Overview of the Author Identification Task at PAN 2014

Efstathios Stamatatos[1], Walter Daelemans[2], Ben Verhoeven[2], Martin Potthast[3], Benno Stein[3], Patrick Juola[4], Miguel A. Sanchez-Perez[5], and Alberto Barrón-Cedeño[6]

[1]University of the Aegean, Greece
[2]University of Antwerp, Belgium
[3]Bauhaus-Universität Weimar, Germany
[4]Duquesne University, USA
[5]Instituto Politécnico Nacional, Mexico
[6]Universitat Politècnica de Catalunya, Spain

Abstract. The author identification task at PAN-2014 focuses on author verification. Similar to PAN-2013 we are given a set of documents by the same author along with exactly one document of questioned authorship, and the task is to determine whether the known and the questioned documents are by the same author or not. In comparison to PAN-2013, a significantly larger corpus was built comprising hundreds of documents in four natural languages (Dutch, English, Greek, and Spanish) and four genres (essays, reviews, novels, opinion articles). In addition, more suitable performance measures are used focusing on the accuracy and the confidence of the predictions as well as the ability of the submitted methods to leave some problems unanswered in case there is great uncertainty. To this end, we adopt the c@1 measure, originally proposed for the question answering task. We received 13 software submissions that were evaluated in the TIRA framework. Analytical evaluation results are presented where one language-independent approach serves as a challenging baseline. Moreover, we continue the successful practice of the PAN labs to examine meta-models based on the combination of all submitted systems. Last but not least, we provide statistical significance tests to demonstrate the important differences between the submitted approaches.

1 Introduction

Authorship analysis has attracted much attention in recent years due to both the rapid increase of texts in electronic form and the need for intelligent systems able to handle this information. Authorship analysis deals with the personal style of authors and includes three major areas:

- *Author identification:* Given a set of candidate authors for whom some texts of undisputed authorship exist, attribute texts of unknown authorship to one of the candidates. This can be applied mainly to forensic applications and literary analysis [13, 31].

- *Author profiling:* The extraction of demographic information such as gender, age, etc. about the authors. This has significant applications mainly in market analysis [28].
- *Author clustering:* The segmentation of texts into stylistically homogeneous parts. This can be applied to distinguish different authors in collaborative writing, to detect plagiarism without a reference corpus (i.e., intrinsic plagiarism detection [35]), and to detect changes in the personal style of a certain author by examining their works chronologically [14].

Author identification is by far the most prevalent field of authorship analysis in terms of published studies. The authorship attribution problem can be viewed as a closed-set classification task where all possible candidate authors are known. This is suitable in many forensic applications where the investigators of a case can provide a specific set of suspects based on certain restrictions (e.g., access to specific material, knowledge of specific facts, etc.). A more general definition of the authorship attribution problem corresponds to an open-set classification task where the true author of the disputed texts is not necessarily included in the set of candidate authors. This setting is much more difficult in comparison to the closed-set attribution scenario, especially when the size of the candidate author set is small [18]. Finally, when the set of candidate authors is singleton, we get the author verification problem. This is an even more difficult attribution task.

The PAN-2014 evaluation lab continues the practice of PAN-2013 and focuses on the author verification problem [15]. First, this is a fundamental problem in authorship attribution [20] and by studying it we can extract more useful conclusions about the performance of certain attribution methods. Any author identification task can be decomposed into a series of author verification problems. Therefore, the ability of an approach to effectively deal with this task means that it can cope with every authorship attribution problem. Moreover, in comparison to PAN-2013, we provide a larger collection of verification problems including more natural languages and genres. Thus, we can study more reliably the performance of the submitted approaches under different conditions and test their ability to be adapted to certain properties of documents. In addition, we define more appropriate performance measures that are suitable for this cost-sensitive task focusing on the ability of the submitted approaches to assign confidence scores in their answers as well as their ability to leave the most uncertain cases unanswered.

Based on the successful practice of PAN-2013, we build a meta-classifier to combine all submitted approaches and examine the performance of this ensemble model in comparison to the individual participants [15]. Moreover, we use one effective model submitted to PAN-2013 as a baseline method. This enables us to have a more challenging baseline (in comparison to random guess) that reflects and can be adapted to the difficulty of a certain corpus. Finally, we provide tests of statistical significance to examine whether there are important differences in the performance of the submitted methods, the baseline, and the meta-classifier.

In the remainder of this paper, Section 2 reviews previous work in author verification, Section 3 analytically describes the evaluation setup used at PAN-2014 and Section 4 presents the evaluation results in detail. A review of the submitted

approaches is included in Section 5 and Section 6 summarizes the main conclusions that can be drawn and discusses future work directions.

2 Relevant Work

The author verification problem was first discussed in [32]. Based on a corpus of newspaper articles in Greek, they used multiple regression to produce a response function for a given author and a threshold value to determine whether or not a questioned document was by that author. False acceptance and false rejection rates were used to evaluate this model. The same metrics were used by [37] to evaluate an authorship verification method based on a rich set of linguistic features.

Perhaps the best-known approach for author verification, the *unmasking* method, was introduced in [19]. The main idea is to build a SVM classifier to distinguish the questioned document from the set of known documents, then to remove the most important features and repeat this process. In case the questioned and known documents are by the same author, the accuracy of the classifier significantly drops after a small number of repetitions while it remains relatively high when they are not by the same author. Accuracy and F_1 were used to evaluate this method that was very effective in long documents but fails when documents are relatively short [33]. Modifications and additional evaluation tests for the unmasking method can be found in [34] and [16].

Luyckx and Daelemans approximated the author verification problem as a binary classification task by considering all available texts by other authors as negative examples [22]. They used recall, precision, and F_1 to evaluate their approach in a corpus of student essays in Dutch. Escalante et al. applied particle swarm model selection to select a suitable classifier for a given author [5]. They used F_1 and balanced error rate (the average of error rates for positive and negative class) to evaluate their approach on two corpora of English newswire stories and Spanish poems. More recently, Koppel and Winter proposed an effective method that attempts to transform authorship verification from a one-class classification task to a multi-class classification problem by introducing additional authors, the so-called impostors, using documents found in external sources (e.g., the Web) [20]. Accuracy and recall-precision graphs were used to evaluate this method.

Author verification was included in previous editions of the PAN evaluation lab. The author identification task at PAN-2011 [1] included 3 author verification problems, each comprising a number of texts (i.e., email messages) of known authorship, all by the same author, and a number of questioned texts (either by the author of the known texts or not). Performance was measured by macro-average precision, recall and F_1. PAN-2013 was exclusively focused on the author verification problem [15]. New training and evaluation corpora were built on three languages (i.e., English, Greek, and Spanish) where each verification problem included at most 10 documents by the same author and exactly one questioned document. Beyond a binary answer for each verification problem, the participants could also produce (optionally) a probability-like score to indicate the confidence of a positive answer. Recall, precision, F_1 and ROC graphs were used to evaluate the performance of the 18

participants. Moreover, a simple meta-model combining all the submitted methods achieved the best overall performance. For the first time, software submissions were requested at PAN-2013 enabling reproducibility of the results and future evaluation on different corpora.

3 Evaluation Setup

PAN-2014 focuses on author verification, similar to PAN-2013. Given a set of known documents all written by the same author and exactly one questioned document, the task is to determine whether the questioned document was written by that particular author or not. Similar to the corresponding task at PAN-2013, best efforts were applied to ensure that all known and questioned documents within a problem are matched for genre, register, theme, and date of writing. In contrast to PAN-2013, the number of known documents is limited to at most 5, while a greater variety of languages and genres is covered. The text length of documents varies from a few hundred to a few thousand words, depending on the genre.

The participants were asked to submit their software and consider as input parameters the language and genre of the documents. For each verification problem, they should provide a score, a real number in [0,1], corresponding to the probability of a positive answer (i.e., the known and the questioned documents are by the same author). In case the participants wanted to leave some verification problems unanswered, they could assign a probability score of exactly 0.5 to those problems.

3.1 Corpus

The PAN-2014 corpus comprises author verification problems in four languages: Dutch, English, Greek, and Spanish. For Dutch and English there are two genres in separate parts of the corpus. An overview of the training and evaluation corpus of the author identification task is shown in Table 1. As can be seen, beyond language and genre there is variety of known texts per problem and text-length. The size of both training and evaluation corpora is significantly larger than the corresponding corpora of PAN-2013. All corpora in both training and evaluation sets are balanced with respect to the number of positive and negative examples.

The Dutch corpus is a transformed version of the CLiPS Stylometry Investigation (CSI) corpus [38]. This recently released corpus contains documents from two genres: essays and reviews, which are the two Dutch genres present in the corpus for this task. All documents were written by language students at the University of Antwerp between 2012 and 2014. All authors are native speakers of Dutch. The CSI corpus was developed for use in computational stylometry research (i.e. detection of age, gender, personality, region of origin, etc.), but has many other purposes as well (e.g., deception detection, sentiment analysis). We adapted the CSI corpus to match the needs of the authorship verification task and ended up with 200 problem sets for the review genre and 192 problem sets in the essay genre. All verification problems include 1-5 known texts. The training and evaluation set each contain half of the problem sets in each genre.

Table 1. Statistics of the training and evaluation corpora used in the author identification task at PAN-2014.

	Language	Genre	#Problems	#Docs	Avg. of known docs per problem	Avg. words per document
Training	Dutch	Essays	96	268	1.8	412.4
	Dutch	Reviews	100	202	1.0	112.3
	English	Essays	200	729	2.6	848.0
	English	Novels	100	200	1.0	3,137.8
	Greek	Articles	100	385	2.9	1,404.0
	Spanish	Articles	100	600	5.0	1,135.6
	Total		696	2,384	2.4	1,091.0
Evaluation	Dutch	Essays	96	287	2.0	398.1
	Dutch	Reviews	100	202	1.0	116.3
	English	Essays	200	718	2.6	833.2
	English	Novels	200	400	1.0	6,104.0
	Greek	Articles	100	368	2.7	1,536.6
	Spanish	Articles	100	600	5.0	1,121.4
	Total		796	2,575	2.2	1,714.9
TOTAL			1,492	4,959	2.3	1,415.0

The English essays corpus was derived from a previously existing corpus of English-as-second-language students. The Uppsala Student English (USE) corpus [2] was originally intended to become a tool for research on foreign languages learning. It consists of university-level full-time students' essays handed by electronic means. In this kind of texts stylistic awareness represents an important writing factor. The USE corpus includes clear borders between writings produced in the framework of three different terms: *a*, *b*, and *c*. Every essay is intended to be produced on personal, formal, or academic style. A total of 440 authors contributed with at least one essay to the corpus, resulting in 1,489 documents. The average size of an essay is 820 words. Typically, one student contributed with more than one essay, often surpassing the different terms. Taking advantage of the USE corpus meta-information, we defined two main constraints: every document in the collection, known or questioned, should contain at least 500 words and the number of known documents in a case must range between one and five. As a result of the first constraint, only 435 authors were considered. We also took advantage of the students' background information to set case-generation rules. Firstly, all the documents in a case must come from students from the same term (i.e., both were written within term *a*, *b*, or *c*). Secondly, we divided the students in age-based clusters. To form negative verification problems, based on the fact that the students' age ranged between 18 and 59 years, an author A was considered as candidate match for author A_q according to the following rules:

- If A_q is younger than 20 years old, A must be younger than 20 as well;

- If A_q is between 20 and 25 years old, A must be exactly the same age;
- If A_q is between 26 and 30 years old, A must be in the same age range; and
- If A_q is older than 30 years old, A must be older than 30 as well.

This combination of age- and term-related constraints allowed us to create cases where the authors come from similar backgrounds. During our generation process, the texts as in the USE corpus were slightly modified. Anonymization labels were substituted by a randomly chosen proper name in English. In order not to provide any hint about a case, the same name was used both in the questioned and known documents. One source USE document could be considered at most twice in the authorship verification corpus: once in a positive case and once in a negative case.

The English novels used in the PAN-2014 corpus represent an attempt to provide a narrower focus in terms of both content and writing style than many similar collections. Instead of simply focusing on a single genre or time period, they focus on a very small subgenre of speculative and horror fiction known generally as the "Cthulhu Mythos". This is specifically a shared-universe genre, based originally on the writings of the American H.P. Lovecraft (for this reason, the genre is also called "Lovecraftian horror"), a shared universe with a theme of human ineffectiveness in the face of a set of powerful named "cosmic horrors". It is also typically characterized by extremely florid prose and an unusual vocabulary. Perhaps most significantly, many of the elements of this genre are themselves unusual terms (e.g., unpronounceable proper names of these cosmic horrors such as "Cthulhu", "Nyarlathotep", "Lloigor", "Tsathoggua", or "Shub-Niggurath"), thus creating a strong shared element that is unusual in regular English prose. Similarly, the overall theme and tone of these stories is strongly negative (many of them, for example, take the form of classical tragedies and end with the death of the protagonist). For this reason, we feel that this testbed provides a number of unusual elements that may be appropriately explored as an example of a tightly controlled genre. The corpus covers an extended length of time, from Lovecraft's original work to modern fan-fiction. Documents were gathered from a variety of on-line sources including Project Gutenberg[1] and FanFiction[2], and edited for uniformity of format; in some cases lengthy works were broken down into subsections based on internal divisions such as chapters or sections.

The Greek corpus comprises newspaper opinion articles published in the Greek weekly newspaper TO BHMA[3] from 1996 to 2012. Note that the training corpus in Greek was formed based on the respective training and evaluation corpora of PAN-2013. The length of each article is at least 1,000 words while the number of known texts per problem varies between 1 to 5. In each verification problem, we included texts that had strong thematic similarities indicated by the occurrence of certain keywords. In contrast to PAN-2013, there was no stylistic analysis of the texts to indicate authors with very similar styles or texts of the same author with notable differences.

[1] http://www.gutenberg.org/
[2] https://www.fanfiction.net/
[3] http://www.tovima.gr

The Spanish corpus refers to the same genre as the Greek corpus. Newspaper opinion articles of the Spanish newspaper El Pais[4] were considered and author verification problems were formed taking into account thematic similarities between articles as indicated by certain keywords used to index the articles in the website of this newspaper. All verification problems for this corpus include exactly five known texts, while the average text length is relatively large, exceeding 1,000 words.

3.2 Performance measures

The probability scores provided by the participants are used to build ROC curves and the area under the curve (AUC) is used as a scalar evaluation measure. This is a well-known evaluation technique for binary classifiers [6]. In addition, the performance measures used in this task should be able to take unanswered problems into account. Similarly to other tasks, like question answering, it is preferred to leave the problem unanswered rather than responding incorrectly when there is great uncertainty. The measures of recall and precision used at PAN-2013 were not able to reward submissions that left problems unanswered while maintaining high accuracy in given answers.

In the current evaluation setup we adopted the c@1 measure, originally proposed for question answering tasks, which explicitly extends accuracy based on the number of problems left unanswered [27]. More specifically, to use this measure we first transform probability scores to binary answers. Every score greater than 0.5 is considered as a positive answer (i.e., the known and questioned documents are by the same author), every score lower than 0.5 is considered as a negative answer (i.e., the known and questioned documents are by different authors) while all scores equal to 0.5 correspond to unanswered problems. Then, c@1 is defined as follows:

$$c@1 = \frac{1}{n}\left(n_c + \frac{n_c}{n} n_u\right)$$

where n is the number of problems, n_c is the number of correct answers, and n_u is the number of problems left unanswered. If a participant would provide an answer different from 0.5 for all problems, then c@1 will be equal to accuracy. If all problems are left unanswered, then c@1 will be zero. If only some problems are left unanswered, this measure will be increased as if these problems were answered with the same accuracy as the rest of the problems. Therefore, this measure rewards participants that maintain a high number of correct answers, for which there is great confidence, and decrease the number of incorrect answers, for uncertain cases, by leaving them unanswered.

To provide a final rank of participants, AUC and c@1 are combined in the final score which is merely the product of these two measures. In addition, the efficiency of the submitted methods is measured in terms of elapsed runtime.

[4] http://elpais.com

3.4 Baseline

The author verification task has a random guess baseline of 0.5 for both AUC and c@1. However, this baseline is not challenging. What we need is a baseline that corresponds to a standard method so that we know what submissions are really better than the state of the art. Moreover, since the evaluation corpus comprises several languages and genres, we need a baseline that can reflect and adapt to the difficulty of a specific corpus.

Based on the submissions of the author identification task at PAN-2013, it is possible to use state-of-the-art methods (in particular, the PAN-2013 winners) and apply them to PAN-2014 corpus. However, since the PAN-2014 task comprises more languages, we need a language-independent approach. In addition, we need a method that can provide both binary answers and probability scores (the latter was optional at PAN-2013). Based on these requirements, we selected the approach of [11] to serve as baseline. More specifically, this approach has the following characteristics:

- It is language-independent.
- It can provide both binary answers and real scores.
- The real scores are already calibrated to probability-like scores for a positive answer (i.e., all scores greater than 0.5 correspond to a positive answer).
- It was the winner of PAN-2013 in terms of overall AUC scores.

It should be noted that this baseline method has not been specifically trained on the corpora of PAN-2014, so its performance is not optimized. It can only be viewed as a general method that can be applied to any corpus. Moreover, this approach does not leave problems unanswered, so it cannot take advantage of the new performance measures.

3.5 Meta-classifier

Following the practice of PAN-2013, we examine the performance of a meta-model that combines all answers given by the participants for each problem. We define a straight-forward meta-classifier that calculates the average of the probability scores provided by the participants for each problem. It can be seen as a heterogeneous ensemble model that combines base classifiers corresponding to different approaches. Note that the average of all the provided answers is not likely to be exactly 0.5; hence, this meta-model very rarely leaves problems unanswered. This meta-model can be naturally extended by allowing all answers with a score between 0.5-a and 0.5+a to become equal to 0.5. However, since the parameter a should be tuned to an arbitrary predefined value or be optimized for each language/genre, we decided not to perform such an extension.

Table 2. Overall evaluation results of the author identification task at PAN-2014.

Rank		FinalScore	AUC	c@1	Runtime	Unansw. Problems
	META-CLASSIFIER	0.566	0.798	0.710		0
1	Khonji & Iraqi	**0.490**	**0.718**	0.683	20:59:40	2
2	Frery et al.	0.484	0.707	**0.684**	00:06:42	28
3	Castillo et al.	0.461	0.682	0.676	03:59:04	78
4	Moreau et al.	0.451	0.703	0.641	01:07:34	50
5	Mayor et al.	0.450	0.690	0.651	05:26:17	29
6	Zamani et al.	0.426	0.682	0.624	02:37:25	0
7	Satyam et al.	0.400	0.631	0.634	02:52:37	7
8	Modaresi & Gross	0.375	0.610	0.614	**00:00:38**	0
9	Jankowska et al.	0.367	0.609	0.602	07:38:18	7
10	Halvani & Steinebach	0.335	0.595	0.564	00:00:54	3
	BASELINE	0.325	0.587	0.554	00:21:10	0
11	Vartapetiance & Gillam	0.308	0.555	0.555	01:07:39	0
12	Layton	0.306	0.548	0.559	27:00:01	0
13	Harvey	0.304	0.558	0.544	01:06:19	100

4 Evaluation Results

We received 13 submissions from research teams in Australia, Canada (2), France, Germany (2), India, Iran, Ireland, Mexico (2), United Arab Emirates, and United Kingdom. The participants submitted and evaluated their author verification software within the TIRA framework [8]. A separate run for each corpus corresponding to each language and genre was performed.

The overall results of the task concerning the performance of the submitted approaches in the whole evaluation corpus are shown in Table 2. These evaluation scores are the result of micro-averaging over the set of 796 verification problems. Put in other words, each verification problem has the same weight in this analysis, so the language and genre information are not taken into account. As can be seen, the overall winner method of Khonji and Iraqi [17] achieved the best results in terms of AUC and was also very effective in terms of c@1. On the other hand, it was one of the less efficient methods requiring about 21 hours for processing the whole evaluation corpus. The second best submission by Frery et al. [7] was much more efficient and achieved the best c@1 score. In general, most of the submitted methods outperformed the baseline. It has to be emphasized that the best five participants were able to leave some problems unanswered. In total 4 out of the 13 participants answered all problems. Moreover, one participant provided binary answers instead of probability scores [36] and one participant did not process the Greek corpus [10]. With respect to the meta-classifier, which is averaging the answers of all 13 participants, its performance is significantly better than each individual system, achieving a final score greater than 0.5.

Table 3. Evaluation results on the evaluation corpus of Dutch essays.

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
META-CLASSIFIER	0.867	0.957	0.906		0
Mayor et al.	**0.823**	**0.932**	0.883	00:15:05	2
Frery et al.	0.821	0.906	**0.906**	00:00:30	0
Khonji & Iraqi	0.770	0.913	0.844	00:58:21	0
Moreau et al.	0.755	0.907	0.832	00:02:09	34
Castillo et al.	0.741	0.861	0.861	00:01:57	2
Jankowska et al.	0.732	0.869	0.842	00:23:26	1
BASELINE	0.685	0.865	0.792	00:00:52	0
Zamani et al.	0.525	0.741	0.708	00:00:27	0
Vartapetiance & Gillam	0.517	0.719	0.719	00:06:37	0
Satyam et al.	0.489	0.651	0.750	00:01:21	0
Halvani & Steinebach	0.399	0.647	0.617	00:00:06	2
Harvey	0.396	0.644	0.615	00:02:19	0
Modaresi & Gross	0.378	0.595	0.635	**00:00:05**	0
Layton	0.307	0.546	0.563	00:55:07	0

Table 4. Evaluation results on the evaluation corpus of Dutch reviews.

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
Satyam et al.	**0.525**	**0.757**	**0.694**	00:00:16	2
Khonji & Iraqi	0.479	0.736	0.650	00:12:24	0
META-CLASSIFIER	0.428	0.737	0.580		0
Moreau et al.	0.375	0.635	0.590	00:01:25	0
Zamani et al.	0.362	0.613	0.590	00:00:11	0
Jankowska et al.	0.357	0.638	0.560	00:06:24	0
Frery et al.	0.347	0.601	0.578	00:00:09	5
BASELINE	0.322	0.607	0.530	00:00:12	0
Halvani & Steinebach	0.316	0.575	0.550	**00:00:03**	0
Mayor et al.	0.299	0.569	0.525	00:07:01	1
Layton	0.261	0.503	0.520	00:56:17	0
Vartapetiance & Gillam	0.260	0.510	0.510	00:05:43	0
Castillo et al.	0.247	0.669	0.370	00:01:01	76
Modaresi & Gross	0.247	0.494	0.500	00:00:07	0
Harvey	0.170	0.354	0.480	00:01:45	0

Table 5. Evaluation results on the evaluation corpus of English essays.

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
META-CLASSIFIER	0.531	0.781	0.680		0
Frery et al.	**0.513**	**0.723**	**0.710**	00:00:54	15
Satyam et al.	0.459	0.699	0.657	00:16:23	2
Moreau et al.	0.372	0.620	0.600	00:28:15	0
Layton	0.363	0.595	0.610	07:42:45	0
Modaresi & Gross	0.350	0.603	0.580	**00:00:07**	0
Khonji & Iraqi	0.349	0.599	0.583	09:10:01	1
Halvani & Steinebach	0.338	0.629	0.538	**00:00:07**	1
Zamani et al.	0.322	0.585	0.550	00:02:03	0
Mayor et al.	0.318	0.572	0.557	01:01:07	10
Castillo et al.	0.318	0.549	0.580	01:31:53	0
Harvey	0.312	0.579	0.540	00:10:22	0
BASELINE	0.288	0.543	0.530	00:03:29	0
Jankowska et al.	0.284	0.518	0.548	01:16:35	5
Vartapetiance & Gillam	0.270	0.520	0.520	00:16:44	0

Table 6. Evaluation results on the evaluation corpus of English novels.

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
Modaresi & Gross	**0.508**	0.711	**0.715**	**00:00:07**	0
Zamani et al.	0.476	0.733	0.650	02:02:02	0
META-CLASSIFIER	0.472	0.732	0.645		0
Khonji & Iraqi	0.458	**0.750**	0.610	02:06:16	0
Mayor et al.	0.407	0.664	0.614	01:59:47	8
Castillo et al.	0.386	0.628	0.615	02:14:11	0
Satyam et al.	0.380	0.657	0.579	02:14:28	3
Frery et al.	0.360	0.612	0.588	00:03:11	1
Moreau et al.	0.313	0.597	0.525	00:11:04	12
Halvani & Steinebach	0.293	0.569	0.515	**00:00:07**	0
Harvey	0.283	0.540	0.525	00:46:30	0
Layton	0.260	0.510	0.510	07:27:58	0
Vartapetiance & Gillam	0.245	0.495	0.495	00:13:03	0
Jankowska et al.	0.225	0.491	0.457	02:36:12	1
BASELINE	0.202	0.453	0.445	00:08:31	0

Table 7. Evaluation results on the evaluation corpus of Greek articles.

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
Khonji & Iraqi	**0.720**	**0.889**	**0.810**	03:41:48	0
META-CLASSIFIER	0.635	0.836	0.760		0
Mayor et al.	0.621	0.826	0.752	00:51:03	3
Moreau et al.	0.565	0.800	0.707	00:05:54	4
Castillo et al.	0.501	0.686	0.730	00:03:14	0
Jankowska et al.	0.497	0.731	0.680	01:36:00	0
Zamani et al.	0.470	0.712	0.660	00:15:12	0
BASELINE	0.452	0.706	0.640	00:03:38	0
Frery et al.	0.436	0.679	0.642	00:00:58	7
Layton	0.403	0.661	0.610	04:40:29	0
Halvani & Steinebach	0.367	0.611	0.600	**00:00:04**	0
Satyam et al.	0.356	0.593	0.600	00:12:01	0
Modaresi & Gross	0.294	0.544	0.540	00:00:05	0
Vartapetiance & Gillam	0.281	0.530	0.530	00:10:17	0
Harvey	0.000	0.500	0.000		100

Table 8. Evaluation results on the evaluation corpus of Spanish articles.

	FinalScore	AUC	c@1	Runtime	Unansw. Problems
META-CLASSIFIER	0.709	0.898	0.790		0
Khonji & Iraqi	**0.698**	**0.898**	**0.778**	04:50:49	1
Moreau et al.	0.634	0.845	0.750	00:18:47	0
Jankowska et al.	0.586	0.803	0.730	01:39:41	0
Frery et al.	0.581	0.774	0.750	00:01:01	0
Castillo et al.	0.558	0.734	0.760	00:06:48	0
Mayor et al.	0.539	0.755	0.714	01:12:14	5
Harvey	0.514	0.790	0.650	00:05:23	0
Zamani et al.	0.468	0.731	0.640	00:17:30	0
Vartapetiance & Gillam	0.436	0.660	0.660	00:15:15	0
Halvani & Steinebach	0.423	0.661	0.640	00:00:27	0
Modaresi & Gross	0.416	0.640	0.650	**00:00:08**	0
BASELINE	0.378	0.713	0.530	00:04:27	0
Layton	0.299	0.553	0.540	05:17:25	0
Satyam et al.	0.248	0.443	0.560	00:08:09	0

Tables 3-8 present the evaluation results on each of the six corpora separately. In all tables, the best performing submission (excluding the meta-classifier and the baseline method) is in boldface. In terms of average performance of all submitted approaches, the corpus of Dutch essays seems to be the easiest while the corpus of Dutch reviews to be the hardest one. The latter can be partially explained by the fact that the corpus provides only one known document per problem and that it contains only short texts. Moreover, the availability of multiple relatively long known documents seems to assist the submitted systems to achieve a better average performance on the Greek and Spanish corpora compared to the English corpora of essays and novels. There is a different winner for each corpus with the exception of [17] who won on both Greek and Spanish corpora. This might indicate a better tuning of their approach for newspaper opinion articles rather than essays, reviews or novels. However, the performance of this submission on all corpora is notable since it is usually included in the first 3-best performing methods with the exception of the English essays where it is ranked 6th (excluding the meta-classifier).

The performance of the baseline method varies. In the English and Spanish corpora it is relatively low. In the Dutch and Greek corpora it is very challenging, outperforming almost half of the participants. In addition, the meta-classifier is very effective on all corpora. However, it is outperformed by some individual participants on three corpora. Another interesting remark is that the problems left unanswered by most participants are not evenly distributed across the corpora. The majority of the problems left unanswered by Castillo et al. [4] refer to Dutch reviews (possibly reflecting the difficulty of this corpus). Similarly, Moreau et al. [25] did not answer many problems of Dutch essays while most of the unanswered problems of Frery et al. [7] belong to English essays and Greek articles. On the other hand, Mayor et al. [23] left at least one problem unanswered in each corpus.

The ROC curves of the best performing participants on the whole evaluation corpus are shown in Figure 1. More specifically, the convex hull of all submitted approaches together with the participants' curves who are part of the convex hull are shown. The overall winning approach of Khonji and Iraqi [17] and the second-best method of Frery et al. [7] dominate the convex hull in case the false positive and false negative errors have the same cost [6]. In low values of FPR in the ROC space, where the cost of false positives is considered higher than the cost of false negatives, the approach of Modaresi and Gross [23] is the best. On the other hand, if the false negatives have larger cost than the false positives, in large values of FPR in the ROC space, the approach of Moreau et al. [25] is the most effective. Note also that the submission by Castillo et al. [4], ranked in the 3rd position in the overall results (see Table 2), is not part of the convex hull meaning that this approach is always outperformed by another approach no matter the cost of the false positives and false negatives.

In addition, Figure 1 depicts the ROC curves of the baseline method and the meta-classifier. The baseline is clearly less effective than the best participants. It outperforms only Frery et al. [7] in very low values of FPR. On the other hand, the meta-classifier clearly outperforms the convex hull of all the submitted methods in the whole range of the curve. This means that the meta-classifier is more effective than any individual submission for any given cost of false positives and false negatives.

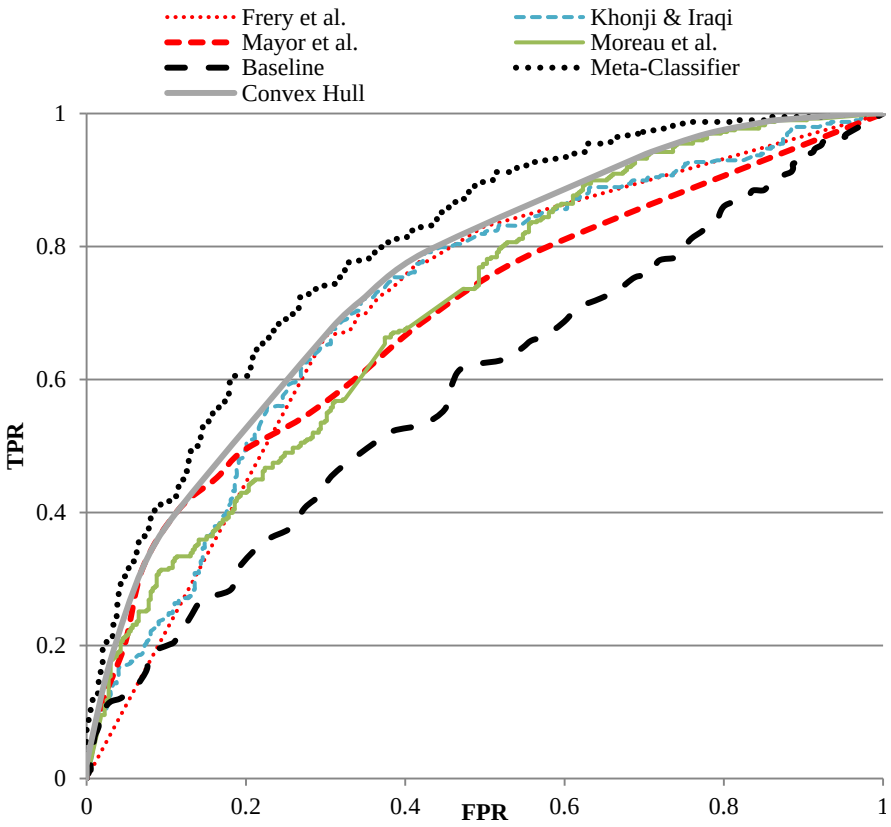

Fig. 1. ROC graphs of the best performing submissions and their convex hull, the baseline method, and the meta-classifier.

We computed statistical significance of performance differences between systems using approximate randomization testing [26][5]. As noted by [39] among others, for comparing outputs from classifiers, frequently used statistical significance tests such as paired t-tests make assumptions that do not hold for precision scores and F-scores. Approximate randomisation testing does not make these assumptions and can handle complicated distributions. We did a pairwise comparison of accuracy of all systems based on this method and the results are shown in Table 9. The null hypothesis is that there is no difference in the output of two systems. When the probability of accepting the null hypothesis is $p < 0.05$ we consider the systems to be significantly different from each other. When $p < 0.001$ the difference is highly significant, when $0.001 < p < 0.01$ the difference is very significant, and when $0.01 < p < 0.05$ the difference is significant.

[5] We used the implementation by Vincent Van Asch available from the CLiPS website http://www.clips.uantwerpen.be/scripts/art

Table 9. Pairwise significance tests for the entire evaluation corpus. Significant differences are marked with asterisks, *** corresponds to highly significant difference ($p < 0.001$), ** corresponds to very significant difference ($0.001 < p < 0.01$), * corresponds to significant difference ($0.01 < p < 0.05$), while = means the difference is not significant ($p > 0.05$).

	Khonji & Iraqi	Frery et al.	Castillo et al.	Moreau et al.	Mayor et al.	Zamani et al.	Satyam et al.	Modaresi & Gross	Jankowska et al.	Halvani & Steinebach	BASELINE	Vartapetiance & Gillam	Layton	Harvey
META-CLASSIFIER	=	*	***	***	***	***	***	***	***	***	***	***	***	***
Khonji & Iraqi		=	**	***	**	**	**	**	**	***	***	***	***	***
Frery et al.			=	*	=	=	=	=	*	***	***	***	**	***
Castillo et al.				=	=	=	=	=	=	*	**	**	*	***
Moreau et al.					=	=	=	=	=	=	*	*	=	***
Mayor et al.						=	=	=	=	**	**	**	**	***
Zamani et al.							=	=	=	**	**	**	**	***
Satyam et al.								=	=	**	**	***	**	***
Modaresi & Gross									=	*	**	*	*	***
Jankowska et al.										=	**	=	=	***
Halvani & Steinebach											=	=	=	***
BASELINE												=	=	*
Vartapetiance & Gillam													=	**
Layton														**

Based on this analysis, it is easy to see that there are no significant differences in systems of neighboring rank. The winner submission of [17] is either very significantly or highly significantly better than the rest of the approaches (with the exception of the second winner [7]). In addition, the meta-classifier is highly significantly better than all the participants except for the first two winners.

5 Survey of Submissions

Among 13 participant approaches, 7 were submitted by teams that had participated also in the PAN-2013 competition. Some of them attempted to improve the method proposed in 2013 [9, 12, 21, 36] and others presented new models [4, 23, 25].

All the submitted approaches can be described according to some basic properties. First, an author verification method is either *intrinsic* or *extrinsic*. For each verification problem, intrinsic methods use only the known texts and the unknown text of that problem to make some analysis and decide whether they are by the same author or not. They don't make use of any other texts by other authors. The majority of submitted approaches falls into this category [4, 7, 9, 10, 12, 21, 24, 25, 29, 36]. On the other hand, extrinsic methods attempt to transform author verification from a one-class classification task (where the known texts are the positive examples and there are no negative examples) to a binary classification task (where documents by other authors play the role of the negative examples). To this end, for each verification problem, extrinsic methods need additional documents by other authors found in external resources. The approaches of [17, 23, 40] belong to this category. The winner submission of PAN-2014 by [17] is a modification of the *Impostors* method [20], similarly to PAN-2013 [30], where a corpus of external documents for each language/genre was used.

Another important characteristic of a verification method is its type of learning. There are *lazy* approaches where the training phase is nearly omitted and all necessary processing is performed at the time they have to decide about a new verification problem. Most of the submitted approaches follow this idea [4, 9, 10, 12, 17, 21, 23, 29, 36, 40]. On the other hand, *eager* methods attempt to build a general model based on the training corpus. For example, [7] builds a decision tree for each corpus, [25] apply a genetic algorithm to find the characteristics of the verification model for each corpus, and [24] use fuzzy C-means clustering to extract a general description of each corpus. Since eager methods perform most of the necessary calculations in the training phase, they are generally more efficient in terms of runtime.

With respect to the features used for text representation, the majority of the participant methods focused on low-level measures. More specifically most of the proposed features are either character measures (i.e., punctuation mark counts, prefix/suffix counts, character n-grams, etc.) or lexical measures (i.e., vocabulary richness measures, sentence/word length counts, stopword frequency, n-grams of words/stopwords, word skip-grams, etc.). There were a few attempts to incorporate syntactic features, namely POS tag counts [17, 25, 40], while one approach was exclusively based on that type of information [10].

6 Discussion

The author identification task at PAN-2014 focused on the author verification problem. The task definition was practically the same as in PAN-2013. However, this year we substantially enlarged both training and evaluation corpora and enriched them to include several languages and genres. In that way, we enabled participants to study

how they can adapt and fine-tune their approaches according to a given language and genre. Another important novelty was the use of different performance measures that put emphasis on both the appropriate ranking of the provided answers in terms of confidence (AUC) as well as the ability of the submitted systems to leave some problems unanswered when there is great uncertainty (c@1). We believe that this combination of performance measures is more appropriate for author verification, a cost-sensitive task.

Similar to PAN-2013, the overall winner was a modification of the *Impostors* method [17]. The performance of this approach was notably stable in all six different corpora despite the fact that it did not leave many problems unanswered. This demonstrates the great potential of extrinsic verification methods. In addition, the significantly larger training corpus allowed participants to explore, for the first time, the use of eager learning methods in the author verification task. Such an approach may be both effective and efficient as it is demonstrated by the overall performance and runtime of the second overall winner [7].

We received 13 software submissions, a reduced figure in comparison to 18 submissions at PAN-2013, possibly due to the greater difficulty of the task. Moreover, this year the evaluation of the submitted systems was performed by participants themselves using the TIRA framework [8]. Seven participants from PAN-2013 submitted their approaches again this year. It is remarkable that those teams that slightly modified their existing approach did not achieve a high performance [9, 12, 21, 36]. On the other hand, the teams that radically changed their approach, including the ability to leave some problems unanswered, achieved very good results [4, 23, 25].

Based on the software submissions at PAN-2013, we were able to define a challenging baseline method that is better than random guessing and can reflect the difficulty of the examined corpus. In many cases, the baseline method was ranked in the middle of the participants list, clearly showing the approaches with notable performance. Given the enhanced set of methods for author verification, collected at PAN-2013 and PAN-2014, we think that it will be possible to further improve the quality of the baseline methods in future competitions. Moreover, following the successful practice of PAN-2013, we examined the performance of a meta-model that combines all submitted systems in a heterogeneous ensemble. This meta-classifier was better than each individual submitted method while its ROC curve clearly outperformed the convex hull of all submitted approaches. This demonstrates the great potential of heterogeneous models in author verification, a practically unexplored area.

For the first time, we applied statistical significance tests on the results of the submitted methods to highlight the real differences between them. According to these tests, there is no significant difference between systems ranked in neighboring positions. However, there are highly significant differences between the winner approach and the rest of the submissions (with the exception of the second winner). We believe that such significance tests are absolutely necessary to extract reliable conclusions and we are going to adopt them in future evaluation labs.

One of our ambitions in this task was to involve experts from forensic linguistics so that they can manually (or semi-automatically) analyze the same corpora and submit their answers. This could serve as another very interesting baseline approach

that would enable the comparison of fully-automated systems with traditional human expert methods. Unfortunately, this attempt was not successful. So far, we were not able to find experts in forensic linguistics willing to participate or to devote the necessary time to solve a large amount of author verification problems under certain time constraints. We are still working on this direction.

We believe that the focus of PAN-2013 and PAN-2014 on the author verification task has produced a significant progress in this field concerning the development of new corpora and new methods as well as in defining an appropriate evaluation framework. Clearly, author verification is far from being a solved task and there are many variations that can be explored in future evaluation labs including cross-topic and cross-genre verification (i.e., where the known and the questioned documents do not match in terms of topic/genre) and very short text verification (i.e., where the documents are tweets or SMS messages).

Acknowledgement

This work was partially supported by the WIQ-EI IRSES project (Grant No. 269180) within the FP7 Marie Curie action and by grant OCI-1032683 from the United States National Science Foundation. The work of the last author is funded by the Spanish Ministry of Education and Science (TACARDI project, TIN2012-38523-C02-00).

References

1. S. Argamon and P. Juola. Overview of the International Authorship Identification Competition at PAN-2011. In V. Petras, P. Forner, P.D. Clough (eds.) *CLEF Notebook Papers/Labs/Workshop*, 2011.
2. M. W. Axelsson. USE--The Uppsala Student English Corpus: An instrument for needs analysis, *ICAME Journal*, 24:155-157, 2000.
3. L. Cappellato, N. Ferro, M. Halvey, and W. Kraaij (eds.). *CLEF 2014 Labs and Workshops, Notebook Papers*. CEUR Workshop Proceedings (CEUR-WS.org), ISSN 1613-0073, 2014.
4. E. Castillo, O. Cervantes, D.Vilariño, D. Pinto, and S. León. Unsupervised Method for the Authorship Identification Task – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
5. H.J. Escalante, M. Montes-y-Gómez and L. Villaseñor-Pineda. Particle Swarm Model Selection for Authorship Verification. In *Proceedings of the 14th Iberoamerican Conference on Pattern Recognition*, pages 563-570, 2009.
6. T. Fawcett. An Introduction to ROC Analysis. *Pattern Recognition Letters*, 27(8):861-874, 2006.
7. J. Fréry, C. Largeron, and M. Juganaru-Mathieu. UJM at CLEF in Author Identification – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
8. T. Gollub, M. Potthast, A. Beyer, M. Busse, F. Rangel, P. Rosso, E. Stamatatos, and B. Stein. Recent Trends in Digital Text Forensics and its Evaluation. In P. Forner, H. Müller, R. Paredes, P. Rosso, and B. Stein (eds), *Information Access*

Evaluation meets Multilinguality, Multimodality, and Visualization. 4th International Conference of the CLEF Initiative, 2013.
9. O. Halvani and M. Steinebach. VEBAV - A Simple, Scalable and Fast Authorship Verification Scheme – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
10. S. Harvey. Author Verification Using PPM with Parts of Speech Tagging – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
11. M. Jankowska, V. Kešelj, and E. Milios. Proximity based One-class Classification with Common N-Gram Dissimilarity for Authorship Verification Task – Notebook for PAN at CLEF 2013.In P. Forner, R. Navigli, and D. Tufis (eds). *CLEF 2013 Evaluation Labs and Workshop –Working Notes Papers*, 2013.
12. M. Jankowska, V. Kešelj, and E. Milios. Ensembles of Proximity-Based One-Class Classifiers for Author Verification – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
13. P. Juola. Authorship Attribution. *Foundations and Trends in IR*, 1:234–334, 2008.
14. P. Juola. An Overview of the Traditional Authorship Attribution Subtask. In *Proc. of CLEF'12*, 2012.
15. P. Juola and E. Stamatatos. Overview of the Author Identification Task at PAN-2013. In P. Forner, R. Navigli, and D. Tufis (eds). *CLEF 2013 Evaluation Labs and Workshop –Working Notes Papers*, 2013.
16. M. Kestemont, K. Luyckx, W. Daelemans, and T. Crombez. Cross-Genre Authorship Verification Using Unmasking. *English Studies*, 93(3):340-356, 2012.
17. M. Khonji and Y. Iraqi. A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF) – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
18. M. Koppel, J. Schler, and S. Argamon. Authorship Attribution in the Wild. *Language Resources and Evaluation*, 45:83–94, 2011.
19. M. Koppel, J. Schler, and E. Bonchek-Dokow. Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research*, 8:1261–1276, 2007.
20. M. Koppel and Y. Winter. Determining if Two Documents are by the Same Author. *Journal of the American Society for Information Science and Technology*, 65(1):178-187, 2014.
21. R. Layton. A Simple Local n-gram Ensemble for Authorship Verification – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
22. K. Luyckx and W. Daelemans. Authorship Attribution and Verification with Many Authors and Limited Data. In *Proceedings of the Twenty-Second International Conference on Computational Linguistics (COLING)*, pages 513-520, 2008.
23. C. Mayor, J. Gutierrez, A. Toledo, R. Martinez, P. Ledesma, G. Fuentes, and I. Meza. A Single Author Style Representation for the Author Verification Task – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].

24. P. Modaresi and P. Gross. A Language Independent Author Verifier Using Fuzzy C-Means Clustering – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
25. E. Moreau, A. Jayapal, and C. Vogel. Author Verification: Exploring a Large set of Parameters using a Genetic Algorithm – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
26. E. W. Noreen. *Computer Intensive Methods for Testing Hypotheses: An Introduction*. Wiley, 1989.
27. A. Peñas and A. Rodrigo. A Simple Measure to Assess Nonresponse. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pages 1415-1424, 2011.
28. F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches. Overview of the Author Profiling Task at PAN 2013. In P. Forner, R. Navigli, and D. Tufis (eds.), *Working Notes Papers of the CLEF 2013 Evaluation Labs*, 2013.
29. Satyam, Anand, A. K. Dawn, and S. K. Saha. A Statistical Analysis Approach to Author Identification Using Latent Semantic Analysis – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
30. S. Seidman. Authorship Verification Using the Impostors Method – Notebook for PAN at CLEF 2013. In P. Forner, R. Navigli, and D. Tufis (eds). *CLEF 2013 Evaluation Labs and Workshop –Working Notes Papers*, 2013.
31. E. Stamatatos. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, 60:538–556, 2009.
32. E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics*, 26(4):471-495, 2000.
33. C. Sanderson and S.Guenter. Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 482–491, 2006.
34. B. Stein, N. Lipka and S. Meyer zu Eissen. Meta Analysis within Authorship Verification. In *Proceedings of the 19th International Conference on Database and Expert Systems Applications*, pages 34-39, 2008.
35. B. Stein, N. Lipka, and P. Prettenhofer. Intrinsic Plagiarism Analysis. *Language Resources and Evaluation*, 45, pages 63-82, 2011.
36. A.Vartapetiance and L.Gillam. A Trinity of Trials: Surrey's 2014 Attempts at Author Verification – Notebook for PAN at CLEF 2014. In Cappellato, et al. [3].
37. H. van Halteren. Linguistic Profiling for Author Recognition and Verification. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics,* 2004.
38. B. Verhoeven and W. Daelemans. CLiPS Stylometry Investigation (CSI) Corpus: A Dutch Corpus for the Detection of Age, Gender, Personality, Sentiment and Deception in Text. In *Proc. of the 9th Int. Conf. on Language Resources and Evaluation* (LREC), 2014.
39. A. Yeh. More Accurate Tests for the Statistical Significance of Result Differences. In *Proceedings of the 18th Conference on Computational Linguistics*, Volume 2, pages 947-953, 2000.

40. H. Zamani, H. Nasr, P. Babaie, S. Abnar, M. Dehghani, and A. Shakery. Authorship Identification Using Dynamic Selection of Features from Probabilistic Feature Set. In *Proc. of the 5th International Conference of the CLEF Initiative*, 2014.