Linda Cappellato and Nicola Ferro and Gareth Jones and Eric San Juan (eds.):
CLEF 2015 Labs and Workshops, Notebook Papers, 8-11 September, Toulouse, France.
CEUR Workshop Proceedings. ISSN 1613-0073, http://ceur-ws.org/Vol-1391/, 2015.

Overview of the Author Identification Task at PAN 2015

Efstathios Stamatatos,[1] Walter Daelemans,[2] Ben Verhoeven,[2] Patrick Juola,[3]
Aurelio López-López,[4] Martin Potthast,[5] and Benno Stein[5]

[1]University of the Aegean, Greece
[2]University of Antwerp, Belgium
[3]Duquesne University, USA
[4]INAOE, Mexico
[5]Bauhaus-Universität Weimar, Germany

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

Abstract This paper presents an overview of the author identification task at PAN-2015 evaluation lab. Similar to previous editions of PAN, this shared task focuses on the problem of author verification: given a set of documents by the same author and another document of unknown authorship, the task is to determine whether or not the known and unknown documents have the same author. However, in contrast to the setup of PAN-2013 and PAN-2014, as well as most previous work in this area, it is no longer assumed that all documents match in genre and topic. In other words, we study cross-topic and cross-genre author verification, a challenging, yet realistic, task. A new evaluation corpus was built, covering the four languages Dutch, English, Greek, and Spanish and comprising a variety of genres and topics. A total of 18 teams participated in this task. Following the practice of previous PAN editions, software submissions were required and evaluated within the evaluation-as-a-service platform TIRA. Based on TIRA, we were able to define challenging baseline models using submissions from the corresponding shared tasks at PAN-2013 and PAN-2014. Analytical evaluation results are given, including statistical significance tests. Moreover, we examine the performance of a heterogeneous ensemble that combines all participant models, and we present a comprehensive review of the submitted methods.

1 Introduction

The main idea behind author identification relies on the assumption that it is possible to reveal the author of a text given (i) a set of candidate authors and (ii) a set of undisputed text samples for each one of them [16, 44]. Writing style is the most crucial information source to solve this task, and it is essential to be able to quantify stylistic choices in texts and measure stylistic similarity between texts. Beyond its traditional literary applications (e.g., verifying the authorship of disputed novels, identifying the author of works published anonymously, etc.) [17, 48] author identification is associated with important forensic applications (e.g. revealing the author of harassing messages in social media, linking terrorist proclamations by their author, etc.) [1, 25].

Author identification can be formulated in various ways, depending on the number of candidate authors and whether the set of candidate authors is closed or open.

One particular variation is the task of *authorship verification*, where there is exactly one candidate author with undisputed text samples, the task is to decide whether an unknown text is by that author or not [12, 23, 27]. In more detail, the authorship verification task corresponds to a one-class classification problem, where the samples of known authorship by the author in question form the target class. All texts written by other authors are viewed as the outlier class, a huge and heterogeneous class, which renders finding representative samples difficult. However challenging, authorship verification is a fundamental problem since any given author identification problem can be decomposed into a set of authorship verification problems. Therefore, it provides an excellent research field to examine competitive approaches aiming at the extraction of reliable and general conclusions [24].

Previous PAN editions focused on the authorship verification task; a number of evaluation corpora covering several natural languages and genres have been created [18, 46]. Moreover, a suitable evaluation framework was developed, highlighting the ability of methods to leave problems unanswered when there is high uncertainty, as well as to assign probability scores to their answers. However, the previous editions of PAN, as well as most work in the literature assume that all texts within a verification case match for both genre and topic. This assumption simplifies the problem, since style is affected by genre in addition to the personal style of each author. Moreover, low-frequency stylistic features are heavily affected by topic nuances. Thus when all documents match in genre and topic, the personal style of the authors would be the major discriminating factor between texts.

PAN-2015 still focuses on authorship verification, but it is no longer assumed that all texts within a verification problem match for genre and topic. This cross-genre and cross-topic variation of the verification task corresponds to a more realistic view of the problem at hand, since, in many applications, it is not possible to obtain text samples of undisputed authorship by certain authors in specific genres and topics. For instance, verifying the authorship of a suicide note, it does not make sense to look for samples of suicide notes by the suspects [6]. In addition, the author of a anonymously published crime fiction novel might be a famous child fiction author who has never published a crime fiction novel before [17].

A new cross-genre and cross-topic corpus was built, covering four languages and a variety of genres and topics. We received 18 software submissions that were evaluated on the TIRA experimentation platform [9, 37]. Following the practice of previous PAN editions, we also examine the performance of baseline models, based on submissions to the corresponding tasks in PAN-2013 and PAN-2014, and train a heterogeneous ensemble classifier that fuses the output of all submitted methods as if they were on.

The remainder of this paper is organized as follows: the next section describes related work in cross-genre and cross-topic author identification. Section 3 presents the evaluation framework of our shared task on author identification at PAN-2015, and Section 4 describes the new evaluation corpus. Section 5 reports on evaluation results obtained, including tests of statistical significance. Then, Section 6 presents a review of the submitted methods, and Section 7 summarizes the main conclusions and discusses directions for future work.

2 Related Work

A review of related work on authorship verification, including the results of previous editions of PAN with respect to this task, is given in [46]. Most of the related work on authorship verification—and author identification in general—concerns only cases where the examined documents match for genre and topic [12, 24, 27, 47]. A notable exception has been reported by Koppel et al. [23], who apply *unmasking* to authorship verification problems where multiple topics were covered by each author, producing very reliable results. Kestemont et al. [19] use the same method in a cross-genre authorship verification experiment on a corpus of prose and theatrical works by a number of authors, demonstrating that unmasking (with default settings) is ineffective in such difficult cases.

In general, a study focusing on cross-genre and cross-topic authorship attribution, where a closed set of candidate authors is used (a simpler case in comparison to authorship verification) is presented in [45]: a corpus of opinion articles covering multiple topics and book reviews, all published in a UK newspaper, was used, and experimental results revealed that character n-gram features are more robust with respect to word features in cross-topic and cross-genre conditions. More recently, Sapkota et al. [39] show that character n-grams corresponding to word affixes, including punctuation marks, are the most significant features in cross-topic authorship attribution. In addition, Sapkota et al. [40] demonstrate that using training texts from multiple topics instead of a single topic can significantly help to correctly recognize the author of texts on another topic.

3 Evaluation Setup

The evaluation setup for this task is basically identical to the one used for PAN-2014. Given a set of documents known to be written by the same author, and exactly one document of unknown authorship, the task is to determine whether the latter document is written by the same author as the former ones. Text length varies from a few hundred to a few thousand words, depending on genre. It is also assumed that positive and negative answers have equal prior probabilities. The only difference to PAN-2014 is that texts within a problem do not necessarily match for genre and/or topic.

Participants are asked to submit software that provides a [0,1]-normalized score corresponding to the probability of a positive answer (i.e., the known documents and the questioned document are by the same author) for each verification problem. It is possible to leave some problems unanswered by assigning a probability score of exactly 0.5. The evaluation of the provided answers is based on two scalar measures: the Area Under the *Receiver Operating Characteristic* Curve (AUC) [7], and c@1 [34]. The former tests the ability of methods to rank scores appropriately, assigning low values to negative problems and high values to positive problems. The latter rewards methods that leave problems unanswered rather than providing wrong answers. Finally, the participating teams are ranked by the final score (AUC $\cdot$ c@1).

Baselines One of the advantages of using TIRA for the evaluation of software submissions is its support for the continuous evaluation of software against newly developed corpora. This enables us to apply software that has been submitted to previous editions

of PAN to the cross-genre and cross-topic corpora of PAN-2015. Furthermore, we can avoid the use of simplistic random-guess baselines (corresponding to a final score of 0.25), and, establish more challenging baselines, adapted to the difficulty of the corpus. These baselines reveals if a newly submitted software performs better than state-of-the-art models. We employ the following three baselines:

- PAN13-BASELINE: The best-performing software submitted to PAN-2013 by Jankowska et al. [15]. This software also served as baseline in PAN-2014 [46].
- PAN14-BASELINE-1: The second-best software submitted to PAN-2014 by Fréry et al. [8].[1]
- PAN14-BASELINE-2: The third-best software submitted to PAN-2014 by Castillo et al. [4]
- PAN15-ENSEMBLE: Following previous PAN editions, we train a meta-model that combines all participant approaches [18, 38, 46]. A heterogeneous ensemble is built based on the average of scores returned by participants for the verification problems of our evaluation corpus.

Note that the baseline obtained from PAN-2013 and PAN-2014 have been trained and fine-tuned using different corpora, and under the assumption that all documents within a problem instance match for genre and topic. Therefore, their performance on cross-genre and cross-topic author verification corpora will not be optimal.

4 Evaluation Corpus

Although it is rather simple to compile a corpus of texts by different authors that belong to different genres/topics (i.e., negative instances of the authorship verification task), it is a lot more challenging to populate the corpus with corresponding positive instances (i.e., texts in different genres/topics by the same author). A new corpus was built that matches the size of the PAN-2014 evaluation corpus, and that covers the same four languages: Dutch, English, Greek, and Spanish. The corpus is divided into a training part, which is released to participants, and a test part, which is used to compute the official evaluation results. Table 1 shows key figures of the corpus.

There are notable differences between the sub-corpora for each language. In the English part, only one known document per problem is provided. In Dutch and Greek parts, the number of known documents per problem vary, whereas, in the Spanish part, there are always four known documents per problem. The documents of Greek and Spanish parts are, on average, longer than those of the Dutch and English parts. For all languages, positive and negative instances are equally distributed.

The Dutch part of our evaluation corpus is a modified version of Verhoeven and Daelemans [50]'s *CLiPS Stylometry Investigation* corpus, which comprises documents from two genres (essays and reviews), written by language students at the University of Antwerp between 2012 and 2014. The English part is a collection dialog lines from plays, excluding speaker names, stage directions, lists of characters, and so on. All positive verification instances comprise parts from different plays by the same author.

[1] Due to some technical problems it was not possible to also test the first PAN-2014 winner

Table 1. Overview of the PAN-2015 cross-genre and cross-topic authorship verification corpus.

	Language	Type	Problems	Documents	Avg. known documents	Avg. words document
Training	Dutch	cross-genre	100	276	1.76	354
	English	cross-topic	100	200	1.00	366
	Greek	cross-topic	100	393	2.93	678
	Spanish	mixed	100	500	4.00	954
Test	Dutch	cross-genre	165	452	1.74	360
	English	cross-topic	500	1000	1.00	536
	Greek	cross-topic	100	380	2.80	756
	Spanish	mixed	100	500	4.00	946
	Σ		1,265	3,701	1.93	641

The English part is the largest one in terms of verification problems. The Greek part is a collection of opinion articles, published at the online forum Protagon,[2] where all documents are categorized into several categories (e.g., Politics, Economy, Science, Health, Media, Sports, etc). For all verification problems of the Greek part, the category of the known documents is the same, but different that of the questioned document. The Spanish part consists of opinion articles taken from a variety of online newspapers and magazines, as well as personal web pages or blogs covering, each covering a variety of topics. It also includes literary essays. This part mixes cross-topic and cross-genre problems, where some problems comprise documents that are noticeable different in both topic and genre, and others match for genre but differ in topic.

5 Evaluation Results

In total, 18 teams submitted their software for this task. The submitted author verification approaches processed each language of the corpus separately using the TIRA experimentation platform. During evaluation, participants did not have access to standard output, standard error, and the evaluation results of their systems. The PAN organizers served as moderators to verify the successful execution of each participant's software. The majority of the 18 teams were able to process all four language parts of the evaluation corpus.

Table 2 compiles the final score (AUC $\cdot$ c@1) of all teams for each language of our corpus, alongside micro-averaged and macro-averaged scores. The performances of the 3 baselines and that of the ensemble can also be seen. Since the English part is much larger with respect to the number of problems, the macro-averaged score provides for a fair overall picture of the capabilities of each team's approach across all languages. On average, the best results were achieved for the cross-topic Greek part. Quite predictably, the cross-genre Dutch part proved to be the most challenging one, followed by the English part (this can be explained by the low number of known documents per problem). Note that the Greek and Spanish parts comprise longer texts (on average more than 500

[2] http://www.protagon.gr

Table 2. Evaluation results for authorship verification at PAN-2015 in terms of AUC · c@1.

Team (sorted alphabetically)	**Dutch**	**English**	**Greek**	**Spanish**	**Micro-avg**	**Macro-avg**
Bagnall [2]	0.451	**0.614**	**0.750**	0.721	**0.608**	**0.628**
Bartoli et al. [3]	0.518	0.323	0.458	**0.773**	0.417	0.506
Castro-Castro et al. [5]	0.247	0.520	0.391	0.329	0.427	0.365
Gómez-Adorno et al. [10]	0.390	0.281	0.348	0.281	0.308	0.323
Gutierrez et al. [11]	0.329	0.513	0.581	0.509	0.479	0.478
Halvani [13]	0.455	0.458	0.493	0.441	0.445	0.462
Hürlimann et al. [14]	0.616	0.412	0.599	0.539	0.487	0.538
Kocher and Savoy [21]	0.218	0.508	0.631	0.366	0.435	0.416
Maitra et al. [28]	0.518	0.347	0.357	0.352	0.378	0.391
Mechti et al. [29]	–	0.247	–	–	0.207	0.063
Moreau et al. [30]	**0.635**	0.453	0.693	0.661	0.534	0.606
Nikolov et al. [31]	0.089	0.258	0.454	0.095	0.217	0.201
Pacheco et al. [33]	0.624	0.438	0.517	0.663	0.480	0.558
Pimas et al. [35]	0.262	0.257	0.230	0.240	0.253	0.247
Posadas-Durán et al. [36]	0.132	0.400	–	0.462	0.333	0.226
Sari and Stevenson [41]	0.381	0.201	–	0.485	0.286	0.250
Solórzano et al. [43]	0.153	0.259	0.330	0.218	0.242	0.235
Vartapetiance and Gillam [49]	0.262	–	0.212	0.348	0.177	0.201
PAN15-ENSEMBLE	0.426	0.468	0.537	0.715	0.475	0.532
PAN14-BASELINE-1 [8]	0.255	0.249	0.198	0.443	0.269	0.280
PAN14-BASELINE-2 [4]	0.191	0.409	0.412	0.683	0.406	0.405
PAN13-BASELINE [15]	0.242	0.404	0.384	0.367	0.358	0.347

words per document), while Dutch and English parts comprise shorter texts (less than 500 words per document).

In terms of micro-averaged and macro-averaged final score, the submissions of Bagnall [2] and Moreau et al. [30] clearly outperform the rest of the participants, respectively. The former seems to be particularly effective for cross-topic verification, but seems to be affected by differences in genre, judging by the low performance on the Dutch part. The latter is very effective for cross-genre verification on the Dutch part, whereas its performance is worse on the English part where only one known document per problem is available. Most of the rest of participants did not manage to achieve notable performances across all four corpora. For example, Bartoli et al. [3] achieves good results for Dutch and Spanish, but fails to be competitive in English and Greek, while the picture is reversed for Kocher and Savoy [21]. Exceptions to the rule are the approaches of Pacheco et al. [33] and Hürlimann et al. [14].

Unlike the evaluation results of PAN-2013 and PAN-2014 [18, 46], the ensemble of all participants is not the best-performing approach. With respect to the micro-averaged and macro-averaged final scores, the ensemble is outperformed by 5 and 4 participants, respectively. An explanation for the mediocre performance of the meta-model can be found in the low average performances of the submitted approaches. This is demonstrated by the fact that all PAN-2014 participants achieved a micro-averaged final score greater than 0.3 while 6 out of 18 PAN-2015 participants achieve a micro-averaged final score lower than 0.3—a score close to the final score of 0.25 of a random-guessing model.

A more detailed picture of the evaluation results can be found in Table 3, where, apart from the final score (FS), also ROC AUC, c@1, the number of Unanswered Problems (UP), and the runtime are reported for Dutch, English, Greek, and Spanish, re-

Table 3. Evaluation results for authorship verification at PAN-2015 per language in terms of final score (FS=AUC · c@1), area under the curve (AUC) of the receiver operating characteristic (ROC), c@1, unanswered problems (UP), and runtime.

(a) English

Team	FS	AUC	c@1	UP	Runtime
Bagnall [2]	**0.614**	**0.811**	**0.757**	3	21:44:03
Castro-Castro et al. [5]	0.520	0.750	0.694	0	02:07:20
Gutierrez et al. [11]	0.513	0.739	0.694	39	00:37:06
Kocher and Savoy [21]	0.508	0.738	0.689	94	00:00:24
PAN15-ENSEMBLE	0.468	0.786	0.596	0	–
Halvani [13]	0.458	0.762	0.601	25	00:00:21
Moreau et al. [30]	0.453	0.709	0.638	0	24:39:22
Pacheco et al. [33]	0.438	0.763	0.574	2	00:15:01
Hürlimann et al. [14]	0.412	0.648	0.636	5	00:01:46
PAN14-BASELINE-2	0.409	0.639	0.640	0	00:26:19
PAN13-BASELINE	0.404	0.654	0.618	0	00:02:44
Posadas-Durán et al. [36]	0.400	0.680	0.588	0	01:41:50
Maitra et al. [28]	0.347	0.602	0.577	10	15:19:13
Bartoli et al. [3]	0.323	0.578	0.559	3	00:20:33
Gómez-Adorno et al. [10]	0.281	0.530	0.530	0	07:36:58
Solórzano et al. [43]	0.259	0.517	0.500	0	00:29:48
Nikolov et al. [31]	0.258	0.493	0.524	16	00:01:36
Pimas et al. [35]	0.257	0.507	0.506	0	00:07:22
PAN14-BASELINE-1	0.249	0.537	0.464	159	00:01:11
Mechti et al. [29]	0.247	0.489	0.506	0	00:04:59
Sari and Stevenson [41]	0.201	0.401	0.500	0	00:05:47
Vartapetiance and G. [49]	0.000	0.500	0.000	500	–

(b) Greek

Team	FS	AUC	c@1	UP	Runtime
Bagnall [2]	**0.750**	0.882	**0.851**	5	10:07:49
Moreau et al. [30]	0.693	**0.887**	0.781	10	07:07:42
Kocher and Savoy [21]	0.631	0.822	0.768	20	00:00:11
Hürlimann et al. [14]	0.599	0.788	0.760	0	00:01:01
Gutierrez et al. [11]	0.581	0.802	0.725	5	00:28:32
PAN15-ENSEMBLE	0.537	0.779	0.690	0	–
Pacheco et al. [33]	0.517	0.773	0.670	3	00:02:01
Halvani [13]	0.493	0.767	0.643	9	00:00:17
Bartoli et al. [3]	0.458	0.698	0.657	1	00:07:45
Nikolov et al. [31]	0.454	0.709	0.640	0	00:01:01
PAN14-BASELINE-2	0.412	0.634	0.650	0	00:01:22
Castro-Castro et al. [5]	0.391	0.621	0.630	0	00:17:59
PAN13-BASELINE	0.384	0.641	0.600	0	00:01:46
Maitra et al. [28]	0.357	0.613	0.582	4	06:22:48
Gómez-Adorno et al. [10]	0.348	0.590	0.590	0	00:09:22
Solórzano et al. [43]	0.330	0.590	0.560	0	00:12:56
Pimas et al. [35]	0.230	0.480	0.480	0	00:03:58
Vartapetiance and G. [49]	0.212	0.460	0.460	0	00:36:30
PAN14-BASELINE-1	0.198	0.484	0.410	28	00:00:30
Mechti et al. [29]	0.000	0.500	0.000	100	–
Posadas-Durán et al. [36]	0.000	0.500	0.000	100	–
Sari and Stevenson [41]	0.000	0.500	0.000	100	–

(c) Dutch

Team	FS	AUC	c@1	UP	Runtime
Moreau et al. [30]	**0.635**	**0.825**	**0.770**	0	08:09:35
Pacheco et al. [33]	0.624	0.822	0.759	30	00:05:08
Hürlimann et al. [14]	0.616	0.808	0.762	1	00:00:38
Maitra et al. [28]	0.518	0.759	0.683	4	02:32:48
Bartoli et al. [3]	0.518	0.751	0.689	1	00:07:01
Halvani [13]	0.455	0.709	0.642	8	00:00:09
Bagnall [2]	0.451	0.700	0.644	2	12:00:43
PAN15-ENSEMBLE	0.426	0.696	0.612	0	–
Gómez-Adorno et al. [10]	0.390	0.625	0.624	0	83:58:15
Sari and Stevenson [41]	0.381	0.613	0.621	4	00:02:04
Gutierrez et al. [11]	0.329	0.592	0.556	5	00:40:32
Vartapetiance and G. [49]	0.262	0.512	0.512	1	00:44:51
Pimas et al. [35]	0.262	0.508	0.515	0	00:02:27
PAN14-BASELINE-1	0.255	0.506	0.503	0	00:00:17
Castro-Castro et al. [5]	0.247	0.503	0.491	0	00:05:51
PAN13-BASELINE	0.242	0.506	0.479	0	00:00:47
Kocher and Savoy [21]	0.218	0.449	0.484	18	00:00:07
PAN14-BASELINE-2	0.191	0.422	0.452	16	00:02:10
Solórzano et al. [43]	0.153	0.397	0.385	4	00:10:25
Posadas-Durán et al. [36]	0.132	0.382	0.346	54	36:39:07
Nikolov et al. [31]	0.089	0.256	0.348	1	00:00:47
Mechti et al. [29]	0.000	0.500	0.000	165	–

(e) Spanish

Team	FS	AUC	c@1	UP	Runtime
Bartoli et al. [3]	**0.773**	**0.932**	**0.830**	0	00:09:16
Bagnall [2]	0.721	0.886	0.814	10	11:21:41
PAN15-ENSEMBLE	0.715	0.894	0.800	0	–
PAN14-BASELINE-2	0.683	0.823	0.830	0	00:04:03
Pacheco et al. [33]	0.663	0.908	0.730	0	00:04:23
Moreau et al. [30]	0.661	0.853	0.775	25	15:27:31
Hürlimann et al. [14]	0.539	0.739	0.730	0	00:01:29
Gutierrez et al. [11]	0.509	0.755	0.674	7	00:24:20
Sari and Stevenson [41]	0.485	0.724	0.670	0	00:03:48
Posadas-Durán et al. [36]	0.462	0.680	0.680	0	02:20:35
PAN14-BASELINE-1	0.443	0.692	0.640	0	00:00:45
Halvani [13]	0.441	0.704	0.627	23	00:00:14
PAN13-BASELINE	0.367	0.656	0.560	0	00:02:37
Kocher and Savoy [21]	0.366	0.650	0.564	20	00:00:22
Maitra et al. [28]	0.352	0.610	0.577	3	10:36:31
Vartapetiance and G. [49]	0.348	0.590	0.590	0	00:48:37
Castro-Castro et al. [5]	0.329	0.558	0.590	0	00:23:54
Gómez-Adorno et al. [10]	0.281	0.530	0.530	0	00:50:41
Pimas et al. [35]	0.240	0.490	0.490	0	00:04:12
Solórzano et al. [43]	0.218	0.454	0.480	0	00:11:18
Nikolov et al. [31]	0.095	0.280	0.340	0	00:01:09
Mechti et al. [29]	0.000	0.500	0.000	100	–

spectively. In these tables, participants as well as the baseline models and the PAN15-ENSEMBLE are ranked according to their final score.

The performance of the baseline models reflects the difficulty of the evaluation corpora. In the Dutch cross-genre part, all three baselines do note improve much above a random-guessing classifier. The PAN13-BASELINE and the PAN14-BASELINE-2 provide relatively good results for the English and Greek cross-topic parts, while the performance of PAN14-BASELINE-1 is considerably lower. This may be explained by the fact that the latter method is based on eager supervised learning, so that it depends too much on the properties of its original training corpus [8]. Both the PAN14-BASELINE-1 and the PAN14-BASELINE-2 are perform significantly better when applied to the mixed Spanish part, where some verification problems match the properties of PAN-2014 corpora. On average, the PAN13-BASELINE and the PAN14-

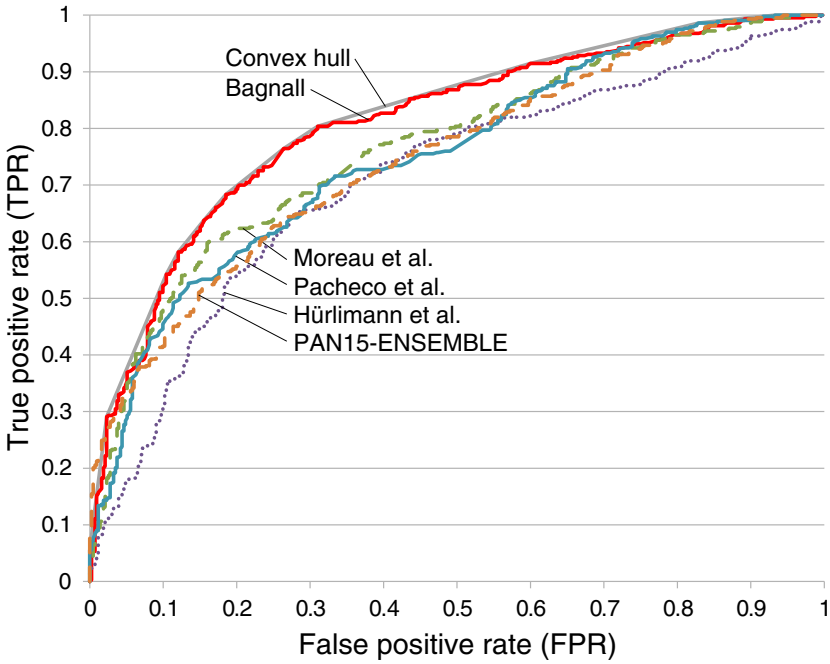

Figure 1. ROC curves of the best-performing approaches, the PAN15-ENSEMBLE, and the convex hull of all 18 participants for the entire evaluation corpus.

BASELINE-2 outperform almost half of the participating teams, demonstrating their potential as generic approaches that can be used on any given corpus. On the other hand, the average performance of the PAN14-BASELINE-1 resembles random-guessing.

Based on the performance of the baseline models and the PAN15-ENSEMBLE, we can divide the 18 submitted approaches into 3 rough categories for each language:

- *Remarkable.* Approaches whose performance is better than PAN15-ENSEMBLE.
- *Good.* Approaches whose performance is higher than PAN13-BASELINE and lower than PAN15-ENSEMBLE.
- *Poor.* Approaches whose performance is lower than PAN13-BASELINE.

ROC Curves To obtained a more insights into the performances of the submitted methods, Figure 1 shows the ROC curves of the top-4 participants alongside the convex hull of all 18 participants, and the ROC curve of PAN15-ENSEMBLE for the entire evaluation corpus (865 verification problems). As can be seen, the convex hull of the submitted methods is almost completely dominated by the winning approach of Bagnall [2]. However, at very low/high FPR values, the approach of Pacheco et al. [33] performs better. These points correspond to the case where false positives/negatives have a very high cost. The performance of the PAN15-ENSEMBLE is also very competitive for such extreme cases, especially for low FPR values.

Table 4. Results of statistical significance tests for the entire evaluation corpus.

	Bartoli et al. [3]	Castro-Castro et al. [5]	Gómez-Adorno et al. [10]	Gutierrez et al. [11]	Halvani [13]	Hürlimann et al. [14]	Kocher and Savoy [21]	Maitra et al. [28]	Mechti et al. [29]	Moreau et al. [30]	Nikolov et al. [31]	Pacheco et al. [33]	Pimas et al. [35]	Posadas-Durán et al. [36]	Sari and Stevenson [41]	Solórzano et al. [43]	Vartapetiance and Gillam [49]	PAN15-ENSEMBLE	PAN13-BASELINE [15]	PAN14-BASELINE-1 [8]	PAN14-BASELINE-2 [4]
Bagnall [2]	***	***	***	***	***	*	***	***	***	**	***	***	***	***	***	***	***	***	***	***	***
Bartoli et al. [3]		=	***	=	*	**	**	=	***	*	***	=	***	***	***	***	***	=	*	***	=
Castro-Castro et al. [5]			***	=	**	*	***	*	***	=	***	=	***	***	***	***	***	=	**	***	=
Gómez-Adorno et al. [10]				**	=	***	=	=	***	***	**	***	*	***	***	**	***	***	=	***	**
Gutierrez et al. [11]					**	*	***	*	***	*	***	=	***	***	***	***	***	=	**	***	=
Halvani [13]						***	=	=	***	***	***	=	**	***	***	***	***	**	=	***	*
Hürlimann et al. [14]							***	***	***	=	***	**	***	***	***	***	***	*	***	***	**
Kocher and Savoy [21]								=	***	***	**	*	=	**	*	*	***	***	=	***	**
Maitra et al. [28]									***	***	***	=	***	***	***	***	***	*	=	***	=
Mechti et al. [29]										***	***	***	***	***	***	***	**	***	***	***	***
Moreau et al. [30]											***	**	***	***	***	***	***	*	***	***	*
Nikolov et al. [31]												***	=	=	=	=	***	***	***	*	***
Pacheco et al. [33]													***	***	***	***	***	=	=	***	=
Pimas et al. [35]														=	=	=	***	***	**	***	***
Posadas-Durán et al. [36]															=	=	***	***	***	*	***
Sari and Stevenson [41]																=	***	***	***	**	***
Solórzano et al. [43]																	***	***	***	**	***
Vartapetiance and Gillam [49]																		***	***	***	***
PAN15-ENSEMBLE																			**	***	=
PAN13-BASELINE [15]																				***	*
PAN14-BASELINE-1 [8]																					***

Statistical significance tests Following PAN-2014 [46], we compute the statistical significance of performance differences between all examined approaches using *approximate randomization testing* [32]. This non-parametric test does not make assumptions that do not hold for the performance measures used, and it can handle complicated distributions. We did a pairwise comparison of the accuracy of all approaches based on this method and the results are shown in Table 4. The null hypothesis is that there is no difference in the output of two approaches. When the probability of accepting the null hypothesis is $p > 0.05$, we consider there to be no significant difference of the output of two approaches (denoted as =). When $0.01 < p < 0.05$ the difference is significant (denoted as *), when $0.001 < p < 0.01$ the difference is very significant (denoted as **), and when $p < 0.001$ the difference is highly significant (denoted as ***).

The overall performance of the winning approach of Bagnall [2] is significantly better compared to the rest of the submissions as well as the baseline methods, and the ensemble of all submissions. It should be noted that in most cases the difference is highly significant. The second best-performing approach by Moreau et al. [30] is also significantly better compared to the remaining approaches, with two exceptions: Castro-Castro et al. [5] and Hürlimann et al. [14]. Beyond the first two winners, it is noteworthy that the approach of Hürlimann et al. [14] is significantly different from the rest of submitted approaches. Moreover, the group of methods from Bartoli et al. [3], Gutierrez et al. [11], Halvani [13], Kocher and Savoy [21], Maitra et al. [28], and Pacheco et al. [33]) achieves reasonably good performances, but in most of their pairwise comparisons, no statistically significant difference between them can be observed.

6 Review of Submitted Methods

The overall best-performing approach of Bagnall [2] in terms of both micro-averaged and macro-averaged final score introduces a character-level Recurrent Neural Network model. The success of this model demonstrates that character-level information can be used in elaborate models to enhance performance compared to naive character n-gram frequencies. The second best-performing approach by Moreau et al. [30] is based on a heterogeneous ensemble combined with stacked generalization. The success of this model verifies the conclusions of previous editions of PAN that different verification models, when combined, can achieve very good results [18, 46]. It should be noted that both winning approaches require remarkably high computational cost. To allow for a quick comparison between the submitted approaches, Table 5 compiles an overview of their basic characteristics. In the remainder of this section, we review the submitted approaches in closer detail.

Verification Model There are two main categories of verification models, namely intrinsic and extrinsic model. The *intrinsic* models only use the texts within a verification problem (the known documents by one author and the unknown document) to arrive at their decision. Usually, they handle the verification task as a one-class classification problem [15, 4, 8]. In addition to that, the *extrinsic* models also use other texts by different authors and attempt to transform the verification task to a binary classification problem [42, 20, 24].

Table 5. Basic characteristics of the submitted approaches.

Team (alphabetically)	Verification model	Learning type	Attribution paradigm	Elaborate Text Analysis
Bagnall [2]	extrinsic	lazy	instance-based	none
Bartoli et al. [3]	intrinsic	eager	instance-based	POS tagging
Castro-Castro et al. [5]	extrinsic	lazy	instance-based	POS tagging, lemmatization
Gómez-Adorno et al. [10]	intrinsic	lazy	profile-based	syntactic parsing
Gutierrez et al. [11]	extrinsic	lazy	instance-based	none
Halvani [13]	intrinsic	lazy	profile-based	none
Hürlimann et al. [14]	intrinsic	eager	instance-based	none
Kocher and Savoy [21]	extrinsic	lazy	profile-based	none
Maitra et al. [28]	intrinsic	eager	instance-based	POS tagging
Mechti et al. [29]	extrinsic	eager	instance-based	POS tagging
Moreau et al. [30]	extrinsic	eager	instance-based	LDA, POS tagging
Nikolov et al. [31]	intrinsic	eager	hybrid	none
Pacheco et al. [33]	extrinsic	eager	profile-based	LDA, POS tagging, lemmatization
Pimas et al. [35]	intrinsic	eager	instance-based	style/grammar checking
Posadas-Durán et al. [36]	intrinsic	lazy	instance-based	syntactic parsing
Sari and Stevenson [41]	intrinsic	eager	hybrid	none
Solórzano et al. [43]	intrinsic	eager	instance-based	POS tagging
Vartapetiance and Gillam [49]	intrinsic	lazy	instance-based	none

Both in PAN-2013 and PAN-2014, the overall best-performing approach employed extrinsic models; more specifically, variations of the *impostors* method [20, 42]. Likewise, most of the best-performing submissions to PAN-2015, including the two top-performing ones, employ extrinsic models [2, 5, 11, 21, 30, 33]. The impostors method is part of the approach of Moreau et al. [30], while Gutierrez et al. [11] and Kocher and Savoy [21] propose modifications thereof. The best-performing intrinsic models are proposed by Bartoli et al. [3], Halvani [13], Hürlimann et al. [14], and Maitra et al. [28]. It should be noted that the performance of the latter approaches on the cross-genre Dutch corpus were remarkable.

Learning algorithm The submitted author verification methods can be further distinguished by their approach to supervised learning, namely *eager* methods and *lazy* methods. The former make use of supervised learning algorithms to extract a general model of the verification problems, based on the training data. Such methods strongly depend on the size, quality and representativeness of the training data. Only a few eager methods were submitted in previous PAN editions, including that of Fréry et al. [8], whereas the majority of submissions to PAN-2015 belong to this category. Well-known and popular supervised machine learning algorithms were used, like SVMs [14, 29, 31, 35, 41], random forest [3, 28, 33], and genetic algorithms [30].

Lazy methods do not apply any eager supervised learning algorithm, but make a decision based on information extracted for each verification problem separately. The winning approach of Bagnall [2], as well as some other submissions that achieve very good performance [11, 13, 21], belong to this category.

Attribution Paradigm In author identification two attribution paradigms are distinguished [44]: the *instance-based* paradigm attempts to capture the style of documents by representing each document separately [2, 3, 11, 14, 30]. The *profile-based* paradigm attempts to capture the style of authors by computing a single representation for all texts written by the same author, a so-called author profile. The latter approach is generally more robust when few texts (in quantity or length) of known authorship are available. In comparison to PAN-2013 and PAN-2014 an increased number of participants followed the profile-based paradigm [10, 13, 21, 33]. Moreover, in a hybrid of the two paradigms, separate representations are extracted for each document written by the same author which are then combined into a single representation [31, 41].

Text Representation Following the practice of participants in previous PAN editions, low-level and language-independent measures are the main kinds of features used to represent the writing style of documents. Typical examples are lengths of words, sentences, and paragraphs, type-token ratio, hapax legomena, and other vocabulary richness and readability measures. A very popular type of features is character n-grams (including unigrams), words, punctuation marks, stopwords, etc. Many submitted approaches rely exclusively on such text representation features, disregarding features that require more sophisticated text analyses [2, 11, 13, 14, 21, 31, 41, 49].

Regarding more sophisticated features, the most popular ones are part-of-speech (POS) n-grams mainly due to the availability of POS taggers of acceptable performance for all four languages of the PAN-2015 corpus [3, 5, 28, 29, 33, 43]. A few participants apply full syntactic parsing, achieving moderate performances at the cost of considerably increased runtime cost [10, 36]. Other features requiring more elaborate text analysis are related to lemmatization [5, 33], style and grammar checking [35], and Latent Dirichlet Allocation [30, 33].

The majority of participants attempt to combine different types of features. Some approaches, however, use only one type of features, for example, the most frequent terms [21], stopword n-grams [49], and character sequences [2]. Also, some of the proposed features do not refer to a single document but capture the difference of a certain feature between two documents (typically one of known and one of unknown authorship). These features are called *differential* features [3] or *joint* features [14], and they are used in combination with eager supervised learning models.

Handling Ambiguous Cases An important issue in author verification is the ability to leave problems unanswered when unsure, rather than providing wrong answers. This capability of an approach is directly measured using c@1. Some of the participants, including the two top-performing ones, attempt to focus on this issue and leave some problems unanswered when the confidence of their answers is low. The most basic approach is to examine the score of each problem leave it unanswered if it lies in a specified range around 0.5 [2, 21, 30]. A more sophisticated model is proposed by Moreau et al. [30], whose classifier determines ambiguous cases that are left unanswered to improve c@1.

7 Conclusions

The shared task on author identification at PAN-2015 focused on the authorship verification problem. In contrast to previous editions of PAN, a major novelty was that cross-genre and cross-topic verification cases were considered. This challenging, yet realistic variation of the problem allowed us to examine whether authorship verification methods are heavily affected by variations in genre and topic among the documents of a verification case. The evaluation results indicate that the cross-genre scenario is more difficult. However, the performance of top-ranked approaches on each language of our corpus is surprisingly high in terms of both AUC and c@1.

The two top-performing methods introduce significant novelties. The winning approach of Bagnall [2] is based on a character-level neural network language model that is used for first time in authorship verification. The success of this model indicates that, beyond the well-known and simplistic character n-gram features, more complex approaches can better exploit character-level information in authorship analysis tasks. The second winning approach of Moreau et al. [30] takes advantage of a heterogeneous ensemble that combines different authorship verification methods, which implements and verifies one of the major conclusion drawn at PAN-2013 and PAN-2014 [18, 46]. Both of these approaches are computationally expensive. However, the increase in runtime cost is not caused by elaborate text analysis methods, such as syntactic parsing or semantic analysis. Rather, they runtime is spend on fine-tuning parameters of the learning algorithms.

We received 18 submissions, which compares to the corresponding tasks at PAN-2013 (18 participants) and PAN-2014 (13 participants). Among them, only five participants also took part in the shared task at PAN-2013 and/or PAN-2014 [11, 13, 30, 35, 49]. These figures verify that there is a lively research community working on author identification tasks, and that PAN has become the major forum of this research. We may also claim that the focus of PAN on the authorship verification problem helped to raise interest of researchers on that fundamental problem and to significantly advance the state of the art. Moreover, the availability of the PAN corpora allowed the development of novel methods that are based on eager supervised learning algorithms. Future PAN editions will examine in more detail whether the size, diversity, and quality of training corpora strongly affect the performance of verification models.

Text-length is one important issue that has not been thoroughly studied within authorship verification research. How long should the texts of known authorship be in order to allow for training reliable verification models? How many words of the unknown documents are really needed to allow for computing an accurate answer? Answers to such and similar questions are critical in case we wish to apply this technology to short texts, like tweets and SMS messages. Another interesting future direction is to study the relationship of authorship verification with other author identification tasks, like *author clustering* (grouping documents by authorship) [26] and *author diarization* (segmenting a multi-author document into authorial components) [22].

Acknowledgements

This work was partially supported by the WIQ-EI IRSES project (Grant No. 269180) within the FP7 Marie Curie action.

Bibliography

[1] Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. Intelligent Systems, IEEE 20(5), 67–75 (2005)

[2] Bagnall, D.: Author Identification using multi-headed Recurrent Neural Networks. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)

[3] Bartoli, A., Dagri, A., Lorenzo, A.D., Medvet, E., Tarlao, F.: An Author Verification Approach Based on Differential Features. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)

[4] Castillo, E., Cervantes, O., Vilariño, D., Pinto, D., León, S.: Unsupervised method for the authorship identification task. In: CLEF 2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2014)

[5] Castro-Castro, D., Pelaez-Brioso, M., Adame-Arcia, Y., noz Guillena, R.M.: Authorship Verification, Combining Linguistic Features and Different Similarity Functions. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)

[6] Chaski, C.E.: Who's at the keyboard: Authorship attribution in digital evidence invesigations. International Journal of Digital Evidence 4 (2005)

[7] Fawcett, T.: An introduction to roc analysis. Pattern Recogn. Lett. 27(8), 861–874 (2006)

[8] Fréry, J., Largeron, C., Juganaru-Mathieu, M.: Ujm at clef in author identification. In: CLEF 2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2014)

[9] 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)

[10] Gómez-Adorno, H., Sidorov, G., Pinto, D., Markov, I.: A Graph Based Authorship Identification Approach. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)

[11] Gutierrez, J., Casillas, J., Ledesma, P., Fuentes, G., Meza, I.: Homotopy Based Classification for Author Verification Task. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)

[12] van Halteren, H.: Linguistic profiling for author recognition and verification. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. ACL '04, Association for Computational Linguistics, Stroudsburg, PA, USA (2004)
[13] Halvani, O.: A Generic Authorship Verification Scheme Based on Equal Error Rates. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[14] Hürlimann, M., Weck, B., van den Berg, E., Suster, S., Nissim, M.: GLAD: Groningen Lightweight Authorship Detection. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[15] Jankowska, M., Keselj, V., Milios, E.: Proximity based one-class classification with Common N-Gram dissimilarity for authorship verification task. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain (2013)
[16] Juola, P.: Authorship Attribution. Foundations and Trends in Information Retrieval 1, 234–334 (2008)
[17] Juola, P.: How a computer program helped reveal J. K. Rowling as author of A Cuckoo's Calling. Scientific American (2013)
[18] Juola, P., Stamatatos, E.: Overview of the author identification task at pan-2013. In: P., T.D.E.F. (ed.) Notebook Papers of CLEF 2013 LABs and Workshops (CLEF-2013) (2013)
[19] Kestemont, M., Luyckx, K., Daelemans, W., Crombez, T.: Cross-genre authorship verification using unmasking. English Studies 93(3), 340–356 (2012)
[20] Khonji, M., Iraqi, Y.: A slightly-modified gi-based author-verifier with lots of features (asgalf). In: CLEF 2014 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2014)
[21] Kocher, M., Savoy, J.: UniNE at CLEF 2015: Author Identification. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[22] 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. pp. 1356–1364 (2011)
[23] Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. J. Mach. Learn. Res. 8, 1261–1276 (2007)
[24] Koppel, M., Winter, Y.: Determining if two documents are written by the same author. Journal of the American Society for Information Science and Technology 65(1), 178–187 (2014)
[25] Lambers, M., Veenman, C.: Forensic authorship attribution using compression distances to prototypes. In: Geradts, Z., Franke, K., Veenman, C. (eds.) Computational Forensics, Lecture Notes in Computer Science, vol. 5718, pp. 13–24. Springer Berlin Heidelberg (2009)
[26] Layton, R., Watters, P., Dazeley, R.: Automated unsupervised authorship analysis using evidence accumulation clustering. Natural Language Engineering 19, 95–120 (2013)

[27] Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. In: Proceedings of the Twenty-Second International Conference on Computational Linguistics (COLING 2008). pp. 513–520. Coling 2008 Organizing Committee, Manchester, UK (2008)
[28] Maitra, P., Ghosh, S., Das, D.: Authorship Verification: An Approach based on Random Forest. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[29] Mechti, S., Jaoua, M., Faiz, R., Bsir, B., Belguith, L.H.: On the Empirical Evaluation of Hybrid Author Identification Method. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[30] Moreau, E., Jayapal, A., Lynch, G., Vogel, C.: Author Verification: Basic Stacked Generalization Applied To Predictions from a Set of Heterogeneous Learners. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[31] Nikolov, S., Tabakova, D., Savov, S., Kiprov, Y., Nakov, P.: SU@PAN'2015: Experiments in Author Veri
cation. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[32] Noreen, E.: Computer-Intensive Methods for Testing Hypotheses: An Introduction. A Wiley-Interscience publication, Wiley (1989)
[33] Pacheco, M.L., Fernandes, K., Porco, A.: Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[34] Peñas, A., Rodrigo, A.: A simple measure to assess non-response. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 1415–1424. HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA (2011)
[35] Pimas, O., Kröll, M., Kern, R.: Know-Center at PAN 2015 Author Identification. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[36] Posadas-Durán, J.P., Sidorov, G., Batyrshin, I., Mirasol-Meléndez, E.: Author Verification Using Syntactic N-grams. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[37] 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 (2014)
[38] Potthast, M., Stein, B., Holfeld, T.: Overview of the 1st International Competition on Wikipedia Vandalism Detection. In: Braschler, M., Harman, D.,

Pianta, E. (eds.) Working Notes Papers of the CLEF 2010 Evaluation Labs (Sep 2010), http://www.clef-initiative.eu/publication/working-notes
[39] Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL. pp. 93–102 (2015)
[40] Sapkota, U., Solorio, T., Montes-y-Gómez, M., Bethard, S., Rosso, P.: Cross-topic authorship attribution: Will out-of-topic data help? In: COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland. pp. 1228–1237 (2014)
[41] Sari, Y., Stevenson, M.: A Machine Learning-based Intrinsic Method for Cross-topic and Cross-genre Authorship Verification. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[42] Seidman, S.: Authorship Verification Using the Impostors Method. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain (2013)
[43] Solórzano, J., Mijangos, V., Pimentel, A., López-Escobedo, F., Montes, A., Sierra, G.: Authorship Verification by Combining SVMs with Kernels Optimized for Different Feature Categories. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[44] Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology 60, 538–556 (2009)
[45] Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy 21, 421–439 (2013)
[46] Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., Sánchez-Pérez, M.A., Barrón-Cedeño, A.: Overview of the author identification task at PAN 2014. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 877–897 (2014)
[47] Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic text categorization in terms of genre and author. Comput. Linguist. 26(4), 471–495 (2000)
[48] Stover, J.A., Winter, Y., Koppel, M., Kestemont, M.: Computational authorship verification method attributes a new work to a major 2nd century african author. Journal of the Association for Information Science and Technology (2015)
[49] Vartapetiance, A., Gillam, L.: Adapting for Subject-Specific Term Length using Topic Cost in Author Verification. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
[50] Verhoeven, B., Daelemans, W.: Clips stylometry investigation (csi) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In: Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014). Reykjavik, Iceland (2014)