Overview of PAN'17
Author Identification, Author Profiling, and Author Obfuscation

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

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

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

Abstract The PAN 2017 shared tasks on digital text forensics were held in conjunction with the annual CLEF conference. This paper gives a high-level overview of each of the three shared tasks organized this year, namely author identification, author profiling, and author obfuscation. For each task, we give a brief summary of the evaluation data, performance measures, and results obtained. Altogether, 29 participants submitted a total of 33 pieces of software for evaluation, whereas 4 participants submitted to more than one task. All submitted software has been deployed to the TIRA evaluation platform, where it remains hosted for reproducibility purposes.

1 Introduction

Digital text forensics is a key area for the application of technologies that analyze writing style. For decades, scientists with various backgrounds, ranging from linguistics over natural language processing to computer security have conducted research to quantify and reliably analyze the writing style of a given text. A general goal is to find a kind of "style fingerprint", which would render its author personally identifiable when other pieces of writing known to be written by the same author are at hand. Collectively termed author identification, several subordinate tasks have been identified and extensively studied under this goal. Besides identifying authors, also the question came up whether authors who share a personal trait also expose this fact via shared writing style characteristics. If true, the author of a text of unknown authorship may still be identified via circumstantial evidence, by narrowing down the list of candidates to those whose profiles match personal traits predicted for the author of the text in question. Predicting one or many of the traits of a text's author is hence called author profiling.

However, despite decades of research the problem of reliably extracting writing style fingerprints from text is only partially solved, and, the problem of identifying style markers that are reliably correlated with personal traits is even far from being solved. At the same time, recent advances in automatic text generation based on deep neural networks have opened the door to mixtures of human-generated and machine-generated texts, rendering future writing style analyses more difficult. In this regard, also the vulnerability of writing style analysis to targeted attacks is being investigated, since text synthesis technology may be applied to alter the writing style of a text in such a way that its author cannot be reliably identified anymore, or, similarly, that wrong personal traits are predicted from it. Any systematic attempt to alter a text in such a way is called author obfuscation. Note that the question whether author identification and profiling dominate author obfuscation or vice versa is open.

At PAN, we have been addressing all of these tasks head-on—in particular, by organizing shared tasks for each of them for the past couple of years. While the specific variants of the tasks in question have changed significantly throughout the years, the underlying goal of getting to the "principles" of writing style technologies and their application remains the same. In the 2017 edition of PAN, we focus on (1) author clustering and style break detection, two tasks that belong to author identification, (2) gender and native language prediction, which belong to author profiling, and (3) author masking as a specific case of author obfuscation. Participation and interest from scientists worldwide has been strong throughout the years, and again a total of 33 teams participated in the current edition. In what follows, we briefly review each shared task and the achieved results.

2 Author Identification

In certain authorship analysis tasks, a given document could be written by multiple authors. In such cases, it is necessary to decompose the document into authorial components [14]. For instance, in intrinsic plagiarism detection, the main part of the document is assumed to be by the alleged author, while the rest of the document has been taken by other authors. In author diarization, several authors collaborated to write a document and the contribution of each one of them should be detected [30]. Tackling such a problem requires to master two basic sub-problems: to properly segment a document into stylistically homogeneous parts, and, to group these parts by authorship. The current edition of PAN focuses on exactly these two tasks, here called *style breach detection* and *author clustering*.

2.1 Style Breach Detection

The style breach detection task at PAN 2017 attaches to a series of subtasks of previous PAN events that focused on intrinsic characteristics of text documents. Including tasks like intrinsic plagiarism detection [17] or author diarization [29], one commonality is that the style of authors has to be extracted and quantified in some way in order to tackle the specific problem types. In a similar way, an intrinsic analysis of the writing style is also key to approach the PAN 2017 style breach detection task, which can be summarized as follows: Given a document, determine whether it is multi-authored, and, if in-fact it is multi-authored, find the borders where authors switch.

From a different perspective, the detection of style breaches, i.e., locating borders between different authors, can be seen as a special case of a general text segmentation problem. However, a crucial difference to existing segmentation approaches is the following: While the latter focus on detecting switches of topics or stories (e.g., [11,27,15]), the aim of style breach detection is to identify borders based on style, disregarding the content. In contrast to the author clustering task described in Section 2.2, the goal is to only find borders; it is irrelevant to identify or cluster authors of segments.

Evaluation Datasets To evaluate the approaches, distinct training and test data sets have been provided, which are based on the Webis-TRC-12 data set [20]. The original corpus contains documents on 150 topics used at the TREC Web Tracks from 2009-2011 [4], whereby professional writers were hired and asked to search for a given topic and to compose a single document from the search results. From these documents, the respective data sets have been generated randomly by varying several configurations:

- number of borders (0–8, 0 for single-author documents, i.e., containing no style breaches)
- number of collaborating authors (1–5)
- average segment length (~ 30–2500 words)
- document length (~ 200–6000 words)
- allow borders either only at the end or within paragraphs
- either uniformly or randomly distribute borders with respect to segment lengths

Parts of the original corpus have already been used and published, and we ensured that the test documents have been created from previously unpublished documents only. Overall, the number of documents in the training data set is 187, whereas the test data set contains 99 documents.

Performance Measures The performance of the submitted algorithms have been measured with two common metrics used in the field of text segmentation. The *WindowDiff* metric [16] proposed for general text segmentation evaluation is computed, because it is widely used for similar problems. It calculates an error rate between 0 and 1 for predicting borders (0 indicates a perfect prediction), by penalizing near-misses less than other/complete misses or extra borders. Depending on the problem types and data sets used, text segmentation approaches report near-perfect windowDiff values of less than 0.01, while on the other side, the error rate exceeds values of 0.6 and higher under certain circumstances [6]. A more recent adaption of the WindowDiff metric is the *WinPR* metric [28]. It enhances WindowDiff by computing the common information retrieval measures precision (WinP) and recall (WinR) and thus allows to give a more detailed, qualitative statement about the prediction. Internally, WinP and WinR are computed based on the calculation of true and false positives/negatives of border positions, respectively. It also assigns higher scores, if predicted borders are closer to the real border position.

Both metrics have been computed on word-level, whereby the participants were asked to provide character positions (i.e., the tokenization was delegated to the evaluator script). For the final ranking of all participating teams, the F-score of WinPR (WinF) is employed.

Table 1. Style breach detection results. Participants are ranked according to their WinF score.

Rank	Participant	WinP	WinR	WinF	WindowDiff
1	Karaś, Śpiewak & Sobecki	0.315	0.586	**0.323**	0.546
–	BASELINE-eq	0.337	**0.645**	0.289	0.647
2	Khan	**0.399**	0.487	0.289	**0.480**
3	Safin & Kuznetsova	0.371	0.543	0.277	0.529
–	BASELINE-rnd	0.302	0.534	0.236	0.598

Results This year, five teams participated in the style breach detection task, whereas three of them submitted their software to TIRA [7,18]. An overview in the nutshell: Karaś, Śpiewak & Sobecki use bags of 3-grams in combination with other common stylometric features that serve as input for a statistical test that determines whether or not two consecutive paragraphs are similar in style. Khan uses feature vectors based on word lists and other basic lexical metrics; borders are then detected by applying a distance function on sliding windows. Finally, Safin & Kuznetsova compute high-dimensional vectors for sentences, using a skip-thought-model [12], which can be seen as a word embedding technique operating on sentences as atomic units. The vectors then serve as input for an outlier detection analysis, similar to identifying suspicious sentences in intrinsic plagiarism detection.

To be able to compare the results, two simple baselines have been computed: *BASELINE-rnd* randomly places 0–10 borders on arbitrary positions inside a document. As a variant, *BASELINE-eq* also decides on a random basis how many borders should be placed (also 0–10), but then places the borders uniformly, i.e., such that all resulting segments are of equal size with respect to tokens contained. Both baselines have been computed based on the average of 100 runs.

The final results of the three submitting teams are presented in Table 1, which shows the average value of each computed measure (note that by doing so WinF is not the result of computing the F-score on the presented WinP and WinR values, but rather the average of the individual WinF scores). Karaś et al. could surpass the baseline equalizing the segment sizes in case of WinF, whereas the baseline using completely random positions could be exceeded by all participants. In comparison to WindowDiff, all approaches perform better than both baselines, whereby Khan achieves the best result. Finally, fine-grained sub performances depending on the data set configuration, e.g., the number of borders, are presented in the respective overview paper of this task [31].

2.2 Author Clustering

Given a small set of short (paragraph-length) documents D, the task is to group them by authorship. More specifically, we adopt the following two scenarios:

- *Complete clustering*. Each document $d \in D$ has to be assigned to exactly one of k clusters, where each cluster corresponds to a distinct author; k is not given.
- *Authorship-link ranking*. Pairs of documents by the same author (authorship-links), $(d_i, d_j) \in D \times D$, have to be extracted and ranked in decreasing order of confidence (a score belonging to [0,1]).

Table 2. The author clustering corpus. Average clusteriness ratio (*r*), number of documents (*N*), number of authors (*k*), number of authorship links, maximum cluster size (maxC), and words per document are given.

	Language	**Genre**	**Problems**	***r***	***N***	***k***	**Links**	**maxC**	**Words**
Training	English	articles	10	0.3	20	5.6	57.3	9.2	52.6
	English	reviews	10	0.3	19.4	6.1	45.4	8.2	62.2
	Dutch	articles	10	0.3	20	5.3	61.6	9.8	51.8
	Dutch	reviews	10	0.4	18.2	6.5	19.7	4.0	140.6
	Greek	articles	10	0.3	20	6.0	38.0	6.7	48.2
	Greek	reviews	10	0.3	20	6.1	41.6	7.5	39.4
Test	English	articles	20	0.3	20	5.7	59.3	9.5	52.5
	English	reviews	20	0.3	20	6.4	43.5	7.9	65.3
	Dutch	articles	20	0.3	20	5.7	49.4	8.3	49.3
	Dutch	reviews	20	0.4	18.4	7.1	19.3	4.1	152.0
	Greek	articles	20	0.3	19.9	5.2	59.6	9.6	46.6
	Greek	reviews	20	0.3	20	6.0	42.2	7.6	37.1

All documents within a clustering problem are single-authored, written in the same language, and belong to the same genre. However, topic and text-length may vary. The main difference when compared to the corresponding PAN-2016 task is that the documents are short, including a few sentences only. This makes the task harder since text-length is crucial when attempting to extract stylometric information.

Evaluation Datasets The datasets used for training and evaluation were extracted from the corresponding PAN-2016 corpora [30]. They include clustering problems in three languages (English, Dutch, and Greek) and two genres (articles and reviews). Each PAN-2016 text was segmented into paragraphs; all paragraphs with less than 100 characters or more than 500 characters were discarded. In each clustering problem, documents by the same authors were selected randomly from all original documents. This means that paragraphs of the same original document or other documents (by the same author) may be grouped. The only exception in this process was the Dutch reviews corpus since its texts were already short (one paragraph each). For this special case, the PAN-2017 datasets were built using the PAN-2016 procedure.

Table 2 shows the details about the training and test datasets. Most of the clustering problems include 20 documents (paragraphs) by an average of 6 authors. In each clustering problem there is an average of about 50 authorship links, whereas the largest cluster contains about 8 documents. Each document has an average of about 50 words. Note that in the case of Dutch reviews these figures deviate from the norm (documents are longer and authorship links are less).

An important factor to each clustering problem is the *clusteriness ratio* $r = k/N$, where N is the size of D. When r is high, most documents belong to single-item clusters, and there are only few authorship links. When r is low, most documents belong to multi-item clusters, and there are plenty of authorship links. In this new corpus r ranges between 0.1 and 0.5 in both training and test datasets, in contrast to the PAN-2016 corpus where $r \geq 0.5$ [30].

Evaluation Framework The same evaluation measures introduced in PAN-2016 are used here [30]. For the complete clustering scenario, Bcubed Recall, Bcubed Precision, and Bcubed F-score are calculated. These are among the best extrinsic clustering evaluation measures [1]. With respect to the authorship-link ranking scenario, Mean Average Precision (MAP), R-precision, and P@10 are used to estimate the ability of systems to rank high correct results.

In order to understand the complexity of tasks and the effectiveness of participant systems, we used a set of baseline approaches and applied them to the evaluation datasets. The baseline methods range from naive to strong and will allow to estimate weaknesses and strengths of participant approaches. More specifically, the following baseline methods were used:

- *BASELINE-Random*. k is randomly chosen from [1,N] and then each $d \in D$ is randomly assigned to one cluster. Authorship links are extracted from the produced clusters and a random score is calculated. The average performance of this method over 30 repetitions is reported. This naive approach can only serve as an indication of the lowest performance.
- *BASELINE-Singleton*. This method sets $k = N$, i.e., all documents are from different authors. It forms singleton clusters and it is used only for the complete clustering scenario. This simple method was found very effective in PAN-2016 datasets, and its performance increases with r [30].
- *BASELINE-Cosine*. Each document is represented by the normalized frequencies of all words occurring at least 3 times in the given collection of documents. Then, for each pair of documents the cosine similarity is used as an authorship-link score. This simple method was found hard-to-beat in PAN-2016 evaluation campaign [30].
- *BASELINE-PAN16*. This is the top-performing method submitted to the corresponding PAN-2016 task. It is based on a character-level recurrent neural network and it is a modification of an effective authorship verification approach [2].

Results We received 6 software submissions that were evaluated in TIRA experimentation platform [7,18]. Table 3 shows the overall evaluation results for both complete clustering and authorship-link ranking on the entire test dataset. The elapsed runtime of each submission is also reported. As can be seen, the method of Gómez-Adorno et al. [8] achieves the best results in both scenarios. Actually, this is the top-performing method taking into account all but one evaluation measures (BCubed precision). By definition, BASELINE-Singleton achieves perfect Bcubed precision since it provides single-item clusters exclusively. BASELINE-PAN16 also tends to optimize precision since it was tuned for another corpus with much higher clusteriness ratio. Within the submitted methods, the approaches of García et al. [5] and Kocher & Savoy [13] are the best ones in terms of Bcubed precision. However, the winning approach of Gómez-Adorno et al. [8] is the only one that achieves both Bcubed recall and precision higher than 0.6. All submitted methods but one surpass baseline approaches in the complete clustering scenario. On the other hand, in the authorship-link ranking scenario, BASELINE-PAN16 is very competitive while BASELINE-Cosine surpasses half of submissions. More detailed evaluation results are presented in [31].

Table 3. Overall evaluation results in author clustering (mean values for all clustering problems). Participants are ranked according to Bcubed F-score.

Participant	**Complete clustering**			**Authorship-link ranking**			**Runtime**
	B^3 F	B^3 rec.	B^3 prec.	MAP	RP	P@10	
Gómez-Adorno et al.	**0.573**	**0.639**	0.607	**0.456**	**0.417**	**0.618**	00:02:06
García et al.	0.565	0.518	0.692	0.381	0.376	0.535	00:15:49
Kocher & Savoy	0.552	0.517	0.677	0.396	0.369	0.509	00:00:42
Halvani & Graner	0.549	0.589	0.569	0.139	0.251	0.263	00:12:25
Alberts	0.528	0.599	0.550	0.042	0.089	0.284	00:01:46
BASELINE-PAN16	0.487	0.331	0.987	0.443	0.390	0.583	50:17:49
Karaś et al.	0.466	0.580	0.439	0.125	0.218	0.252	00:00:26
BASELINE-Singleton	0.456	0.304	**1.000**	–	–	–	–
BASELINE-Random	0.452	0.339	0.731	0.024	0.051	0.209	–
BASELINE-Cosine	–	–	–	0.308	0.294	0.348	–

In general, almost all submitted approaches are quite efficient and can process all evaluation datasets quickly. The approaches of García et al. [5] and Halvani & Graner [10] are relatively slower compared to the other submissions but, however, much more efficient than BASELINE-PAN16. The most successful approaches use low-level (character or lexical) features [5,8,10,13]. Relatively simple clustering algorithms, like hierarchical agglomerative clustering [8] or β-compact graph-based clustering [5] provided the best results in the complete clustering scenario. A more detailed survey of submissions is included in [31].

3 Author Profiling

Author profiling aims at classifying authors in different classes depending on their sociolect aspects, namely how they share language. This allows the identification of author traits such as age, gender, native language, language variety, or personality type. Author profiling is growing in interest, specially due to the rise of social media, where authors may hide personal information, or even lie. Author profiling may help to improve marketing segmentation, security, and it allows the use of the language as evidence in possible cases of abuse or harassing messages.

In previous editions at PAN we have mainly focused on age and gender identification in different genres or in a cross-genre environment. The Author Profiling shared task at PAN'17 focuses on the following aspects:

- *Gender and language variety identification.* As in previous editions, the task contains gender prediction. Instead of age identification, the aim this year is at discriminating among different varieties of the same languages (also known as dialects).
- *Demographic idiosyncrasies.* This is the first time the gender dimension is studied together with the language variety, which may provide insights on the difficulty of the task depending on geographical and cultural idiosyncrasies.
- *Multilinguality.* Participants are provided with data in Arabic, English, Spanish and Portuguese.

Table 4. Languages and varieties. There are 500 authors per variety and gender, 300 for training and 200 for test. Each author contains 100 tweets.

(AR) Arabic	(EN) English	(ES) Spanish	(PT) Portuguese
Egypt	Australia	Argentina	Brazil
Gulf	Canada	Chile	Portugal
Levantine	Great Britain	Colombia	
Maghrebi	Ireland	Mexico	
	New Zealand	Peru	
	United States	Spain	
		Venezuela	
4,000	6,000	7,000	2,000

3.1 Evaluation Framework

The evaluation data has been collected from Twitter in four different languages, namely Arabic, English, Spanish and Portuguese. The authors have been annotated with their gender and language variety. The gender annotation has been carried out with the help of dictionaries of proper nouns, and the variety has been based on the geographical retrieval of the tweets. For each author, we considered exactly 100 tweets. The dataset is balanced by gender and variety. There are 500 authors per variety and gender. The dataset has been split in a 60/40 proportion with 300 authors for training and 200 for test. The corresponding languages and varieties are shown in Table 4, together with the total number of authors for each subtask.

For evaluation, the accuracy for variety, gender and joint identification per language is calculated. Then, we average the results obtained per language (Eq. 1).

$$\begin{aligned}\overline{gender} &= \frac{gender\_ar + gender\_en + gender\_es + gender\_pt}{4} \\ \overline{variety} &= \frac{variety\_ar + variety\_en + variety\_es + variety\_pt}{4} \\ \overline{joint} &= \frac{joint\_ar + joint\_en + joint\_es + joint\_pt}{4}\end{aligned} \tag{1}$$

The final ranking is calculated as the average of the previous values (Eq. 2):

$$ranking = \frac{\overline{gender} + \overline{variety} + \overline{joint}}{3} \tag{2}$$

In order to understand the complexity of the subtasks in each language and with the aim at comparing the performance of the participants approaches, we propose the following baselines:

- *BASELINE-stat.* It is a statistical baseline that emulates the random choice. This baseline depends on the number of classes: 2 in case of gender identification, and from 2 to 7 in case of variety identification.
- *BASELINE-bow.* This method represents documents as a bag-of-words with the 1,000 most common words in the training set, weighting by absolute frequency

Table 5. Joint accuracies per language and global ranking as average per language of gender, variety and joint identification.

Ranking	Team	Global	Arabic	English	Spanish	Portuguese
1	nissim17	**0.8361**	**0.6831**	**0.7429**	**0.8036**	0.8288
2	martinc17	0.8285	0.6825	0.7042	0.7850	0.8463
3	miranda17	0.8258	0.6713	0.7267	0.7621	0.8425
4	miura17	0.8162	0.6419	0.6992	0.7518	**0.8575**
5	lopezmonroy17	0.8111	0.6475	0.7029	0.7604	0.8100
6	markov17	0.8097	0.6525	0.7125	0.7704	0.7750
7	poulston17	0.7942	0.6356	0.6254	0.7471	0.8188
8	sierraloaiza17	0.7822	0.5694	0.6567	0.7279	0.8113
	BASELINE-LDR	0.7325	0.5888	0.6357	0.6943	0.7763
9	romanov17	0.7653	0.5731	0.6450	0.6846	0.7775
10	benajiba17	0.7582	0.5688	0.6046	0.7021	0.7525
11	schaetti17	0.7511	0.5681	0.6150	0.6718	0.7300
12	kodiyan17	0.7509	0.5688	0.6263	0.6646	0.7300
13	zampieri17	0.7498	0.5619	0.5904	0.6764	0.7575
14	kheng17	0.7176	0.5475	0.5704	0.6400	0.6475
15	ganesh17	0.6881	0.5075	0.4713	0.5614	0.7300
16	kocher17	0.6813	0.5206	0.4650	0.4971	0.7575
17	akhtyamova17	0.6270	0.2875	0.4333	0.5593	0.6675
	BASELINE-bow	0.6195	0.1794	0.4713	0.5561	0.7588
18	khan17	0.4952	0.3650	0.1900	0.2189	0.5488
	BASELINE-stat	0.2991	0.1250	0.0833	0.0714	0.2500
19	ribeirooliveira17	0.2087	-	-	-	0.7538
20	alrifai17	0.1701	0.5638	-	-	-
21	bouzazi17	0.1027	-	0.2479	-	-
22	castrocastro17	0.0695	-	0.1017	-	-

of occurrence. The texts are preprocessed in order to lowercase words, remove punctuation signs and numbers, and remove the stop words for the corresponding language.

- *BASELINE-LDR[23].* This method represents documents on the basis of the probability distribution of occurrence of their words in the different classes.

3.2 Results

This year 22[1] have been the teams who participated in the shared task. In this section a summary of the obtained results is shown. In Table 5 the overall performance per language and users' ranking are shown. We can observe that the best results were achieved in Portuguese (85.75%), followed by Spanish (80.36%), English (74.29%) and Arabic (68.31%). The difference on accuracy among languages is very significant. Most of the participants obtained better results than both baselines. However, in case of Portuguese only 9 teams outperformed the bag-of-words baseline, showing the power of simple

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

Table 6. Best results per language and task.

Language	*Joint*	Gender	Variety
Arabic	0.6831	0.8031	0.8313
English	0.7429	0.8233	0.8988
Spanish	0.8036	0.8321	0.9621
Portuguese	0.8575	0.8700	0.9838

words to discriminate among varieties and genders in that language. On the contrary, this baseline shows its inefficiency in case of Arabic, where the accuracy drops to values close to the statistical baseline.

In Table 6 the best results per language and task are shown. We can observe that for both the gender and variety subtasks, the best results were achieved in Portuguese, followed by Spanish, English and Arabic. In case of gender identification, the accuracies are between 80.31% in case of Arabic and 87% in case of Portuguese, whereas the difference is higher for language variety identification, where the worst results obtained in Arabic is 83.13% (4 varieties), against a 98.38% obtained in Portuguese (2 varieties). Results for Spanish (7 varieties) (96.21%) are close to Portuguese, while in English (6 varieties) they fall to 89.88%. A more in-depth analysis of the results and the different approaches can be found in [25].

4 Author Obfuscation

Author obfuscation is the youngest branch of PAN's main tasks, and perhaps also one of the most difficult ones. Its goal is to attack the approaches to the other tasks by altering their text input in a way that will cause them return an incorrect answer. The difficulty arises not so much from making changes to the texts that have such an effect on the attacked approaches, but to do so in a way so that the text input can still be understood by a human and so that its original message is not twisted beyond recognition. The latter two severely limit the potential changes that can be made, rendering any form of obfuscation a form of style paraphrasing where the goal is to change the writing style of a piece of writing without changing a text's pragmatics.

The author obfuscation task at PAN 2017 concerns author masking, where the specific goal is to attack authorship verification technology. For the latter, the task is to verify whether a given pair of texts has been written by the same author, whereas for the former, the task is to alter the writing style of a designated text from a given pair written by the same author in order to prevent verification algorithms from arriving at just that decision. As a shared task, author masking has been organized for the first time at PAN 2016 [19]. We continue with author masking in much the same way as before. Since the setup did no change significantly, just to be self-contained, the following gives only a brief recap.

4.1 Evaluation Datasets

The evaluation data consist of the English portion of the joint datasets of the PAN 2013-2015 authorship verification tasks, separated by training datasets and test datasets. The datasets cover a broad range of genres, namely computer science textbook excerpts, essays from language learners, horror fiction novel excerpts, and dialog lines from plays. The joint training dataset was handed out to participants, while the joint test dataset was held back and only accessible 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 and paraphrased version, perhaps using the remaining texts from the same author to learn what style changes are at least necessary to make the writing styles of the two texts the most dissimilar.

4.2 Performance Measures

We call an obfuscation software

- **safe**, if its obfuscated texts can not be attributed to their original authors anymore,
- **sound**, if its obfuscated texts are textually entailed by their originals, and
- **sensible**, if its obfuscated texts are well-formed and inconspicuous.

Any evaluation of an author obfuscation approach must at least cover these three dimensions, whereas the assessment and quantification of especially the latter two is still an open problem. To cut a long story short, in this shared task, we evaluate the safety of a submitted approach by feeding the obfuscated evaluation dataset that it produces to as many pre-trained authorship verification approaches as possible. Fortunately, with 44 authorship verifiers, the number of such approaches available to us is rather high, allowing for meaningful conclusions. This is made possible due to the fact that TIRA has been employed at PAN since before 2013, so that all authorship verification approaches submitted to the corresponding shared tasks are available to us in working condition. By counting the number of cases where a true positive prediction of an authorship verifier is flipped to a false negative prediction because of applying a to-be-evaluated obfuscator beforehand, we can calculate the relative impact the obfuscator has on the verifier. When doing so not just for one verifier, but for 44 state-of-the-art verifiers, this tells a lot about the ability of the obfuscator to fulfill its purpose of obfuscating the writing style of texts in a way that cannot be defeated by any verifier known to date.

Regarding soundness and sensibleness of an author obfuscation approach, we rely own judgment as well as on peer-review. Here, we grade a selection of Likert scale of 1-5 with regard to sensibleness, and on 3-point scale with regard to soundness. In the past, the participants who participated in peer-evaluation came up with similar grade scales, obtained results commensurate with ours.

4.3 Results

A detailed evaluation of the results of a total of 5 obfuscation approaches, two of which have been submitted this year, and three of which past year can be found in the task overview paper [9].

5 Summary

Altogether, PAN presented its participants again with a set of challenging shared tasks, including new ones as well as "classical" ones which were given new spin. Multilingual as well as multigenre corpora have been prepared, which will henceforth serve as new benchmark datasets for their tasks. At the same time, the software underlying each of the submitted approaches has been collected and hosted on the TIRA experimentation platform, ensuring replicability of results as well as reproducibility, e.g., by allowing for their reevaluation using new datasets as they arrive in the future. In this regard, for future work, we will continue to develop PAN's shared tasks, providing new and challenging datasets as well as inventing new tasks belonging to author identification, author profiling, and author obfuscation.

Acknowledgements

Our special thanks go to all of PAN's participants, to Symanto Group[2] for sponsoring PAN and to MeaningCloud[3] 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. Amigó, E., Gonzalo, J., Artiles, J., Verdejo, F.: A comparison of extrinsic clustering evaluation metrics based on formal constraints. Information Retrieval 12(4), 461–486 (2009)
2. Bagnall, D.: Authorship Clustering Using Multi-headed Recurrent Neural Networks—Notebook for PAN at CLEF 2016. In: Balog et al. [3], http://ceur-ws.org/Vol-1609/
3. Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.): CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal. CEUR Workshop Proceedings, CEUR-WS.org (2016), http://www.clef-initiative.eu/publication/working-notes
4. Clarke, C.L., Craswell, N., Soboroff, I., Voorhees, E.M.: Overview of the TREC 2009 web track. Tech. rep., DTIC Document (2009)
5. García, Y., Castro, D., Lavielle, V., noz, R.M.: Discovering Author Groups Using a β-compact Graph-based Clustering. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Working Notes. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)
6. Glavaš, G., Nanni, F., Ponzetto, S.P.: Unsupervised text segmentation using semantic relatedness graphs. Association for Computational Linguistics (2016)
7. Gollub, T., Stein, B., Burrows, S.: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Hersh, B., Callan, J., Maarek, Y., Sanderson, M. (eds.) 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12). pp. 1125–1126. ACM (Aug 2012)

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

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

8. Gómez-Adorno, H., Aleman, Y., no, D.V., Sanchez-Perez, M.A., Pinto, D., Sidorov, G.: Author Clustering using Hierarchical Clustering Analysis. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Working Notes. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)
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. Halvani, O., Graner, L.: Author Clustering based on Compression-based Dissimilarity Scores. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Working Notes. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)
11. Hearst, M.A.: TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational linguistics 23(1), 33–64 (1997)
12. Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A., Fidler, S.: Skip-thought vectors. In: Advances in neural information processing systems (NIPS). pp. 3294–3302 (2015)
13. Kocher, M., Savoy, J.: UniNE at CLEF 2017: Author Clustering. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Working Notes. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)
14. Koppel, M., Akiva, N., Dershowitz, I., Dershowitz, N.: Unsupervised decomposition of a document into authorial components. In: Lin, D., Matsumoto, Y., Mihalcea, R. (eds.) Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL). pp. 1356–1364 (2011)
15. Misra, H., Yvon, F., Jose, J.M., Cappe, O.: Text segmentation via topic modeling: an analytical study. In: Proceedings of CIKM 2009. pp. 1553–1556. ACM (2009)
16. Pevzner, L., Hearst, M.A.: A critique and improvement of an evaluation metric for text segmentation. Computational Linguistics 28(1), 19–36 (2002)
17. 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)
18. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
19. 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/
20. 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
21. 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 Workshop Proceedings, CEUR-WS.org (Sep 2015)

22. 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 Workshop Proceedings, CEUR-WS.org (Sep 2014)
23. Rangel, F., Rosso, P., Franco-Salvador, M.: A low dimensionality representation for language variety identification. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing. Springer-Verlag, LNCS, arXiv:1705.10754 (2016)
24. 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)
25. 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)
26. 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 et al. [3]
27. Riedl, M., Biemann, C.: TopicTiling: a text segmentation algorithm based on LDA. In: Proceedings of ACL 2012 Student Research Workshop. pp. 37–42. Association for Computational Linguistics (2012)
28. Scaiano, M., Inkpen, D.: Getting more from segmentation evaluation. In: Proceedings of the 2012 conference of the north american chapter of the association for computational linguistics: Human language technologies. pp. 362–366. Association for Computational Linguistics (2012)
29. 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/
30. 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)
31. 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, CLEF and CEUR-WS.org (Sep 2017)