especially with regard to their presence on social media. Our recent seminal work on celebrity profiling [34], and this task at PAN 2019 introduce this particular group of people to computational linguistics. This task focuses on determining four demographics of celebrities based on their Twitter timelines:

Table 2. Accuracy per subtask and language, and global ranking as average.

Ranking	Team	Bots vs. Human		Gender		Average
		EN	ES	EN	ES	
1	Pizarro	0.9360	0.9333	0.8356	0.8172	0.8805
2	Srinivasarao & Manu	0.9371	0.9061	0.8398	0.7967	0.8699
3	Bacciu et al.	0.9432	0.9078	0.8417	0.7761	0.8672
4	Jimenez-Villar et al.	0.9114	0.9211	0.8212	0.8100	0.8659
5	Fernquist	0.9496	0.9061	0.8273	0.7667	0.8624
6	Mahmood	0.9121	0.9167	0.8163	0.7950	0.8600
7	Ipsas & Popescu	0.9345	0.8950	0.8265	0.7822	0.8596
8	Vogel & Jiang	0.9201	0.9056	0.8167	0.7756	0.8545
9	Johansson & Isbister	0.9595	0.8817	0.8379	0.7278	0.8517
10	Goubin et al.	0.9034	0.8678	0.8333	0.7917	0.8491
11	Polignano & de Pinto	0.9182	0.9156	0.7973	0.7417	0.8432
12	Valencia et al.	0.9061	0.8606	0.8432	0.7539	0.8410
13	Kosmajac & Keselj	0.9216	0.8956	0.7928	0.7494	0.8399
14	Fagni & Tesconi	0.9148	0.9144	0.7670	0.7589	0.8388
	char nGrams	0.9360	0.8972	0.7920	0.7289	0.8385
15	Glocker	0.9091	0.8767	0.8114	0.7467	0.8360
	word nGrams	0.9356	0.8833	0.7989	0.7244	0.8356
16	Martinc et al.	0.8939	0.8744	0.7989	0.7572	0.8311
17	Sanchis & Velez	0.9129	0.8756	0.8061	0.7233	0.8295
18	Halvani & Marquardt	0.9159	0.8239	0.8273	0.7378	0.8262
19	Ashraf et al.	0.9227	0.8839	0.7583	0.7261	0.8228
20	Gishamer	0.9352	0.7922	0.8402	0.7122	0.8200
21	Petrik & Chuda	0.9008	0.8689	0.7758	0.7250	0.8176
22	Oliveira et al.	0.9057	0.8767	0.7686	0.7150	0.8165
	W2V	0.9030	0.8444	0.7879	0.7156	0.8127
23	De La Peña & Prieto	0.9045	0.8578	0.7898	0.6967	0.8122
24	López Santillán et al.	0.8867	0.8544	0.7773	0.7100	0.8071
	LDSE	0.9054	0.8372	0.7800	0.6900	0.8032
25	Bolonyai et al.	0.9136	0.8389	0.7572	0.6956	0.8013
26	Moryossef	0.8909	0.8378	0.7871	0.6894	0.8013
27	Zhechev	0.8652	0.8706	0.7360	0.7178	0.7974
28	Giachanou & Ghanem	0.9057	0.8556	0.7731	0.6478	0.7956
29	Espinosa et al.	0.8413	0.7683	0.8413	0.7178	0.7922
30	Rahgouy et al.	0.8621	0.8378	0.7636	0.7022	0.7914
31	Onose et al.	0.8943	0.8483	0.7485	0.6711	0.7906
32	Przybyla	0.9155	0.8844	0.6898	0.6533	0.7858
33	Puertas et al.	0.8807	0.8061	0.7610	0.6944	0.7856
34	Van Halteren	0.8962	0.8283	0.7420	0.6728	0.7848
35	Gamallo & Almatarneh	0.8148	0.8767	0.7220	0.7056	0.7798
36	Bryan & Philipp	0.8689	0.7883	0.6455	0.6056	0.7271
37	Dias & Paraboni	0.8409	0.8211	0.5807	0.6467	0.7224
38	Oliva & Masanet	0.9114	0.9111	0.4462	0.4589	0.6819
39	Hacohen-Kerner et al.	0.4163	0.4744	0.7489	0.7378	0.5944
40	Kloppenburg	0.5830	0.5389	0.4678	0.4483	0.5095
	MAJORITY	0.5000	0.5000	0.5000	0.5000	0.5000
	RANDOM	0.4905	0.4861	0.3716	0.3700	0.4296
41	Bounaama & Amine	0.5008	0.5050	0.2511	0.2567	0.3784
42	Joo & Hwang	0.9333	-	0.8360	-	0.4423
43	Staykovski	0.9186	-	0.8174	-	0.4340
44	Cimino & Dell'Orletta	0.9083	-	0.7898	-	0.4245
45	Ikae et al.	0.9125	-	0.7371	-	0.4124
46	Jeanneau	0.8924	-	0.7451	-	0.4094
47	Zhang	0.8977	-	0.7197	-	0.4044
48	Fahim et al.	0.8629	-	0.6837	-	0.3867
49	Saborit	-	0.8100	-	0.6567	0.3667
50	Saeed & Shirazi	0.7951	-	0.5655	-	0.3402
51	Radarapu	0.7242	-	0.4951	-	0.3048
52	Bennani-Smires	0.9159	-	-	-	0.2290
53	Gupta	0.5007	-	0.4044	-	0.2263
54	Qurdina	0.9034	-	-	-	0.2259
55	Aroyehun	0.5000	-	-	-	0.1250

– Their `gender`, as male, female, or, for the first time, non-binary.
– Their precise `birth year` within a novel, variable-bucket evaluation scheme.
– Their degree of `fame`, as rising, star, or superstar.
– Their `occupation`, as in "claim to fame", categorized as sports, performer, creator, politics, manager, science, professional, or religious.

This is the first installment of celebrity profiling at PAN, with 92 registrations, 12 active participants and seven submitted solutions.

Table 3. Results on both test datasets for the celebrity profiling task.

Team	Test dataset 1					Test dataset 2				
	cRank	gender	age	fame	occup	cRank	gender	age	fame	occup
radivchev	**0.593**	**0.726**	**0.618**	0.551	**0.515**	**0.559**	**0.609**	**0.657**	**0.548**	0.461
morenosandoval	0.541	0.644	0.518	**0.563**	0.469	0.497	0.561	0.516	0.518	0.418
martinc	0.462	0.580	0.361	0.517	0.449	0.465	0.594	0.347	0.507	**0.486**
fernquist	0.424	0.447	0.339	0.493	0.449	0.413	0.465	0.467	0.482	0.300
petrik	0.377	0.595	0.255	0.480	0.340	0.441	0.555	0.360	0.526	0.385
asif	–	–	–	–	–	0.402	0.588	0.254	0.504	0.427
bryan	–	–	–	–	–	0.231	0.335	0.207	0.289	0.165
baseline-rand	0.223	0.344	0.123	0.341	0.125	–	–	–	–	–
baseline-uniform	0.138	0.266	0.117	0.099	0.152	–	–	–	–	–
baseline-mv	0.136	0.278	0.071	0.285	0.121	–	–	–	–	–

3.1 Datasets

The complete dataset for this task contained the Twitter timelines of 48,335 celebrity accounts, annotated with the four social variables gender, birth year, fame, and occupation. We constructed the dataset by matching all verified Twitter accounts to their respective Wikidata entries [34], omitting all memorial and business accounts. This method yielded 71,706 entries for verified, notable, and living humans with an estimated error rate of 0.6%. From these, we sampled all accounts which had Wikidata entries indicating gender, year of birth, and occupation and which had English as their main language marked in their Twitter profile, leaving 48,335 authors, each with an average 2,181 tweets. The training dataset comprised 33,836 authors and the test dataset 14,499 authors; 956 authors were sampled from the latter as small-scale test dataset. To label them, gender and year of birth were extracted from their respective Wikidata items; the 1,379 listed different occupations were grouped into eight categories. Fame was determined based on their number of followers: rising (below 1000), star, and superstar (>100,000). These boundaries reflect the standard deviation of a Gaussian distribution overlaid on the logarithm of the follower distribution across all datasets.

3.2 Evaluation Framework

The performance measure for this task is *cRank*, the harmonic mean of the measures employed for each individual demographic:

$$\text{cRank} = \frac{4}{\frac{1}{\text{F}_{1,\text{fame}}} + \frac{1}{\text{F}_{1,\text{occupation}}} + \frac{1}{\text{F}_{1,\text{gender}}} + \frac{1}{\text{F}_{1,\text{birth year}}}}.$$

For gender, fame, and occupation, performance is estimated as multi-class F_1. Since the dataset features a realistic distribution of the social variables, we favored micro- over macro-averaged F_1. For age, we chose a lenient approach: Instead of grouping the year of birth into fixed age buckets, participants were asked to determine a precise year, whereas we applied a variable-bucket strategy during evaluation. Here, the predicted year of birth of an author is correct if it is within an ε-environment of the truth. The threshold ε is between 2 and 9 years, increasing linearly with the true age of the author.

3.3 Results

Altogether, seven participants successfully submitted software to the celebrity profiling task. Table 3 lists the performance of their methods for cRank and the individual measures. A notable observation is that performance is more varied on the more difficult test dataset 1, where leading approaches perform better on the more difficult dataset while others perform weaker. Additionally, while the ordering of participants by cRank is the same for both datasets, it differs for individual demographics. We provide more insights into participants' performance and the analysis of the results in the extended task overview [35].

4 Cross-Domain Authorship Attribution

Authorship attribution [5,9,31] continues to be an important problem in information retrieval and computational linguistics, but also in applied areas such as law and journalism, where knowing the author of a document (such as a ransom note) may enable, e.g., law enforcement to save lives. The most common framework for testing candidate algorithms is the closed-set attribution task: given a sample of reference documents from a restricted and finite set of candidate authors, the task is to determine the most likely author of a previously unseen document of unknown authorship. This task may be quite challenging in cross-domain conditions, when documents of known and unknown authorship come from different domains (e.g., thematic area, genre). In addition, it is often more realistic to assume that the true author of a disputed document is not necessarily included in the list of candidates [10].

This year, we again focus on the attribution task in the context of transformative literature, more colloquially know as 'fanfiction'. Fanfiction refers to

a rapidly expanding body of fictional narratives typically produced by non-professional authors who self-identity as 'fans' of a particular oeuvre or individual work [4]. When sharing their texts, fanfiction writers explicitly acknowledge taking inspiration from one (or more) literary domains that are known as 'fandoms'. From the perspective of writing style, fanfiction offers valuable benchmark data: the writings are unmediated and unedited before publication, meaning that they should accurately reflect an individual author's writing style. In the previous edition, this task dealt with authorship attribution in fanfiction, and specifically attribution across different domains or fandoms. This year, we have further increased the difficulty of the task, by focusing on *open-set* attribution conditions, meaning that the true author of a test text is not necessarily included in the list of candidate authors. More formally, an open cross-domain authorship attribution problem can be expressed as a tuple (A, K, U), with A as the set of candidate authors, K as the set of reference (known authorship) texts, and U as the set of unknown authorship texts. For each candidate author $a \in A$, we are given $K_a \subset K$, a set of texts unquestionably written by a. Each text in U should be assigned to exactly one $a \in A$ or the system should refrain from an attribution, if the target author of a text in U is not in A. From a text categorization point of view, K is the training corpus and U is the test corpus. Let D_K be the set of fandoms of texts in K. Then, all texts in U belong to a single (target) fandom $d_U \notin D_K$.

4.1 Datasets

This year's shared task worked with datasets in four major Indo-European languages: English ("en"), French ("fr"), Italian ("it"), and Spanish ("sp"). For each language, 10 "problems" were constructed on the basis of a larger dataset obtained from archiveofourown.org in 2017. Per language, five problems were released as a development set to the participants, in order to calibrate their systems. The final evaluation of the submitted systems was carried out on the five remaining problems (which were not publicly released before the final results were communicated). Each problem had to be solved fully independently from the other problems by a system. Importantly, the development material could not be treated as mere training material for supervised learning approaches, because the sets of candidate authors of the development and the evaluation corpora are not overlapping. Therefore, approaches should not be designed to particularly handle the candidate authors of the development corpus but should focus on their scalability to other author sets.

One "problem" corresponds to a single open-set attribution task, where we distinguish between the "source" and "target" material. The "source" material in each problem contains exactly 7 training texts for exactly 9 candidate authors. In the "target" material, these 9 authors are represented by at least one test text (but potentially more). Additionally, the target material also contains so-called "adversaries", which were not written by one of the candidate authors (indicated by the author label "<UNK>"). The proportion of the number of target texts

written by the candidate authors in problems, as opposed to <UNK> documents, was varied across the problems in the development dataset, in order to discourage systems from opportunistic guessing.

Let U_K be the subset of U that includes all test documents actually written by the candidate authors while U_U is the subset of U containing the rest of test documents not written by any candidate author. Then, the *adversary ratio* $a = |U_U|/|U_K|$ determines the likelihood of a test document to belong to one of the candidates. If $a = 0$ (or close to 0), then it is essentially a closed-set attribution scenario, since all test documents belong to the candidate authors (or very few are actually written by adversaries). If $a = 1$, then it is equally probable for a test document to be written by a candidate author or by another author. If $a > 1$, then it is more likely for a test document to be written by an adversary not included in the list of candidates.

In this edition of the authorship attribution task, we examine cases where a ranges from 0.2 to 1.0. In more detail, as can be seen in Table 4, the development dataset comprises 5 problems per language that correspond to $a = [0.2, 0.4, 0.6, 0.8, 1.0]$. This dataset was released in order for the participants to develop and calibrate their submissions. The final evaluation dataset also includes 5 problems per language but with fixed $a = 1$. Thus, the participants are guided to develop generic approaches (varying likelihood a test document is written by a candidate or an adversary). In addition, it is possible to estimate the effectiveness of submitted methods when $a < 1$ by ignoring their answers for specific subsets of U_U in the evaluation dataset.

Table 4. Details about the fanfiction datasets built for the cross-domain authorship attribution task. $|A|$ refer to the size of candidates list, $|K_a|$ is the amount of training documents per author, $|U|$ is the amount of test documents, a is the adversary ratio, and $|d|$ denotes the average length (in words) of documents.

	Language	Problems	$\|A\|$	$\|K_a\|$	$\|U\|$	a	$\|d\|$
Development	English	5	9	7	137-561	0.2-1.0	804
	French	5	9	7	38-430	0.2-1.0	790
	Italian	5	9	7	46-196	0.2-1.0	814
	Spanish	5	9	7	112-450	0.2-1.0	846
Evaluation	English	5	9	7	98-180	1.0	817
	French	5	9	7	48-290	1.0	790
	Italian	5	9	7	34-302	1.0	821
	Spanish	5	9	7	172-588	1.0	838

4.2 Evaluation Framework

The submissions were separately evaluated in each attribution problem based on their open-set macro-averaged F_1 score (calculated over the training classes,

i.e., when <UNK> is excluded) [11]. Participants were ranked according to their average open-set macro-F_1 across all attribution problems of the evaluation corpus. A reference implementation was made available to the participants. As customary, we provide the implementation of three baseline methods that offered an estimation of the overall difficulty of the problem given the state of the art in the field. These implementations were in Python (2.7+) and relied heavily on Scikit-learn and its base packages [14,15] as well as NLTK [1]:

1. BASELINE-SVM: a language-independent authorship attribution approach that frames attribution as a conventional text classification problem [30]. It is based on a character 3-gram representation and a linear SVM classifier with a reject option. It estimates the probabilities of output classes and assigns an unknown document to the <UNK> class when the difference of the top two candidates is less than a threshold.
2. BASELINE-COMPRESSOR: a language-independent approach that uses text compression to estimate the distance of an unknown document to each of the candidate authors (originally proposed by [32] and reproduced by [17]). It assigns an unknown document to the <UNK> class when the difference between the two most likely candidates is lower than a threshold.
3. BASELINE-IMPOSTERS: an implementation of the language-independent "imposters" approach for authorship verification [7,10], based on character tetragram features. During a bootstrapped procedure, the technique iteratively compares an unknown text to each candidate author's stylistic profile, as well as to a set of imposter documents, on the basis of a random feature set. If the highest ranking candidate author does not pass a fixed similarity threshold after this procedure, the document is assigned to the <UNK> class and left unattributed. We included a set of 5,000 problem-external documents per language written by "imposter" authors (the authorship of these texts is also encoded as "<UNK>".)

4.3 Evaluation Results

In total, 12 methods were submitted to the task. The task overview paper contains a more comprehensive overview and discussion of the submitted methods [6]. Table 5 shows an overview of the evaluation results of participants and their ranking according to their macro-F_1 (averaged across all attribution problems of the dataset). As can be seen, all but one submission surpass the three baseline methods. In general, the submitted methods and the baselines achieve better macro-recall than macro-precision. The two top-performing submissions obtain very similar macro-F_1 score. However, the winning approach of Muttenthaler et al. has better macro-precision while Neri et al. achieve better macro-recall. The winning approach also proved to be runtime-efficient.

Table 5. The final evaluation results of the cross-domain authorship attribution task. Participants and baselines are ranked according to macro-F_1.

Submission	Macro-Precision	Macro-Recall	Macro-F_1	Runtime
Muttenthaler et al.	**0.716**	0.742	**0.690**	00:33:17
Bacciu et al.	0.688	**0.768**	0.680	01:06:08
Custódio et al.	0.664	0.717	0.65	01:21:13
Bartelds & de Vries	0.657	0.719	0.644	11:19:32
Rodríguez et al.	0.651	0.713	0.642	01:59:17
Isbister	0.629	0.706	0.622	01:05:32
Johansson	0.593	0.734	0.616	01:05:30
Basile	0.616	0.692	0.613	00:17:08
Van Halteren	0.590	0.734	0.598	37:05:47
Rahgouy et al.	0.601	0.633	0.580	02:52:03
Gagala	0.689	0.593	0.576	08:22:33
baseline-svm	0.552	0.635	0.545	
baseline-compressor	0.561	0.629	0.533	
baseline-impostors	0.428	0.580	0.395	
Kipnis	0.270	0.409	0.259	20:20:21

5 Style Change Detection

Style change detection tasks at previous PAN editions [8,29,33] aimed to analyze multi-authored documents. In 2016, the task was to identify and group text fragments of individual authors [29], whereas, in 2017, the goal was to determine whether a given document is multi-authored, and if this is the case, to find the borders where authors switch [33]. These tasks showed that accurately identifying individual authors and their contributions within a single document is a complex task. Hence, last year, we substantially relaxed the problem by transforming it into a binary classification task that predicts whether a given document is single- or multi-authored [8]. Considering the promising results achieved by the submitted approaches, we continue last year's task and additionally ask participants to predict the number of involved authors. Hence, this year's style change detection task was defined as follows: given a document, (1) is the document written by one or more authors (i.e., are there style changes or not?), and, (2) if the document is multi-authored, how many authors have collaborated?

5.1 Evaluation Dataset

The datasets provided for training, validation, and testing of the approaches were curated based on data of the StackExchange Q&A platform.[3] We extract user questions and answers from 15 heterogeneous sites, which cover topics ranging from cooking to philosophy. The datasets are assembled by varying the following parameters:

- number of style changes (including 0 for single-authored documents)
- number of collaborating authors (1–5)
- document length (300–1500 tokens)
- allowing changes only at the end or within paragraphs
- uniform or random distribution of changes with respect to segment lengths

The split between training, validation, and test was performed by employing approximate 50/25/25% stratified random sampling. An overview of the datasets is depicted in Table 6, where we list the number of documents for the different number of authors (absolute numbers and relative share in the respective dataset) and the average number of tokens per document for single- and multi-authored documents.

Table 6. Overview style change detection datasets, where SA and MA refer to single-authored and multi-authored documents, respectively, and text length is measured by the average number of tokens per document.

Dataset	Docs	Authors					Text 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

5.2 Performance Measures

The style change detection task comprises answering two questions individually: distinguishing single- from multi-author documents and predicting the number of authors in case of a multi-authored document. Hence, the performance measure employed to assess the quality of the participant's approaches naturally incorporates the performance of the two sub-tasks. Particularly, we employ *accuracy*

[3] https://stackexchange.com/.

for the binary classification task of distinguishing between single-authored from multi-authored documents. For measuring the prediction performance regarding the actual number of authors, we reason that in this classification task, we are not only interested in measuring the number of correctly classified documents, but also aim to incorporate the extent to which the prediction differed from the actual class. As our classes employed are integers (the number of authors), we incorporate the distance between the predicted and the actual class in the performance measure. Hence, we employ the *Ordinal Classification Index (OCI)* [3] as an error measure for ordinal data in classification tasks. This index is based on the confusion matrices resulting from the classification task employed and yields a value between 0 and 1, with 0 being the best value (perfect prediction). Besides measuring accuracy and the ordinal classification index individually, we also combine those two measures into a single rank measure:

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

5.3 Results

The style change detection task received two software submissions, which were evaluated on the TIRA experimentation platform. We depict the participant's results in Table 7, where we list accuracy, the ordinal classification index and the proposed overall rank measure. As can be seen, Nath achieves higher scores for both sub-tasks and hence, also in the combined rank measure. More details on the approaches taken can be found in the task overview [36].

Table 7. Overall results for the style change detection task

Participant	Accuracy	OCI	Rank
Zuo	0.6041	0.8086	0.3978
Nath	0.8479	0.8652	0.4913

6 Summary and Outlook

This year's PAN lab has been quite a success in terms of establishing new tasks for the coming years, community interest and scale, and quality of the newly developed benchmarking resources. While not every task attracted a large number of participants, we hope to continue to develop each one by introducing the new concept of an ongoing online task. Based on the TIRA evaluation platform [18], it becomes manageable to basically keep a task running, accepting new participants with little to no overhead on our part, while giving those who did not find the time to participate ahead of the submission deadline for PAN 2019

to do so afterwards, thereby making an early contribution for PAN 2020. If such a routine could be established, the development of new shared tasks would become more disentangled from a rigid timeline of deadlines. Rather, the only deadline remaining would be a cut-off date for the next PAN workshop that participants who want their submissions published have to meet, whereas they can plan and pursue their submission in their own time throughout the year. Still, many demand deadlines, so that a regular engagement of participants by organizers will continue to be an important part of organizing a shared task. We hope that, using the concept of ongoing online tasks, even tasks that did not attracts lots attention in terms of participants, but that are still of general interest and importance, will get a chance of being promoted. That said, we still plan to nurture our large tasks and to grow them even further, if possible.

Acknowledgments. The work of Paolo Rosso was partially funded by the Spanish MICINN under the research project MISMIS-FAKEnHATE on Misinformation and Miscommunication in social media: FAKE news and HATE speech (PGC2018-096212-B-C31). Our special thanks goes to all PAN participants for providing high-quality submission, to Symanto (https://www.symanto.net) for sponsoring the PAN Lab 2019 and to The Logic Value (https://thelogicvalue.com) for sponsoring the author profiling shared task award.

References

1. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media, Sebastopol (2009)
2. Cappellato, L., Ferro, N., Losada, D., Müller, H. (eds.): CLEF 2019 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, September 2019
3. Cardoso, J., Sousa, R.: Measuring the performance of ordinal classification. Int. J. Pattern Recognit Artif Intell. **25**(08), 1173–1195 (2011)
4. Hellekson, K., Busse, K. (eds.): The Fan Fiction Studies Reader. University of Iowa Press, Iowa City (2014)
5. Juola, P.: Authorship attribution. Found. Trends Inf. Retrieval **1**(3), 233–334 (2006)
6. Kestemont, M., Stamatatos, E., Manjavacas, E., Daelemans, W., Potthast, M., Stein, B.: Overview of the cross-domain authorship attribution task at PAN 2019. In: Cappellato et al. [2]
7. Kestemont, M., Stover, J.A., Koppel, M., Karsdorp, F., Daelemans, W.: Authenticating the writings of Julius Caesar. Expert Syst. Appl. **63**, 86–96 (2016). https://doi.org/10.1016/j.eswa.2016.06.029
8. Kestemont, M., et al.: Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In: Cappellato, L. et al. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs, Avignon, France, 10–14 September 2018, pp. 1–25 (2018)
9. Koppel, M., Schler, J., Argamon, S.: Computational methods in authorship attribution. J. Am. Soc. Inf. Sci. Technol. **60**(1), 9–26 (2009)
10. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. J. Assoc. Inf. Sci. Technol. **65**(1), 178–187 (2014)

11. Júnior, P.R.M., et al.: Nearest neighbors distance ratio open-set classifier. Mach. Learn. **106**(3), 359–386 (2017)
12. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of Workshop at International Conference on Learning Representations (ICLR 2013) (2013)
13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
14. Oliphant, T.: NumPy: A Guide to NumPy. Trelgol Publishing (2006). http://www.numpy.org/
15. Pedregos, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
16. Pizarro, J.: Using n-grams to detect bots on Twitter: notebook for PAN at CLEF 2019. In: Cappellato et al. [2]
17. Potthast, M., et al.: Who wrote the web? Revisiting influential author identification research applicable to information retrieval. In: Ferro, N., et al. (eds.) ECIR 2016. LNCS, vol. 9626, pp. 393–407. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-30671-1_29
18. 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 - Lessons Learned from 20 Years of CLEF. Springer, Heidelberg (2019)
19. Potthast, M., Rosso, P., Stamatatos, E., Stein, B.: A decade of shared tasks in digital text forensics at PAN. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11438, pp. 291–300. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15719-7_39
20. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at PAN 2015. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop - Working Notes Papers, 8–11 September, Toulouse, France. CEUR-WS.org (2015)
21. Rangel, F., Rosso, P.: On the implications of the general data protection regulation on the organisation of evaluation tasks. Lang. Law= Linguagem e Direito **5**(2), 95–117 (2018)
22. Rangel, F., Rosso, P.: Overview of the 7th author profiling task at PAN 2019: bots and gender profiling. In: Cappellato et al. [2]
23. Rangel, F., et al.: Overview of the 2nd author profiling task at PAN 2014. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop - Working Notes Papers, 15–18 September, Sheffield, UK. CEUR-WS.org (2014)
24. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: Gelbukh, A. (ed.) CICLing 2016. LNCS, vol. 9624, pp. 156–169. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75487-1_13
25. Rangel, F., Rosso, P., G'omez, M.M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at PAN 2018: multimodal gender identification in Twitter. In: CLEF 2018 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org (2017)
26. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the author profiling task at PAN 2013. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop - Working Notes Papers, 23–26 September, Valencia, Spain, September 2013

27. 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. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org, September 2017
28. 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: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org., September 2016
29. 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., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 332–350. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_28
30. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34**(1), 1–47 (2002)
31. Stamatatos, E.: A survey of modern authorship attribution methods. J. Am. Soc. Inf. Sci. Technol. **60**, 538–556 (2009)
32. Teahan, W.J., Harper, D.J.: Using compression-based language models for text categorization. In: Croft, W.B., Lafferty, J. (eds.) Language Modeling for Information Retrieval. INRE, vol. 13, pp. 141–165. Springer, Dordrecht (2003). https://doi.org/10.1007/978-94-017-0171-6_7
33. Tschuggnall, M., et al.: Overview of the author identification task at PAN-2017: style breach detection and author clustering. In: Cappellato, L. et al. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs, pp. 1–22 (2017)
34. Wiegmann, M., Stein, B., Potthast, M.: Celebrity profiling. In: 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019). Association for Computational Linguistics, July 2019
35. Wiegmann, M., Stein, B., Potthast, M.: Overview of the celebrity profiling task at PAN 2019. In: Cappellato et al. [2]
36. Zangerle, E., Tschuggnall, M., Specht, G., Potthast, M., Stein, B.: Overview of the style change detection task at PAN 2019. In: Cappellato et al. [2]