Overview of PAN 2018

Author Identification, Author Profiling, and Author Obfuscation

Efstathios Stamatatos[1], Francisco Rangel[2,3], Michael Tschuggnall[4], Benno Stein[5], Mike Kestemont[6], Paolo Rosso[3], and Martin Potthast[5]

[1]Dept. of Information & Communication Systems Eng., University of the Aegean, Greece
[2]Autoritas Consulting, S.A., Spain
[3]PRHLT Research Center, Universitat Politècnica de València, Spain
[4]Department of Computer Science, University of Innsbruck, Austria
[5]Web Technology & Information Systems, Bauhaus-Universität Weimar, Germany
[6]University of Antwerp, Belgium

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

Abstract PAN 2018 explores several authorship analysis tasks enabling a systematic comparison of competitive approaches and advancing research in digital text forensics. More specifically, this edition of PAN introduces a shared task in cross-domain authorship attribution, where texts of known and unknown authorship belong to distinct domains, and another task in style change detection that distinguishes between single-author and multi-author texts. In addition, a shared task in multimodal author profiling examines, for the first time, a combination of information from both texts and images posted by social media users to estimate their gender. Finally, the author obfuscation task studies how a text by a certain author can be paraphrased so that existing author identification tools are confused and cannot recognize the similarity with other texts of the same author. New corpora have been built to support these shared tasks. A relatively large number of software submissions (41 in total) was received and evaluated. Best paradigms are highlighted while baselines indicate the pros and cons of submitted approaches.

1 Introduction

Nowadays, a huge amount of digital texts is produced daily in Internet media. In many cases, the originality and credibility of this information is questionable. In addition, information about the authors of texts may be missing or hidden behind aliases. It is, therefore, essential to attempt to estimate credibility of texts and unmask author information in order to avoid social media misuse, enhance cyber-security, and enable digital text forensics. PAN is an evaluation lab dedicated to study originality (is this information new or re-used?), trust (can we trust this information?), and authorship (who wrote this?) of digital texts. Several shared tasks have been organized during the last 10 years covering many aspects of this field.

PAN 2018 follows the trend of recent years and focuses on authorship analysis exclusively. This research area attempts to reveal information about the authors of texts based mainly on their stylistic preferences. Every author has her unique characteristics

(stylistic fingerprint) but she also shares some properties with other people of similar background (age, gender, education, etc.) It is quite challenging to define or measure both personal style (for each individual author) and collective style (males, females, young people, old people, etc.). In addition, it remains unclear what one should modify in her texts in order to attempt to hide her identity or to mimic the style of another author. This edition of PAN deals with these challenging issues.

Author identification puts emphasis on the personal style of individual authors. The most common task is authorship attribution where there is a set of candidate authors (suspects), with samples of their texts, and one of them is selected as the most likely author of a text of disputed authorship [31]. This can be a closed-set (one of the suspects is surely the true author) or an open-set (the true author may not be among the suspects) attribution case. This edition of PAN focuses on closed-set *cross-domain authorship attribution*, that is, when the texts unquestionably written by the suspects and the texts of disputed authorship belong to different domains. This is a realistic scenario suitable for several applications. For example, imagine the case of a crime novel published anonymously when all candidate authors have only published fantasy novels [13] or a disputed tweet when the available texts written by the suspects are newspaper articles. To be able to control the domain of texts, we turned to so-called *fanfiction* [11]. This term refers to the large body of contemporary fiction that is nowadays created by non-professional authors ('fans'), who write in the tradition of a well-known source work, such as the *Harry Potter* series by J.K. Rowling, that is sometimes called the 'canon'. These writings or 'fics' within such a 'fandom' heavily borrow characters, motives, settings, etc. from the source fandom. Fanfiction provides excellent material to study cross-domain authorship attribution since most fans are active in multiple fandoms.

Another important dimension in author identification is to intrinsically analyse a document, possibly written by multiple authors and identify the contribution of each co-author. The previous edition of PAN aimed to find the exact border positions within a document where the authorship changes. Taking the respective results into account which have shown that the problem is quite hard [39], we substantially relaxed the task this year and broke it down to the simple question: Given a document, are there any style changes or not? An alternative formulation would thus be to solely predict whether a document is written by a single author or by multiple collaborators, whereby it is irrelevant to the task to identify the exact border positions between authors. While the evaluation of the two preceding tasks relied on the Webis-TRC-12 data set [21], we created a novel data set by utilizing the StackExchange network[1]. Containing millions of publicly available questions and answers regarding several topics and subtopics, it represents a rich source which we exploited to build a comprehensive, but still realistic data set for the style change detection task.

When the collective style of groups of authors is considered, *author profiling* attempts to predict demographic and social characteristics, like age, gender, education, and personality traits. It is a research area associated with important applications in social media analytics and marketing as well as cyber forensics. In this edition of PAN, for the first time, multimodal information is considered. Both texts and images posted by social media users are used to predict their gender.

[1] https://stackexchange.com, visited June 2018

Finally, *author obfuscation* views authorship analysis from a different perspective. Given that author identification tools are available and are able to recognize the similarity within a set of texts of a certain author, the task examines what should be changed in one of these texts, maintaining its meaning, so that the author identification tools are confused. This task is strongly associated with maintaining privacy in online texts to ensure that anyone can freely express her opinion, even in countries and conditions where freedom of speech is restricted.

2 Previous Work

Two previous editions of PAN included shared tasks in authorship attribution [1,12]. However, they only examined the case where both training and test documents belong to the same domain. A relatively limited number of cross-domain authorship attribution studies has been published in the last decade. Most frequently, emphasis is put on cross-topic conditions using novels, journalistic texts, or scientific books belonging to clearly distinguished thematic areas [16,32,30,33]. Another trend is to examine cross-genre conditions using mainly literature works or social media texts (aiming to link accounts by the same user in different social networks) [14,17]. Novels in English and Spanish have also been used in the extreme case of cross-language authorship attribution [3]. To the best of our knowledge, so far there is no authorship attribution study focusing on fanfiction in cross-domain conditions.

With respect to intrinsic analyses of texts, PAN included several shared tasks in the last years. Starting from intrinsic plagiarism detection [19], the focus went from clustering authors wihin documents [35] to the detection of positions where the style, i.e., the authorship, changes [39]. Nevertheless, especially for the latter tasks the performances of submitted approaches were inferior to what was expected or even to simple baselines (e.g., [6]). Thereby approaches utilized typical stylometric features such as bags of character n-grams, frequencies of function words and other lexcial metrics, processed by algorithms operating on top to detect outliers (e.g., [28]). In general, only few approaches target a segmentation by other criteria than topic, i.e., especially by authors (e.g., [8,7,38]). With respect to the proposed style change detection task at PAN'18, i.e., to solely separate single-authored documents from multi-authored ones, no prior studies exist to the best of our knowledge.

In all previous editions of PAN, author profiling tasks focused on textual information exclusively aiming at recognizing specific demographic and social characteristics, like age, gender, native language, and personality traits of authors [25,23,22,27,26]. Most of the author profiling corpora are based on online texts, like blogs, tweets, reviews, etc.

Regarding author masking, this is the third time this task has been offered in a row [9,20]. Given the significant challenge this task offers because of the need to paraphrase a given text under the constraint to change its writing style sufficiently, we have not changed it much compared to previous years, but have kept it as is so that new submissions are immediately comparable to those of previous years: With two additional submissions this year, the total number of automatic obfuscation approaches aiming at masking authors are now up to a total of 9 submission. Instead of changing the task,

we continue to investigate new ways of evaluating and measuring the performance of obfuscation approaches, which, too, provides for an excellent challenge.

3 Author Identification

3.1 Cross-domain Authorship Attribution

Fanfiction presents an interesting benchmark case for computational authorship identification. Most of the fanfiction is nowadays produced on online platforms (such as fanfiction.net or archiveofourown.org) that are not strongly moderated, so that they accurately reflect an author's individual style. Interestingly, many fans are moreover active across different fandoms a fact that facilitate the study of authorship attribution in cross-domain conditions. Because of the explicit intertextuality (i.e. borrowings from the original canon), it can be anticipated that the style and content of the original canons have a strong influence on the fanfics, because these often aim to imitate the style of the canon's original authors. Fanfiction thus allows for exciting authorship research: do fanfiction authors generally succeed in imitating the author's stylome or does their individual fingerprint still show in the style of their fics?

Closed-set authorship attribution attempts to identify the most likely author of a text. 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 becomes quite challenging when documents of known and unknown authorship come from different domains (e.g., thematic area, genre), i.e., cross-domain authorship attribution. In this edition of PAN all documents of unknown authorship are fics of the same fandom (target fandom) while the documents of known authorship by the candidate authors are fics of several fandoms (other than the target-fandom). This can be more accurately described as *cross-fandom attribution in fanfiction*. The participants are asked to prepare a method that can handle multiple cross-fandom attribution problems. t In more detail, a cross-domain authorship attribution problem is a tuple (A, K, U), where A is the set of candidate authors, K is the set of reference (known authorship) texts, and U is 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 $a \in A$. 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 $d_U \notin D_K$.

Corpora For this edition of PAN, we have collected a large number of fanfics and their associated metadata from the authoritative community platform *Archive of Our Own*, a project of the Organization for Transformative Works ([2]). We limited our initial selection to fanfics in English (en), French (fr), Italian (it), Polish (pl), and Spanish (sp) that counted at least 500 tokens, according to the platform's own internal word count. Across all datasets, 'Harry Potter - J. K. Rowling' was typically the most frequent

[2] https://github.com/radiolarian/AO3Scraper

Table 1. The cross-domain authorship attribution corpus.

	Language	Problems	Authors	Training texts per author	Test texts per author Min	Test texts per author Max	Text length (avg. words)
Development	English	2	5,20	7	1	22	795
	French	2	5,20	7	1	10	796
	Italian	2	5,20	7	1	17	795
	Polish	2	5,20	7	1	21	800
	Spanish	2	5,20	7	1	21	832
Evaluation	English	4	5,10,15,20	7	1	17	820
	French	4	5,10,15,20	7	1	20	782
	Italian	4	5,10,15,20	7	1	29	802
	Polish	4	5,10,15,20	7	1	42	802
	Spanish	4	5,10,15,20	7	1	24	829

fandom. We therefore selected fanfics from this fandom as the target domain of all attribution problems. Only authors were admitted who contributed at least 7 texts to the non-target fandoms and at least 1 text to the target fandom.

For each language we constructed two separate datasets: a development set that participants could use to calibrate their system and an evaluation set on which the competing systems were eventually evaluated. Crucially, there was no overlap in authors between the development set and the test set (to discourage systems from overfitting on the characteristics of specific authors in the development set). To maximize the comparability of the data sets across languages, we randomly sampled 20 authors for each language and exactly 7 training texts from the non-target fandoms from their entire oeuvre. No sampling was carried out in the test material so that the number of test texts varies per author or problem. No texts shorter than 500 tokens were included and to normalize the length of longer fics, we only included the middle 1,000 tokens of texts that were longer than 1,000 tokens. Tokenization was done using NLTK's 'WordPunct-Tokenizer' [2]; our scripts heavily used the scikit-learn library [18]. The word count statistics are presented in the overview table below (Table 1). All texts were encoded as plain text (UTF8). To investigate the effect of the number of authors in an attribution problem, we provide several (downsampled) versions, containing random subsets of 5, 10, 15 and 20 authors respectively. For the early-bird evaluation, we only considered the problems of maximal number of authors (20) for each language.

Evaluation Framework Given that we deal with a closed-set classification task and the fact that the evaluation dataset is not equally distributed over the candidate authors, we decided to use the macro-averaged F1 score as an evaluation measure. Given an authorship attribution problem, for each candidate author recall and precision of the provided answers are calculated and a F1 score is provided. Then, the average F1 score over all candidate authors is used to estimate the performance of submissions for that attribution problem. Finally, submissions are ranked according to their mean macro-averaged F1 score over all available attribution problems.

To estimate the difficulty of a cross-domain authorship attribution problem and provide a challenging baseline for participants, we developed a simple but quite effective approach [30,29,33]. This method is based on character n-gram features and a support

Table 2. The evaluation results of the cross-domain authorship attribution task.

Submission	Overall	English	French	Italian	Polish	Spanish	Runtime
Custódio & Paraboni	**0.685**	0.744	**0.668**	0.676	0.482	**0.856**	00:04:27
Murauer et al.	0.643	**0.762**	0.607	0.663	0.450	0.734	00:19:15
Halvani & Graner	0.629	0.679	0.536	**0.752**	0.426	0.751	00:42:50
Mosavat	0.613	0.685	0.615	0.601	0.435	0.731	00:03:34
Yigal et al.	0.598	0.672	0.609	0.642	0.431	0.636	00:24:09
Martín dCR et al.	0.588	0.601	0.510	0.571	**0.556**	0.705	00:11:01
PAN18-BASELINE	0.584	0.697	0.585	0.605	0.419	0.615	00:01:18
Miller et al.	0.582	0.573	0.611	0.670	0.421	0.637	00:30:58
Schaetti	0.387	0.538	0.332	0.337	0.388	0.343	01:17:57
Gagala	0.267	0.376	0.215	0.248	0.216	0.280	01:37:56
López-Anguita et al.	0.139	0.190	0.065	0.161	0.128	0.153	00:38:46
Tabealhoje	0.028	0.037	0.048	0.014	0.024	0.018	02:19:14

vector machine (SVM) classifier. First, all character 3-grams that occur at least 5 times in the training texts of an attribution problem are extracted and used as features to represent both training and test texts. Then, a SVM with linear kernel is trained based on the training texts and can be used to predict the most likely author of the test texts. As shown in previous work, this simple model can be very effective in cross-domain conditions given that the number of features is appropriately defined for each specific attribution problem [32]. However, in this shared task, we use a simple version where the cutoff frequency threshold (i.e., practically, this defines the number of features) is the same (5) for any attribution problem. This approach is called PAN18-BASELINE in the rest of this paper. A Python implementation of this approach has been released to enable participants experiment with its possible variations.

Evaluation Results We received 11 submissions from research teams from several countries (Austria, Brazil, Germany, Iran (2), Israel (2), Mexico, the Netherlands, Spain, and Switzerland). All software submissions were deployed and evaluated in TIRA experimentation framework. Each submission had to analyse all attribution problems included in the evaluation corpus and it was given information about the language of the texts of each problem. Table 2 presents the mean macro-averaged F1 scores for all participants in the whole evaluation dataset and for the subset of problems in each of the five available languages.

As can be seen, 6 submissions were able to surpass the baseline, another one was very close to it and 4 submissions were clearly below it. The overall top-performing submission by Custódio & Paraboni was also the most effective one for French and especially Spanish (with a remarkable difference from the second-best approach). Moreover, the method of Halvani & Graner achieved quite remarkable results for Italian in comparison to the rest of submissions. The most difficult cases appear to be the Polish ones while the highest average results are obtained for English and Spanish. With respect to the total runtime cost of the submitted approaches, in general, the top-performing methods are also relatively fast. On the contrary, most of the methods that perform significantly lower than the baseline are also the least efficient ones.

Table 3. Performance of the cross-domain authorship attribution submissions per candidate set size.

Submission	20 Authors	15 Authors	10 Authors	5 Authors
Custódio & Paraboni	**0.648**	**0.676**	**0.739**	**0.677**
Murauer et al.	0.609	0.642	0.680	0.642
Halvani & Graner	0.609	0.605	0.665	0.636
Mosavat	0.569	0.575	0.653	0.656
Yigal et al.	0.570	0.566	0.649	0.607
Martín dCR et al.	0.556	0.556	0.660	0.582
PAN18-BASELINE	0.546	0.532	0.595	0.663
Miller et al.	0.556	0.550	0.671	0.552
Schaetti	0.282	0.352	0.378	0.538
Gagala	0.204	0.240	0.285	0.339
López-Anguita et al.	0.064	0.065	0.195	0.233
Tabealhoje	0.012	0.015	0.030	0.056

Table 3 shows the performance (macro-averaged F1 score) of the submitted methods for a varying candidate set size (from 20 authors to 5 authors). Apparently, the overall top-performing method of Custódio & Paraboni remains the most effective one for each of the examined candidate set sizes. In most cases, the ranking of participants is very similar to their overall ranking. It's also remarkable that the PAN18-BASELINE is especially effective when there are only a few (5) authors. In general, the performance of submissions improves when the candidate set becomes shorter. However, it seems that the best-performing approaches are less accurate in problems with 5 candidate authors in comparison to problems with 10 authors.

The winning method of Custódio and Paraboni [15] is an ensemble of three simple authorship attribution approaches based on character and word n-gram features and a distorted version of texts [33]. In each attribution, the most likely model is selected. The success of this approach provides evidence that the combination of several independent attribution methods is a very promising direction. Similar conclusions were drawn in previous shared tasks on author verification [34]. The second-best method according to the overall ranking is a variation of the PAN18-BASELINE that uses dynamic adaptation of parameter values for each attribution problem separately. The third-best submission is based on text compression. Apparently, methods using simple and language-independent features are more effective in this task in comparison to more sophisticated approaches based on linguistic analysis and deep learning. A more comprehensive review of submitted methods is included in the task overview paper [15].

3.2 Style Change Detection

The style change detection task at PAN 2018 attaches to a series of subtasks of previous PAN events that focused on intrinsic characteristics of text documents [19,35,39]. Considering the relatively low accuracies achieved by participants of those tasks we therefore proposed a substantially simplified task at PAN 2018 while still beeing a continuation of the previous year's style breach detection task: Given a text document, participants should apply intrinsic analyses to decide whether it is written by one or more authors, i.e., if there exist any style changes or not. With respect to the intendend

Table 4. Overview of the style change detection data set.

	Training				Validation				Test			
		Authors				Authors				Authors		
Topic/Site	**Problems**	**1**	**2**	**3**	**Problems**	**1**	**2**	**3**	**Problems**	**1**	**2**	**3**
bicycles	160	80	47	33	82	41	28	13	70	35	27	8
christianity	358	179	107	72	176	88	48	40	172	86	45	41
gaming	178	89	47	42	86	43	23	20	78	39	21	18
history	354	177	104	73	178	89	54	35	170	85	46	39
islam	166	83	49	34	86	43	31	12	72	36	20	16
linguistics	144	72	46	26	72	36	22	14	64	32	12	20
meta	196	98	56	42	94	47	30	17	90	45	30	15
parenting	178	89	54	35	92	46	32	14	78	39	27	12
philosophy	468	234	146	88	232	116	63	53	224	112	65	47
poker	100	50	35	15	48	24	14	10	42	21	13	8
politics	204	102	57	45	102	51	34	17	90	45	22	23
project man.	104	52	24	28	50	25	12	13	44	22	14	8
sports	102	51	34	17	54	27	20	7	40	20	12	8
stackoverflow	112	56	23	33	60	30	16	14	48	24	12	12
writers	156	78	43	35	80	40	25	15	70	35	18	17
	2980	**1490**	**872**	**618**	**1492**	**746**	**452**	**294**	**1352**	**676**	**384**	**292**

task simplification, it was thereby irrelevant to identify the number of style changes, the specific positions, or to build clusters of authors.

Evaluation Data To evaluate the approaches, three distinct data sets for training, validation and testing have been created using an approximate 50/25/25 split, whereby the solutions for the first two were provided. All data set are based on user posts from 15 heterogeneous sites of the Q&A network StackExchange[3], covering different topics (e.g., *programming*, *politics*, *sports* or *religion*) and subtopics (e.g., *law*, *economy* or *european union* for the *politics* topic). Using the questions and answers of users belonging to the same topic and subtopic, the final documents have been assembled by varying the following parameters:

- number of style changes (including 0 for single-authored documents)
- number of collaborating authors (1–3)
- document length (300–1000 tokens)
- allow changes only at the end or within paragraphs
- uniform or random distribution of changes with respect to segment lengths

An overview of the dataset showing the number of problems per topic, i.e., StackExchange site, is depicted in Table 4. In total 2980 training, 1492 validation and 1352 test documents have been created, whereby each text consists of the same topic/subtopic and thus making the task single-genre and single-topic. Finally, for each data set and topic the number of single-authored documents is equal to the number of multi-authored documents, resulting in a 50% accuracy baseline for random guessing. A detailed description of the data set and the creation thereof is presented in the respective task overview paper [15].

[3] https://stackexchange.com, visited June 2018

Results This year, six teams participated in the style change detection task, whereby five of them submitted their software to TIRA. The performance was thereby measured by computing the accuracy of correct predictions.

At a glance, most approaches applied a binary classification based on different more or less complex models computed from stylometric features, and only one approach used an algorithmic method based on similarity measures. The best performing approach by Zlatkova et al. utilizes a stacking technique to combine an ensemble of multiple learners. Using several feature groups (e.g., including word n-grams and typcial beginnings and endings), they at first build four different classifiers (i.e., an SVM, Random Forest, AdaBoost Trees and a multilayer perceptron) for each group to compute weighted models. Finally, a logistic regression combines these models together with a tf-idf-based gradient boosting approach to predict the final output. Safin and Ogaltsov also rely on an ensemble of three classifiers trained from common text statistics like number of sentences or punctuation frequencies, character n-grams and word n-grams. The final prediction is then calculated by a weighted sum of the classifier predictions, whereby the weightings have been tuned during preliminary experiments.

The approaches by Hosseinia et al. and Schaetti make use of different neural networks. Hosseinia et al. use two parallel recurrent neural networks (RNN) solely based on features extracted from the grammatical structure, i.e., the parse tree of sentences. To predict the appearance of style changes, they reverse the sentence order of a document, compute the respective parse tree features and integrate several similarity measures in their fusion layer to compare the reverse-order features with the original ones. On the other hand, Schaetti utilizes a character-based convolutional neural network (CNN) with three convolutional layers and 25 filters each, which does the final classification using a binary, linear layer. To train the network with more examples, the original training corpus was artificially extended by approximately a factor of 10 by sampling new documents from the available training corpus.

Finally, Khan used an algorithmic approach that at first splits a document into single sentences, builds groups thereof and computes simple word-based features. Using a sliding window technique, two consecutive sentence groups are then compared by calculating a matching score, whereby a tuned threshold determines the existence of a style change.

To be able to compare the results, three baselines have been used: (i) rnd1-BASELINE is simply guessing, (ii) rnd2-BASELINE uses a slightly enhanced guessing technique by incorporating the statistics of the training/validation datasets, which reveal that longer documents are a bit more likely to be multi-authored, and (iii) C99-BASELINE utilizes the C99 text segmentation algorithm [4] by predicting style changes if C99 found more than one segment and no changes in case it yielded only a single segment.

The final results of the five submitting teams are presented in Table 5. Zlatkova et al. could achieve the significantly best accuracy by predicting correctly 89% of all documents across all topics and subtopics. Moreover, all approaches could outperform all baselines. With respect to the runtime the two best performing approaches also needed significantly more time (due to the ensemble techniques and parse tree generation, respectively), compared to the other participants who could produce predictions within

Table 5. Evaluation results of the style change detection task.

Submission	Accuracy	Runtime
Zlatkova et al.	**0.893**	01:35:25
Hosseinia & Mukherjee	0.825	10:12:28
Safin & Ogaltsov	0.803	00:05:15
Khan	0.643	00:01:10
Schaetti	0.621	00:03:36
C99-BASELINE	0.589	00:00:16
rnd2-BASELINE	0.560	-
rnd1-BASELINE	0.500	-

minutes for the roughly 1,300 documents in the test data set. Finally, fine-grained performances depending on the different topics, subtopics and data set configurations are presented in the respective overview paper of this task [15].

4 Author Profiling

The objective of author profiling is to classify authors depending on their sociolect aspect, that is, how language is shared by people. This may allow to identify personal traits such as age, gender, native language, language variety or personality type. The interest in author profiling can be seen in the number of participants in this shared task over the last years[4], as well as the number of investigations in the field[5]. Its importance relies on the possibility of improving marketing segmentation, security or forensics. For example, using the language as evidence to detect possible cases of abuse or harassing messages, and then to profile the authors.

The Author Profiling shared task at PAN 2018 focuses on the following aspects:

- *Gender identification.* As in previous editions, the task addresses gender identification, but from a new multimodal perspective.
- *Multimodality.* Besides textual data, images can be used to profile the authors. This multimodal perspective allows to investigate whether images can help to improve gender identification beyond considering only textual features.
- *Multilinguality.* Data is provided in Arabic, English and Spanish.
- *Twitter.* Data was collected from Twitter, where its idiosyncratic characteristics may show the daily real use of the language.

[4] In the six editions of the author profiling shared task we have had respectively 21 (2013: age and gender identification [25]), 10 (2014: age and gender identification in different genre social media [23]), 22 (2015: age and gender identification and personality recognition in Twitter [22]), 22 (2016: cross-genre age and gender identification [27]), 22 (2017: gender and language variety identification [26], and 23 (2018: multimodal gender identification [24]) participating teams.

[5] The search of "author profiling" raises 1,560 results in Google Scholar: https://scholar.google.es/scholar?q="author+profiling"

4.1 Evaluation Framework

To build the PAN-AP-2018 corpus we have used a subset from the PAN-AP-2017 corpus in Arabic, English and Spanish. For each author, we tried to collect all the images shared in her timeline. Since some authors did not share images (other users closed their accounts), the PAN-AP-2018 corpus contains the subset of authors from the PAN-AP-2017 corpus that still exist and have shared at least 10 images. In Table 6 the corpus figures are shown. The corpus is completely balanced per gender and each author is composed of exactly 100 tweets.

Table 6. Number of authors per language and subset, half of them per gender. Each author is composed of 100 tweets and 10 images.

	(AR) Arabic	(EN) English	(ES) Spanish
Training	1,500	3,000	3,000
Test	1,000	1,900	2,200

The participants were asked to send three predictions per author (namely *modalities*), by using: *a)* a textual-based approach; *b)* an image-based approach; *c)* a combination of both approaches. The participants were allowed to approach the task in any language and to use any of these three approaches, although we encouraged them to participate in all languages and *modalities*[6].

The accuracy has been used for evaluation. For each language, we obtain the accuracy for each *modality*. The accuracy obtained with the combined approach has been selected as the accuracy for the given language. If the author only used the textual approach, this accuracy has been used. The final ranking has been calculated as the average accuracy per language, as shown in the following equation:

$$ranking = \frac{gender_{ar} + variety_{en} + gender_{es}}{3} \quad (1)$$

4.2 Results

This year 23 have been the teams who participated in the shared task. In Table 7 the overall performance per language and user's ranking are shown. The best results have been obtained in English (85.84%), followed by Spanish (82%) and Arabic (81.80%). As can be observed, all of them are over 80% of accuracy and most of the systems over 70% of accuracy.

The overall best result (81.98%) has been obtained by the authors in [36]. They have approached the task with deep neural networks. For textual processing, they used word embeddings from a stream of tweets with FastText skip-grams and trained a Recurrent Neural Network. For images, they used a pre-trained Convolutional Neural Network.

[6] From the 23 participants, 22 participated in Arabic and Spanish, and all of them in English. All of them approached the task with textual features, and 12 also used images.

They combined both approaches with fusion component. The authors in [5] have obtained the second best result on average (81.70%) by approaching the task only from the textual perspective. They used SVM with different types of word and character n-grams. The third best overall result (80.68%) has been obtained by the authors in [37]. They used SVM with combinations of word and character n-grams for texts and a variant of the Bag of Visual Words for images, combining both predictions with a convex linear combination. Nevertheless, there is no statistical significance among the three of them. With respect to the different languages, the best results have been obtained by the same authors. For instance, the best result in Arabic (81.80%) has been obtained by the authors in [37], the best ones in English (85.84%) by the authors in [36], and the best ones in Spanish (82%) by the authors in [5]. It is worth to mention that the only result that is significantly higher is the one obtained in English (85.84%).

Table 7. Accuracy per language and global ranking as average per language.

Ranking	Team	Arabic	English	Spanish	Average
1	Takahashi *et al.*	0.7850	**0.8584**	0.8159	**0.8198**
2	Daneshvar	0.8090	0.8221	**0.8200**	0.8170
3	Tellez *et al.*	**0.8180**	0.8068	0.7955	0.8068
4	Ciccone *et al.*	0.7940	0.8132	0.8000	0.8024
5	Kosse *et al.*	0.7920	0.8074	0.7918	0.7971
6	Nieuwenhuis & Wilkens	0.7870	0.8095	0.7923	0.7963
7	Sierra-Loaiza *et al.*	0.8100	0.8063	0.7477	0.7880
8	Martinc *et al.*	0.7780	0.7926	0.7786	0.7831
9	Veenhoven *et al.*	0.7490	0.7926	0.8036	0.7817
10	ópez-Santillán *et al.*	0.7760	0.7847	0.7677	0.7761
11	Hacohen-Kerner *et al.* (A)	0.7570	0.7947	0.7623	0.7713
12	Gopal-Patra *et al.*	0.7680	0.7737	0.7709	0.7709
13	Hacohen-Kerner *et al.* (B)	0.7570	0.7889	0.7591	0.7683
14	Stout *et al.*	0.7640	0.7884	0.7432	0.7652
15	Von Däniken *et al.*	0.7320	0.7742	0.7464	0.7509
16	Schaetti	0.7390	0.7711	0.7359	0.7487
17	Aragon & Lopez	0.6670	0.8016	0.7723	0.7470
18	Bayot & Gonçalves	0.6760	0.7716	0.6873	0.7116
19	Garibo	0.6750	0.7363	0.7164	0.7092
20	Sezerer *et al.*	0.6920	0.7495	0.6655	0.7023
21	Raiyani *et al.*	0.7220	0.7279	0.6436	0.6978
22	Sandroni-Dias & Paraboni	0.6870	0.6658	0.6782	0.6770
23	Karlgren *et al.*	-	0.5521	-	-

In Table 8 the best results per language and *modality* are shown. Results obtained with the textual approach are higher than the ones obtained with images, although very similar in case of English. It should be highlighted that the best results where obtained by combining texts and images, especially in the case of English where the improvement is higher. A more in-depth analysis of the results and the different approaches can be found in [24].

Table 8. Best results per language and modality.

Language	Textual	Images	Combined
Arabic	0.8170	0.7720	0.8180
English	0.8221	0.8163	0.8584
Spanish	0.8200	0.7732	0.8200

5 Author Obfuscation

The author obfuscation task at PAN 2018 focuses on *author masking*, which can be viewed as an attack to existing authorship verification technology. More specifically, given a pair of texts written by the same author, the task is to change the style of one of these texts so that verification algorithms are led astray and cannot detect the unique authorship anymore. Pan 2018 features the third edition of this task, whose specification follows the evaluation framework of the two previous editions [20,9]. In order to be self-contained, the following paragraphs will repeat basic information of both the data and the setup.

5.1 Evaluation Datasets

The evaluation data consist of the English portion of the combined datasets of the PAN 2013-2015 authorship verification tasks, separated by training datasets and test datasets. The datasets cover a broad range of genres: excerpts from computer science textbooks, essays from language learners, excerpts from horror fiction novels, and dialog lines from plays. As usual, the (combined) training dataset was handed out to participants, while the (combined) test dataset was held back, being accessible only via the TIRA experimentation platform. The test dataset contains a total of 464 problem instances, each consisting of a to-be-obfuscated text and one or more other texts from the same author. The approaches submitted by participants were supposed to process each problem instance and to return for each of the to-be-obfuscated texts a paraphrased version. The paraphrasing procedure was allowed to exploit the other texts from the same author in order to learn about potential style modifications that may render the writing styles of the two texts dissimilar.

5.2 Performance Measures

To measure an algorithmically achieved obfuscation performance we propose to distinguish the following three orthogonal dimensions. We call an obfuscation (similarly: an obfuscation software)

- **safe**, if the obfuscated text cannot be attributed to the original authors,
- **sound**, if the obfuscated text is textually entailed by the original text, and
- **sensible**, if the obfuscated text is well-formed and inconspicuous.

From these dimensions the safety can be automatically calculated using the TIRA versions of 44 authorship verification approaches that are at our disposal: in this regard,

we count the number of cases for which a true positive prediction of an authorship verifier is flipped to a false negative prediction after having applied the to-be-evaluated obfuscator. This is repeated for all 44 state-of-the-art verifiers.

With the current state of the art the soundness and the sensibleness of an author obfuscation approach can hardly assessed automatically; the values for these dimensions are hence based on human judgment (our as well as peer-review judgements). For this purpose, we grade a selection on a Likert scale of 1-5 with regard to sensibleness, and on 3-point scale with regard to soundness.

5.3 Results

We received 2 submissions for the author obfuscation task in addition to the 7 from the previous two years. A detailed evaluation of the results of these methods together with baselines (submissions from previous two years) is still underway at the time of writing this paper, since it requires the re-execution of the 44 authorship verifiers that have been submitted to the PAN authorship verification tasks. Evaluation results and analysis will be included in the task overview paper [10].

6 Summary

PAN 2018 shared tasks attracted a relatively large number of participants (41 submissions in total for all the tasks), comparable to previous editions of this evaluation lab. This demonstrates that there is a large and active research community in digital text forensics and PAN has become the main forum of this community. New datasets were built to support the PAN 2018 shared tasks covering several languages. One more year we required software submissions and all participant methods were evaluated in TIRA, ensuring replicability of results and facilitating the re-evaluation of these approaches using other datasets in the future.

Fanfiction texts provide an excellent material for evaluating authorship analysis methods. Focusing on cross-domain authorship attribution we were able to study how differences in fandom affect the effectiveness of attribution techniques. In general, submissions that do not require a deep linguistic analysis of texts were found to be both the most effective and the most efficient ones for this task. Heterogeneous ensembles of simple base methods and compression models outperformed more sophisticated approaches based on deep learning. Furthermore, the candidate set size is inversely correlated with the attribution accuracy especially when more than 10 authors are considered.

With the relaxation of the style change detection task at PAN 2018 we achieved to not only attract more participants, but also to significantly improve the performances of the submitted approaches. On a novel data set created from a popular Q&A network containing more than 4,000 problems, all participants achieved to surpass all provided baselines significantly by applying various techniques from machine learning ensembles to deep learning. Achieved accuracies of up to nearly 90% over the whole data set represent a good starting point to further develop and tighten the style change detection task in future PAN editions.

Author profiling was for another edition of PAN the most popular task with 23 submissions. The combination of information coming from texts and images posted by social media users seems to slightly improve the results of gender recognition. It is also notable that textual information and images when considered separately achieve comparable results. It remains to be seen whether they can be combined more effectively.

A key conclusion for author masking so far is that the task continues to be of interest to the community, albeit, it cannot compete in terms of number of participants with the other tasks. This is by no means to the detriment of the task, since we believe that the detection and prediction tasks of PAN can only truly be appreciated if the risks posed by an adversary are taken into account. In this regard, each of the aforementioned tasks have the potential of being attacked in the future, either by well-equipped individuals, or even at large by initiatives to subvert online surveillance. In this regard, we plan on recasting the obfuscation task next year, making it a bit easier to participate, yet extending its reach to other tasks.

Acknowledgments Our special thanks go to all of PAN's participants, to Symanto Group[7] for sponsoring PAN and to MeaningCloud[8] for sponsoring the author profiling shared task award. The work at the Universitat Politècnica de València was funded by the MINECO research project SomEMBED (TIN2015-71147-C2-1-P).

References

1. Argamon, S., Juola, P.: Overview of the International Authorship Identification Competition at PAN-2011. In: Petras, V., Forner, P., Clough, P. (eds.) Notebook Papers of CLEF 2011 Labs and Workshops, 19-22 September, Amsterdam, Netherlands (Sep 2011), http://www.clef-initiative.eu/publication/working-notes
2. Bird, S., Klein, E., Loper, E.: Natural Language Processing with Python. O'Reilly Media (2009)
3. Bogdanova, D., Lazaridou, A.: Cross-language authorship attribution. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014. pp. 2015–2020 (2014)
4. Choi, F.Y.: Advances in domain independent linear text segmentation. In: Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference (NAACL). pp. 26–33. Association for Computational Linguistics, Seattle, Washington, USA (April 2000)
5. Daneshvar, S.: Gender identification in twitter using n-grams and lsa. In: ANNOUNCED, T.B. (ed.) Working Notes Papers of the CLEF 2018 Evaluation Labs (Sep 2018)
6. Daniel Karaś, M.S., Sobecki, P.: OPI-JSA at CLEF 2017: Author Clustering and Style Breach Detection. In: Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)
7. Giannella, C.: An improved algorithm for unsupervised decomposition of a multi-author document. Technical Papers, The MITRE Corporation (February 2014)
8. Glover, A., Hirst, G.: Detecting stylistic inconsistencies in collaborative writing. In: The New Writing Environment, pp. 147–168. Springer (1996)

[7] https://www.symanto.net/

[8] https://www.meaningcloud.com/

9. Hagen, M., Potthast, M., Stein, B.: Overview of the Author Obfuscation Task at PAN 2017: Safety Evaluation Revisited. 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 (Sep 2017)
10. Hagen, M., Potthast, M., Stein, B.: Overview of the Author Obfuscation Task at PAN 2018. In: Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
11. Hellekson, K., Busse, K. (eds.): The Fan Fiction Studies Reader. University of Iowa Press (2014)
12. Juola, P.: An Overview of the Traditional Authorship Attribution Subtask. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy (Sep 2012), http://www.clef-initiative.eu/publication/working-notes
13. Juola, P.: The rowling case: A proposed standard analytic protocol for authorship questions. Digital Scholarship in the Humanities 30(suppl-1), i100–i113 (2015)
14. Kestemont, M., Luyckx, K., Daelemans, W., Crombez, T.: Cross-genre authorship verification using unmasking. English Studies 93(3), 340–356 (2012)
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. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
16. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research 8, 1261–1276 (2007)
17. Overdorf, R., Greenstadt, R.: Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. Proceedings on Privacy Enhancing Technologies 2016(3), 155–171 (2016)
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
19. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection. In: Notebook Papers of the 5th Evaluation Lab on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN). Amsterdam, The Netherlands (September 2011)
20. Potthast, M., Hagen, M., Stein, B.: Author Obfuscation: Attacking the State of the Art in Authorship Verification. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016), http://ceur-ws.org/Vol-1609/
21. Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: Fung, P., Poesio, M. (eds.) Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 13). pp. 1212–1221. Association for Computational Linguistics (Aug 2013), http://www.aclweb.org/anthology/P13-1119
22. 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 (Sep 2015)
23. 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: 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 (Sep 2014)

24. Rangel, F., Rosso, P., y 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)
25. 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 (Sep 2013)
26. 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 (Sep 2017)
27. 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. CEUR-WS.org (Sep 2016)
28. Safin, K., Kuznetsova, R.: Style Breach Detection with Neural Sentence Embeddings. In: Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)
29. Sapkota, U., Bethard, S., Montes, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 93–102 (2015)
30. Sapkota, U., Solorio, T., Montes, M., Bethard, S., Rosso, P.: Cross-topic authorship attribution: Will out-of-topic data help? In: Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1228–1237 (2014)
31. Stamatatos, E.: Intrinsic Plagiarism Detection Using Character n-gram Profiles. In: Stein, B., Rosso, P., Stamatatos, E., Koppel, M., Agirre, E. (eds.) SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09). pp. 38–46. Universidad Politécnica de Valencia and CEUR-WS.org (Sep 2009), http://ceur-ws.org/Vol-502
32. Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy 21, 421–439 (2013)
33. Stamatatos, E.: Authorship attribution using text distortion. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 1138–1149. Association for Computational Linguistics (2017)
34. Stamatatos, E., amd Ben Verhoeven, W.D., Juola, P., López-López, A., Potthast, M., Stein, B.: Overview of the Author Identification 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 (Sep 2015)
35. Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Clustering by Authorship Within and Across Documents. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016), http://ceur-ws.org/Vol-1609/
36. Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., Ohkuma, T.: Text and image synergy with feature cross technique for gender identification. In: ANNOUNCED, T.B. (ed.) Working Notes Papers of the CLEF 2018 Evaluation Labs (Sep 2018)
37. Tellez, E.S., Miranda-Jiménez, S., Moctezuma, D., Graff, M., Salgado, V., Ortiz-Bejar, J.: Gender identification through multi-modal tweet analysis using microtc and bag of visual words. In: ANNOUNCED, T.B. (ed.) Working Notes Papers of the CLEF 2018 Evaluation Labs (Sep 2018)

38. Tschuggnall, M., Specht, G.: Automatic decomposition of multi-author documents using grammar analysis. In: Proceedings of the 26th GI-Workshop on Grundlagen von Datenbanken. CEUR-WS, Bozen, Italy (October 2014)
39. 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: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, vol. 1866. CLEF and CEUR-WS.org (Sep 2017), http://ceur-ws.org/Vol-1866/