Overview of the PAN/CLEF 2015 Evaluation Lab

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

[1]Dept. of Information & Communication Systems Eng., University of the Aegean, Greece
[2]Web Technology & Information Systems, Bauhaus-Universität Weimar, Germany
[3]Autoritas Consulting, S.A., Spain
[4]Natural Language Engineering Lab, Universitat Politècnica de València, Spain

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

Abstract This paper presents an overview of the PAN/CLEF evaluation lab. During the last decade, PAN has been established as the main forum of text mining research focusing on the identification of personal traits of authors left behind in texts unintentionally. PAN 2015 comprises three tasks: plagiarism detection, author identification and author profiling studying important variations of these problems. In plagiarism detection, community-driven corpus construction is introduced as a new way of developing evaluation resources with diversity. In author identification, cross-topic and cross-genre author verification (where the texts of known and unknown authorship do not match in topic and/or genre) is introduced. A new corpus was built for this challenging, yet realistic, task covering four languages. In author profiling, in addition to usual author demographics, such as gender and age, five personality traits are introduced (openness, conscientiousness, extraversion, agreeableness, and neuroticism) and a new corpus of Twitter messages covering four languages was developed. In total, 53 teams participated in all three tasks of PAN 2015 and, following the practice of previous editions, software submissions were required and evaluated within the TIRA experimentation framework.

1 Introduction

Nowadays, huge volumes of electronic texts are produced daily and the need to automatically handle this information significantly increases. Topic, genre, and sentiment can be used to assign texts into predefined categories by exploiting their word usage, form and structure. Beyond such general characteristics, personal traits of authors left behind in texts unintentionally can also be used to extract useful information from texts.

Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN), a series of evaluation labs, focuses on that direction. During the last decade, PAN has been established as the main forum of innovative research in textual plagiarism detection and authorship analysis by producing large volumes of challenging corpora and introducing novel evaluation frameworks. PAN/CLEF 2015 edition comprises 3 tasks:

- *Plagiarism detection*: Given a document, identify all plagiarized sources and boundaries of re-used passages.

- *Author identification*: Given a document, identify its author.
- *Author profiling*: Given a document, extract information about the author (e.g. gender, age).

The last editions of PAN also focused on the same tasks [13,44]. However, every year important novelties are introduced. In more detail, in plagiarism detection community-driven corpus construction is introduced as a new way of developing evaluation resources characterized by diversity. This helps to drive the plagiarism detection task toward a truly community-driven evaluation.

The author identification task focuses on the authorship verification problem. Given a set of documents all by the same author and another questioned document, the task is to determine whether the author of the known documents is also the author of the questioned document. In contrast to most previous work in this area (including PAN-2013 and PAN-2014 editions), it is not assumed that all documents within a problem belong to the same genre/topic [21,64]. New corpora in several languages are built to enable the evaluation of submitted methods in challenging, yet realistic, cross-genre and cross-topic conditions.

The author profiling task at PAN-2015 enriches the author's demographics that are extracted from texts. In addition to gender and age (similar to PAN-2013 and PAN-2014 editions), personality traits are introduced. More specifically, the Big Five personality traits of Twitter users are examined (openness, conscientiousness, extraversion, agreeableness, and neuroticism). New corpora are produced for this task covering several European languages.

In total, 53 submissions were received for the three tasks (13, 18, and 22, respectively). Following the successful practice of PAN-2014, all participants were requested to submit their software to be evaluated within the TIRA experimentation platform [14] where participants are able to remotely run their software and evaluate its output [44]. The role of task organizers is then reduced to review evaluation results and assist participants to solve execution errors. TIRA ensures credibility and reproducibility of the reported results and supports continuous experimentation of the submitted methods using new corpora.

The remainder of this paper is organized as follows. Sections 2, 3, and 4, comprise relevant work, the evaluation setup, and results of plagiarism detection, author identification, and author profiling tasks, respectively. Finally, section 5 summarizes the main conclusions of the evaluation lab.

2 Plagiarism Detection

This section gives a brief report on the results of the plagiarism detection task at PAN 2015. An extended version of this report can be found in [46,15], where a more in-depth analysis of the obtained results is given. This year marks the beginning of a complete task overhaul, introducing community-driven corpus construction as a new way of developing evaluation resources with diversity. This lays the groundwork to drive the plagiarism detection task toward a truly community-driven evaluation, where ideally all aspects of the task are self-organized. This complements our previous efforts to improve the reproducibility of shared tasks by means of software submission using the TIRA experimentation platform.

2.1 Related Work

Research on plagiarism detection has a long history, both within PAN and without. Within PAN, we have been the first to organize shared tasks on plagiarism detection [50], whereas since then, we have introduced a number of variations of the task as well as new evaluation resources: the first shared task in 2009 focused on two sub-problems of plagiarism detection, namely the traditional external plagiarism detection [67], where a reference collection is used to identify plagiarized passages, and intrinsic plagiarism detection [32,66], where no such reference collection is at hand and plagiarism has to be identified from writing style changes within a document. For the first shared task in 2009, we have created the first standardized, large-scale evaluation corpus for plagiarism detection [49]. As part of this effort, we have devised the novel performance measures which for the first time took into account task-specific characteristics of plagiarism detection, such as detection granularity. Finally, in the first three years of PAN, we have also introduced cross-language plagiarism detection as a sub-task of plagiarism detection for the first time [40], and introduced corresponding problem instances into the corpus. Altogether, in the first three years, we successfully acquired and evaluated the plagiarism detection approaches of 42 research teams from around the world, some participating more than once. Many insights have been gained from this experience which informed our subsequent activities [50,39,41].

Starting in 2012, we have completely overhauled our evaluation approach to plagiarism detection based on the insights gained from the previous years [42]. Since then, we have separated external plagiarism detection into the two tasks of source retrieval and text alignment. The former task deals with information retrieval approaches to retrieve potential sources for a suspicious document from a large text collection, such as the web, which has been indexed with traditional retrieval models. The latter task of text alignment focuses on the problem of extracting matching passages from pairs of documents, if there are any. Both tasks have never been studied in this way before, whereas most of the existing body of work can be considered to deal mostly with text alignment.

For source retrieval, we went to considerable lengths to set up a realistic evaluation environment: we have obtained and indexed the entire English portion of the ClueWeb09 corpus, building the research search engine ChatNoir [47]. ChatNoir served two purposes, namely as an API for plagiarism detectors for those who cannot afford to index the ClueWeb themselves, but also as an end user search engine for authors which were hired to construct a new, realistic evaluation resource for source retrieval. We have hired 18 semi-professional authors from the crowdsourcing platform oDesk (now Upwork) and asked them to write essays of length at least 5000 words on pre-defined topics obtained from the TREC web track. To write their essays, the authors were asked to conduct research using ChatNoir, reusing text from the web pages they found. This way, we have created realistic information needs which in turn lead the authors to use our search engine in a realistic way to fulfill their task. This has lead to new insights into the nature of how humans reuse text, some building up a text as they go, whereas others first collect a lot of text and then boil it down to the final essay [48]. Finally, we have devised and developed new evaluation measures for source retrieval that for the first time take into account the retrieval of near-duplicate results when calculating precision and recall [43,45].

Regarding text alignment, we focus on the text reuse aspects of the task by boiling down the problem to its very core, namely comparing two text documents to identify reused passages of text. In this task, we have started in 2012 to experiment with software submissions for the first time, which lead to the development of the TIRA experimentation platform [14]. We have continued to employ this platform as a tool to collect softwares also for source retrieval and the entire PAN evaluation lab as of 2013, thus improving the reproducibility of PAN's shared tasks for the foreseeable future [13,44]. Altogether, in the second three-year cycle of this task, we have acquired and evaluated the plagiarism detection approaches of 20 research teams on source retrieval and 31 research teams on text alignment [42,43,45].

2.2 Community-Driven Construction of Evaluation Resources

Traditionally, the evaluation resources required to run a shared task are created by its organizers—but the question remains: why? Several reasons come to mind:

- *Seniority.* Senior community members may have the best vantage point in order to create representative evaluation resources.
- *Closed data access.* Having access to an otherwise closed data source (e.g., from a company) gives some community members an advantage over others in creating evaluation resources with a strong connection to the real world.
- *Task inventorship.* The inventor of a new task (i.e., a task that has not been considered before), is in a unique position to create normative evaluation resources, shaping future evaluations.
- *Being first to the table.* The first one to pick up the opportunity may take the lead in constructing evaluation resouces (e.g., because a task has never been organized before, or, to mitigate a lack of evaluation resources).

All of the above reasons are sufficient for an individual or a small group of researchers to become organizers of a shared task, and, to create corresponding evaluation resources. However, from reviewing dozens of shared tasks that have been organized in the human language technologies, we can conclude that neither of them is a necessary condition [44].

We question the traditional connection of shared task organization and evaluation resource creation, since this imposes several limitations on scale and diversity and therefore the representativeness of the evaluation resources that can be created:

- *Scale.* The number of man-hours that can be invested in the creation of evaluation resources is limited by the number of organizers and their personal commitment. This limits the scale of the evaluation resources. Crowdsourcing may be employed as a means to increase scale in many situations, however, this is mostly not the case where task-specific expertise is required.
- *Diversity.* The combined task-specific capabilities of all task organizers may be limited regarding the task's domain. For example, the number of languages spoken

by task organizers is often fairly small, whereas true representativeness across languages would require evaluation resources from at least all major language families spoken today.

By involving participants in a structured way into the creation of evaluation resources, task organizers may build on their combined expertise, man-power, and diversity.

2.3 Text Alignment Corpus Construction

In text alignment, given a pair of documents, the task is to identify all contiguous passages of reused text between them. The challenge with this task is to identify passages of text that have been obfuscated, sometimes to the extent that, apart from stop words, little lexical similarity remains between an original passage and its reused counterpart. Consequently, for task organizers, the challenge is to provide a representative corpus of documents that emulate this situation.

For the previous editions of PAN, we have created such corpora ourselves, whereas obfuscated text passages have been generated automatically, semi-automatically via crowdsourcing [5], and by collecting real cases. Until now, however, we neglected participants of our shared task as potential helpers in creating evaluation resources. Given that a stable community has formed around this task in previous years, and that the corpus format has not changed in the past three years, we felt confident to experiment with this task and to switch from algorithm development to corpus construction.

Corpus Construction Task The task was to construct an evaluation corpus for text alignment, where two possibilities to accomplish this task were given as follows:

- *Corpus collection.* Gather real-world instances of text reuse or plagiarism, and annotate them.
- *Corpus generation.* Given pairs of documents, generate passages of reused or plagiarized text between them. Apply a means of obfuscation of your choosing.

The task definition is cast as open as possible, imposing no particular restrictions on the way in which participants approach this task, which languages they consider, or which kinds of obfuscation they collect or generate. To ensure compatibility among each other and with previous corpora, however, the format of all submitted corpora had to conform with that of the existing corpora. By fixing the corpus format, future editions of the text alignment task may build on the evaluation resources created within this task without further effort, and the softwares that have been submitted in previous editions of the text alignment task and are now available at the TIRA experimentation platform may be re-evaluated on the new corpora. The latter in fact forms part of the analysis of the submitted corpora. To ensure compatibility, we handed a corpus validation tool that checked all format restrictions.

Corpus Validation and Analysis The creation of new evaluation corpora must be done with the utmost care, since corpora are barely double-checked or questioned again once they have been accepted as authoritative. This presents the organizers of a corpus construction task with the new challenge of evaluating submitted corpora, where the evaluation of a corpus should aim at establishing its validity.

Unlike with traditional shared tasks, the validity of a corpus can not only be established via an automatically computed performance measure, but requires manual reviewing effort. As part of their participation, all participants who submitted a corpus therefore had to peer-review the corpora submitted by other participants. Furthermore, we also publicly invited community members of PAN to volunteer to review submitted corpora. The following instructions were handed out to the reviewers:

> The peer-review is about dataset validity, i.e. the quality and realism of the plagiarism cases. Conducting the peer-review includes:
>
> - *Manual* review of as many examples as possible from all datasets and all obfuscation strategies therein
> - Make observations about how the dataset has been constructed
> - Make observations about potential quality problems or errors
> - Make observations on the realism of each dataset and each obfuscation strategy
> - Write about your observations in your notebook (make sure to refer to examples from the datasets for your findings).

Handing out the complete submitted corpora for peer-review, however, is out of the question, since this would defeat the purpose of subsequent shared task evaluations by revealing the ground truth prematurely. Therefore, the organizers of a corpus construction task serve as mediators, splitting submitted corpora into training and test datasets, and handing out only the training portion for peer-review. The participants who submitted a given corpus, however, may never be reliably evaluated based on their own corpus. Also, colluding participants may not be ruled out entirely.

Finally, when a shared task has previously invited software submission, this creates ample opportunity to re-evaluate the existing softwares on the submitted corpora. This allows for evaluating submitted corpora in terms of difficulty of detecting enclosed plagiarism cases: the performances of existing software on submitted corpora, when compared to their respective performances on previously used corpora, allow for a relative assessment of corpus difficulty. In our case, more than 30 text alignment softwares have been submitted since 2012.

Submitted corpora A total of 8 corpora have been submitted to the PAN 2015 text alignment corpus construction task. The corpora are of varying sizes and diversity: some corpora feature languages, such as Chinese, Persian, and Urdu, which were previously unobtainable to us. Some corpora feature real plagiarism cases, other automatically generated plagiarism.

A survey of the peer-reviews conducted by participants as well as the outlined evaluation of corpus difficulty based on software submitted to previous editions of the PAN text alignment task is forthcoming and will form part of the task overview paper [46].

3 Author Identification

The main idea behind author identification is that it is possible to reveal the author of a text given a set of candidate authors and undisputed text samples for each one of them [19,61]. The most crucial information for this task refers to writing style and it is essential to be able to quantify stylistic choices in texts and measure stylistic similarity between texts. 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.) and literary applications (e.g., verifying the authorship of disputed novels, identifying the author of works published anonymously, etc.) [10,20]

The author identification task has several variations depending mainly on the number of candidate authors and whether the set of candidate authors is closed or open. One particular variation of the task is authorship verification where there is only one candidate author for whom there are undisputed text samples and we have to decide whether an unknown text is by that author or not [24,16,29]. 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 positive class. All texts written by other authors can be viewed as the negative class, a huge and heterogeneous class from which it is not easy to find representative samples. However challenging, authorship verification is a very significant task since any given author identification problem can be decomposed into a set of authorship verification problems. The verification task is a fundamental problem in authorship analysis and provides an excellent research field to examine competitive approaches aiming at the extraction of reliable and general conclusions [25].

Previous PAN/CLEF editions have focused on the authorship verification task and achieved to produce appropriate evaluation corpora covering several natural languages and genres [21,64]. 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, most previous work in this field assumes that all texts within a verification problem match for both genre and thematic area. This assumption makes things easier 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.

PAN/CLEF 2015 also focuses on authorship verification but it no longer makes the assumption that all texts within a problem match for genre and thematic area. This cross-genre and cross-topic variation of the verification task corresponds to a more realistic view of the issue at hand since in many applications it is not possible to require undisputed text samples by certain authors in specific genres and topics. For instance, when one wants to verify the authorship of a suicide note it does not make sense to look for samples of suicide notes by the suspects [10]. In addition, the author of a crime fiction novel published anonymously could be a famous author of child fiction[20].

3.1 Related Work

Most of previous work in authorship verification (and more general in authorship analysis) only concern the case where the examined documents match for genre and topic [65,16,29,25]. A notable exception is reported in [24] where the *unmasking* method was applied to author verification problems where multiple topics were covered producing very reliable results. Kestemont *et al.* used the same method in a cross-genre experiment based on a corpus of prose and theatrical works by the same authors demonstrating that unmasking (with default settings) is not so effective in such difficult cases.

Stamatatos [62] presents 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). 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, it was shown that character n-grams corresponding to word affixes and including punctuation marks are the most significant features in cross-topic authorship attribution [57]. In addition, Sapkota *et al.* demonstrated that using training texts from multiple topics (instead of a single topic) can significantly help to correctly recognize the author of texts on another topi [58].

3.2 Evaluation Setup

The evaluation setup for this task is practically identical to the one used in PAN-2014. 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. Text length varies from a few hundred to a few thousand words, depending on genre. The only difference with PAN-2014 is that texts within a problem do not match for genre and/or thematic area.

Participants are asked to submit their software that 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) 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) and c@1 [37]. 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 participant teams are ranked by the final score (AUC · c@1).

Baselines One of the advantages of using TIRA for the evaluation of software submissions is that it supports the continuous evaluation of software in newly developed corpora. This enables us to apply methods submitted in previous editions of PAN to the cross-genre and cross-topic corpora of PAN-2015. That way, we can avoid the use of simplistic random-guess baselines (corresponding to final score = 0.25) and establish more challenging baselines that can be adapted to the difficulty of the corpus. In more detail, one of the best performing methods submitted to the author verification task at PAN-2013 (the winner approach when AUC is considered) [18] is also applied to evaluation corpora. In the reminder this approach is called PAN13-BASELINE. In addition, the second winner [12] and the third winner [6] of the corresponding PAN-2014 task are also used as baseline methods. For the rest of this paper, these two methods are called PAN14-BASELINE-1 and PAN14-BASELINE-2, respectively. It should be underlined that these methods have been trained and fine-tuned using different corpora and under

Table 1. The new cross-genre and cross-topic author identification corpus.

Language	Type	#problems	#docs	Known docs/ problem (avg.)	Words/doc (avg.)
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
Evaluation					
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 assumption that all documents within a problem match for genre and topic. Therefore, their performance on cross-genre and cross-topic author verification corpora is by no means optimized.

3.3 Corpus

Although it is relatively simple to compile a corpus of texts by different authors belonging to different genres/topics (negative instances of the verification task) it is particularly challenging to populate the corpus with appropriate positive instances (texts in different genres/topics by the same author). A new corpus was built that matches the volume of PAN-2014 and covers the same four languages: Dutch, English, Greek, and Spanish. The corpus is divided into a training part and an evaluation part as can be seen in Table 1. There are important differences between the sub-corpora for each language. In Dutch part, the known and unknown documents within a problem differ in genre while in English, Greek, and Spanish parts they differ in topic. In the English part only one known document per problem is provided. In Dutch and Greek parts the number of known documents per problem varies while in the Spanish part four known texts per problem are available. In all parts of the corpus, positive and negative instances are equally distributed.

The Dutch corpus is a transformed version of the *CLiPS Stylometry Investigation* corpus that includes documents written by language students at the University of Antwerp between 2012 and 2014 in two genres: essays and reviews [69]. The English corpus is a collection of dialogue from plays where the lines spoken by actors on the stage were extracted. Character names, stage directions, lists of characters, and so forth, were all removed. All positive verification instances comprise parts from different plays by the same author. English part is the largest in terms of verification problems. The Greek corpus is a collection of opinion articles published in the online forum Protagon[1] where all documents are categorized into several thematic categories (e.g. Politics, Economy, Science, Health, Media, Sports, etc). The Spanish corpus consists of opinion articles taken from a variety of online newspapers and magazines, as well as

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

Table 2. Author identification results in terms of final score (AUC · c@1).

Team (alphabetically)	Dutch	English	Greek	Spanish	Micro-avg	Macro-avg
Bagnall	0.451	**0.614**	**0.750**	0.721	**0.608**	**0.628**
Bartoli *et al.*	0.518	0.323	0.458	**0.773**	0.417	0.506
Castro-Castro *et al.*	0.247	0.520	0.391	0.329	0.427	0.365
Gómez-Adorno *et al.*	0.390	0.281	0.348	0.281	0.308	0.323
Gutierrez *et al.*	0.329	0.513	0.581	0.509	0.479	0.478
Halvani	0.455	0.458	0.493	0.441	0.445	0.462
Hürlimann *et al.*	0.616	0.412	0.599	0.539	0.487	0.538
Kocher & Savoy	0.218	0.508	0.631	0.366	0.435	0.416
Maitra *et al.*	0.518	0.347	0.357	0.352	0.378	0.391
Mechti *et al.*	-	0.247	-	-	0.207	0.063
Moreau *et al.*	**0.635**	0.453	0.693	0.661	0.534	0.606
Nikolov *et al.*	0.089	0.258	0.454	0.095	0.217	0.201
Pacheco *et al.*	0.624	0.438	0.517	0.663	0.480	0.558
Pimas *et al.*	0.262	0.257	0.230	0.240	0.253	0.247
Posadas-Durán *et al.*	0.132	0.400	-	0.462	0.333	0.226
Sari & Stevenson	0.381	0.201	-	0.485	0.286	0.250
Solórzano *et al.*	0.153	0.259	0.330	0.218	0.242	0.235
Vartapetiance & Gillam	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	0.255	0.249	0.198	0.443	0.269	0.280
PAN14-BASELINE-2	0.191	0.409	0.412	0.683	0.406	0.405
PAN13-BASELINE	0.242	0.404	0.384	0.367	0.358	0.347

personal web pages or blogs covering a variety of topics. It also includes literary essays. This is a mixed corpus meaning that in some verification problems there is a noticeable difference in topic and/or genre while in other problems the documents match for genre and they only differ in nuances of the topic.

3.4 Evaluation Results

In total, 18 teams submitted their software for this task. The submitted author verification approaches processed each part of the corpus separately. The majority of them were able to process all four parts of the evaluation corpus, one for each language. Table 2 provides the final score (AUC · c@1) for each part of corpus together with micro-averages and macro-averages (a more detailed view in the evaluation results can be found in [63]). Note that the English part is much larger with respect to the number of problems. Thus, macro-average provides a fair picture of the ability of submitted methods to handle all four sub-corpora. In average, the best results were produced for the cross-topic Greek part. Quite predictably, the cross-genre Dutch part proved to be the most challenging followed by the English part (this can be explained by the low number of known documents per problem). Note also that Greek and Spanish parts comprise longer texts (in average more than 500 words per document) while Dutch and English parts include shorter texts (less than 500 words per document).

The best performing approach, in terms of both micro-average and macro-average of final score, introduces a character-level Recurrent Neural Network model [3]. This method seems to be particularly effective for cross-topic verification cases while, based on the relatively low performance on the Dutch part, it seems to be affected by differences in genre. The second best overall performing approach by Moreau *et al.* is based on a heterogeneous ensemble combined with stacked generalization [34]. The success of this model verifies the conclusions of previous editions of PAN that different verification models when combined can achieve very good results [21,64].

In contrast to previous PAN editions, the majority of participants used eager supervised learning methods (e.g. SVMs, random forests) to model the verification process based on the training corpus. The best performing submitted methods belong to this category with the notable exception of the winner approach. The most successful methods also adopt the extrinsic verification paradigm where the one-class verification problem is transformed to a binary classification task by making use of texts from other authors [64]. The vast majority of submitted approaches attempt to combine a variety of text representation features. Most of them can be extracted from texts without any elaborate text analysis (e.g., word/sentence/paragraph length, character and word n-grams, etc.) The most common elaborate type of features depends on POS tagging. Only a couple of methods make use of full syntactic parsing. A more detailed review of the submitted approaches is given in [63].

The performance of the baseline models reflects the difficulty of the evaluation corpora. In the cross-genre Dutch part, all three baselines resemble a random-guessing classifier. PAN13-BASELINE and PAN14-BASELINE-2 provide relatively good results for the cross-topic English and Greek corpora while the performance of PAN14-BASELINE-1 is notably low. This may be explained by the fact that the latter method is based on eager supervised learning so it depends too much on the properties of the training corpus [12]. Both PAN14-BASELINE-1 and PAN14-BASELINE-2 are remarkably improved when applied to the mixed Spanish corpus where some verification problems match the properties of PAN-2014 corpora. In average, PAN13-BASELINE and PAN14-BASELINE-2 outperform almost half of the participant teams demonstrating their potential as generic approaches that can be used in any given corpus. On the other hand, the average performance of PAN14-BASELINE-2 resembles random-guessing.

Combining all participants Following the successful practice of previous PAN editions, we developed a simple meta-model combining all participant methods. This heterogeneous ensemble is based on the average of scores produced by all 18 participants for each verification problem. The evaluation results of this approach can also be seen in Table 2. In contrast to the corresponding results of PAN-2013 and PAN-2014 [21,64], the ensemble of all participants is not the best performing approach. When the micro-average and macro-average of final score is concerned, the ensemble is outperformed by 5 and 4 participants, respectively. This moderate performance of the meta-model can be partially explained by the low average performance of the submitted methods. This is demonstrated by the fact that all PAN-2014 participants acquired a micro-average final score greater than 0.3 while 6 out of 18 PAN-2015 participants get a micro-average final score lower than 0.3 (recall that the final score of a random-guessing model is 0.25).

4 Author Profiling

Author profiling distinguishes between classes of authors studying their sociolect aspect, that is, how language is shared by people. This helps in identifying profiling aspects such as gender, age, native language, or personality type. Author profiling is a problem of growing importance in applications in forensics, security, and marketing.

E.g., from a forensic linguistics perspective one would like being able to know the linguistic profile of the author of a harassing text message (language used by a certain type of people) and identify certain characteristics (language as evidence). Similarly, from a marketing viewpoint, companies may be interested in knowing, on the basis of the analysis of blogs and online product reviews, the demographics of people that like or dislike their products.

4.1 Related Work

Pennebaker's [38] investigated how the style of writing is associated with personal attributes such as age, gender and personality traits, among others. In [2] the authors approached the task of gender identification from the British National Corpus and achieved approximately 80% accuracy. Similarly in [17] and [4] the authors investigated age and gender identification from formal texts. Recently most investigations focus on social media. For example, in [23] and [59] the authors investigated the style of writing in blogs. On the other hand, Zhang and Zhang [71] experimented with short segments of blog post and obtained 72.1% accuracy for gender prediction. Similarly, Nguyen et al. [35] studied the use of language and age among Dutch Twitter users. Since 2013 a shared task on author profiling has been organised at PAN [56,55]. It is worth mentioning the second order representation based on relationships between documents and profiles used by the best performing team at the PAN-AP 2013 and 2014 [28,27]. Recently, the EmoGraph [53] graph-based approach tried to capture how users convey verbal emotions in the morphosyntactic structure of the discourse, obtaining competitive results with the best performing systems at PAN 2013. Moreover with the PAN-AP-2013 dataset, the authors in [70] investigate a high variety of different features and show the contribution of IR-based features in age and gender identification and in [30] the authors approached the task with 3 million features in a MapReduce configuration, obtaining high accuracies with fractions of processing time.

With respect to automatically recognising personality from text, Argamon *et al.* [68] focused on two of the Big Five traits (Extraversion and Emotional Stability), measured by means of self-reports. They used Support Vector Machines (SVMs), trained on word categories and relative frequency of function words, to recognize these two traits. In a similar way, Oberlander and Nowson [36] worked on the classification of personality types of bloggers extracting patterns in a bottom-up fashion. Mairesse *et al.* [31], investigated systematically the usefulness of different sets of textual features exploiting psycholinguistic dictionaries such as LIWC and MRC. They extracted personality models from self-reports and observed data, and reported that the openness to experience trait yield the best performance. In more recent years, the interest in personality recognition has grown in two areas: the analysis of human behaviour and social network analysis. Several studies have started exploring the wealth of behavioral data made available by cameras, microphones [33], wearable sensors [22], and mobile phones [11] linking personality traits to dimensions such as face to face interaction, speech video and text transcriptions. From the other hand, researchers have also focused on personality prediction from corpora of social network data, like Twitter and Facebook, exploiting either linguistic features in status updates, social features such as friends count, and daily activity [51,9]. Kosinski *et al.* [26] made an extensive analysis of different features,

Table 3. Distribution of Twitter users with respect to age classes per language.

	Training				Early birds				Test			
	EN	ES	IT	DU	EN	ES	IT	DU	EN	ES	IT	DU
18-24	58	22			16	6			56	18		
25-34	60	56			16	14			58	44		
35-49	22	22			6	6			20	18		
50+	12	10			4	4			8	8		
Σ	152	110	38	34	42	30	12	10	142	88	36	32

including the size of friendship network, uploaded photos count and events attended, finding the correlations with the personality traits of 180000 Facebook users. They reported very good results in the automatic prediction of Extraversion. Bachrach et al. made an extensive analysis of the network traits (i.e. such as size of friendship network, uploaded photos, events attended, times user has been tagged in photos) that correlate with personality of 180000 Facebook users. They predicted personality scores using multivariate linear regression, and reported good results on extraversion. Schwartz *et al.* [60] analyzed 700 million words, phrases, and topic instances collected from the Facebook messages of 75000 volunteers, who also filled a standard Big Five personality test. In 2013 [8] and 2014 [7] evaluation campaigns on personality recognition have been organised in the framework of the workshop on computational personality recognition.

4.2 Experimental Settings

In the Author Profiling task at PAN 2015 participants approached the task of identifying age, gender and personality traits from Twitter in four different languages: English, Spanish, Dutch and Italian. The corpus was annotated with the help of an online questionnaire. In this test, users reported their age and gender and self-assessed their personality traits with the BFI-10 online test[2] [52]. For labelling age, the following classes were considered: 18-24; 25-34; 35-49; 50+. The dataset was split into training, early birds and test, as in previous editions. The number of authors per language and age class can be seen in Table 3. The corpus is balanced per gender but imbalanced per age.

We have used two different measures for evaluation: accuracy and Root Mean Square Error (RMSE). For the identification of age and gender, and also for the joint identification, the accuracy measure was used. The accuracy is calculated as the ratio between the number of authors correctly predicted and the total number of authors. RMSE was used to evaluate personality prediction. It measures how far is the predicted value to the actual value for each trait. RMSE is calculated as in Formula 1, where n is the number of authors, f_i the actual value for trait i and $\widehat{f_i}$ the predicted one.

[2] In order to address ethical and privacy issues, authors were asked for their permission to use the tweets when answering the personality test. The dataset was anonymised, password protected, and released to task participants only.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\widehat{f_i} - f_i)^2}{n}} \tag{1}$$

We averaged the five RMSEs in order to obtain a global measure for personality prediction. The overall performance per language was obtained as the average between the joint identification accuracy and the (1-RMSE) for the personality recognition, as indicated in Formula 2.

$$rank = \frac{(1 - RMSE) + joint\_accuracy}{2} \tag{2}$$

Finally, the global ranking was obtained as the arithmetic average of the global measures per language.

4.3 Evaluation Results

This year 22 have been the teams who submitted software and notebook papers. In this section we show a summary of the obtained results. In Table 4 the overall performance per language and users' ranking are shown. The approach of [1] performs best overall and it is on the top 3 in every language. The authors combine the second order representation that allowed them to obtain the best results in PAN task in 2013 and 2014 together with Latent Semantic Analysis. We can observe that the highest accuracies were obtained in the Dutch dataset, with values over 90% in some cases, although it is the dataset with the lower number of authors. On the other hand, the worst results were obtained in the English dataset, although it has the highest number of authors. This may be due to the absence of age identification in Dutch that makes the task easier for that language. Something similar happens with more related languages such as Italian and Spanish, where accuracies in the first one are higher.

In Figure 1 the distribution of accuracies per language is shown. As can be seen, results in Spanish are the most sparse ones. Concretely, participants obtained accuracies from 0.5049 to 0.8215. Furthermore, results are more concentrated below the median (0.6547). Except in Dutch, there are slightly more extreme results in the lower bound. However in Dutch, the outliers occur in the upper bound, for instance with accuracies over 90%.

In Table 5 the best results per language and task are shown. In comparison to previous years of PAN, systems obtained much higher accuracy value in both age and gender identification. This may suggest that, although the shorter length of individual tweets and their informality, the amount of tweets per author is good enough to profile age and gender with high accuracy. With respect to personality recognition, we can see that the best results were obtained for Italian and Dutch. This is contrary to what we may have expected due to the smaller number of authors for both languages both in training and test. With respect to each trait, it seems that the *Stable* one is the most difficult to predict as opposed to maybe *Conscientious* and *Openness*. A more in-depth analysis of the results and the different approaches can be found in [54].

Table 4. Global ranking as average of each language global accuracy.

Ranking	Team	Global	English	Spanish	Italian	Dutch
1	alvarezcarmona15	0.8404	0.7906	0.8215	0.8089	0.9406
2	gonzalesgallardo15	0.8346	0.7740	0.7745	0.8658	0.9242
3	grivas15	0.8078	0.7487	0.7471	0.8295	0.9058
4	kocher15	0.7875	0.7037	0.7735	0.8260	0.8469
5	sulea15	0.7755	0.7378	0.7496	0.7509	0.8637
6	miculicich15	0.7584	0.7115	0.7302	0.7442	0.8475
7	nowson15	0.7338	0.6039	0.6644	0.8270	0.8399
8	weren15	0.7223	0.6856	0.7449	0.7051	0.7536
9	poulston15	0.7130	0.6743	0.6918	0.8061	0.6796
10	maharjan15	0.7061	0.6623	0.6547	0.7411	0.7662
11	mccollister15	0.6960	0.6746	0.5727	0.7015	0.8353
12	arroju15	0.6875	0.6996	0.6535	0.7126	0.6843
13	gimenez15	0.6857	0.5917	0.6129	0.7590	0.7790
14	bartoli15	0.6809	0.6557	0.5867	0.6797	0.8016
15	ameer15	0.6685	0.6379	0.6044	0.7055	0.7260
16	cheema15	0.6495	0.6130	0.6353	0.6774	0.6723
17	teisseyre15	0.6401	0.7489	0.5049	0.6024	0.7042
18	mezaruiz15	0.6204	0.5217	0.6215	0.6682	0.6703
19	bayot15	0.6178	0.5253	0.5932	0.6644	0.6881
	ashraf15	-	0.5854	-	-	-
	kiprov15	-	0.7211	0.7889	-	-
	markov15	-	0.5890	0.5874	-	0.6798

Table 5. Best results per language and tasks

	Age and Gender			**Personality Traits**					
Language	*Joint*	Gender	Age	*RMSE*	E	S	A	C	O
English	0.7254	0.8592	0.8380	0.1442	0.1250	0.1951	0.1305	0.1101	0.1198
Spanish	0.7727	0.9659	0.7955	0.1235	0.1319	0.1631	0.1034	0.1017	0.1108
Italian	-	0.8611	-	0.1044	0.0726	0.1555	0.0527	0.1093	0.0972
Dutch	-	0.9688	-	0.0563	0.0750	0.0637	0.0000	0.0619	0.0354

5 Conclusions

PAN/CLEF 2015 evaluation lab attracted a high number of teams from all around the world. This demonstrates that the topics of the shared tasks are of particular interest for researchers. New corpora have been developed covering multiple languages for plagiarism detection, author identification and author profiling. These new resources together with the produced evaluation results largely define the state of the art in the respective areas.

In the last editions of PAN, the same basic tasks are repeated. However, each year variations of these tasks are taken into account and significant novelties are introduced. This practice enables us to establish a suitable evaluation framework composed by large scale corpora and appropriate evaluation measures without having to start from scratch every year. In addition, it permits participants from past years to improve their method and adopt it in order to handle the peculiarities of certain variations of tasks.

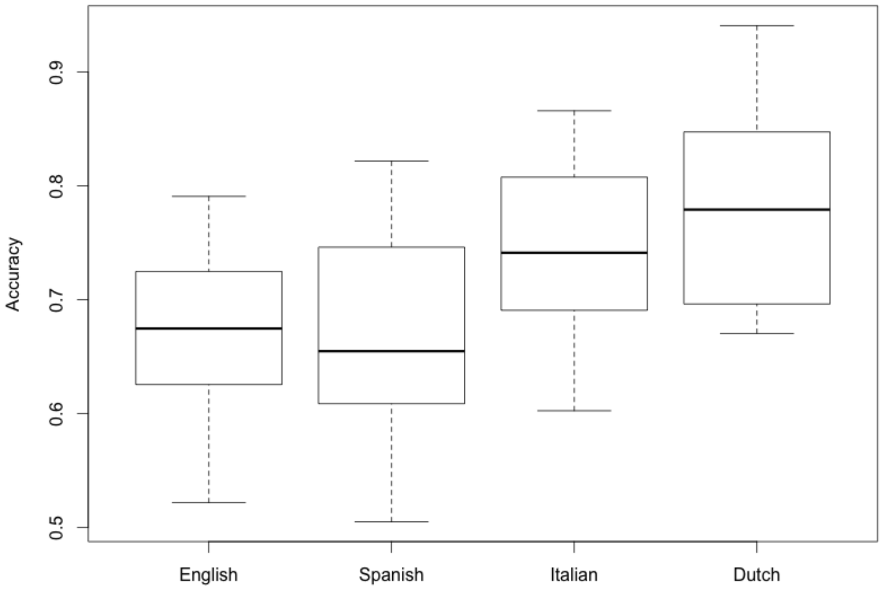

Figure 1. Distribution accuracies per language.

PAN requires software submissions to be evaluated within the TIRA experimentation platform. This procedure proved to be quite successful. It ensures credibility and reproducibility of the reported results while it enables to perform cross-year experiments where the submitted methods of one year are evaluated on a corpus of another year. That way, it is possible to establish challenging baselines (applying past methods to new corpora) and combine different models for the same task.

Acknowledgements

We thank the organizing committees of PAN's shared tasks Fabio Celli, Walter Daelemans, Ben Verhoeven, Patrick Juola, and Aurelio López-López. Our special thanks go to all of PAN's participants. This work was partially supported by the WIQ-EI IRSES project (Grant No. 269180) within the FP7 Marie Curie action.

References

1. Álvarez-Carmona, M.A., López-Monroy, A.P., Montes-Y-Gómez, M., Villaseñor-Pineda, L., Jair-Escalante, H.: INAOE's Participation at PAN'15: Author Profiling task—Notebook for PAN at CLEF 2015. In: CLEF 2013 Working Notes. CEUR. (2015)

2. Argamon, S., Koppel, M., Fine, J., Shimoni, A.R.: Gender, Genre, and Writing Style in Formal Written Texts. TEXT 23, 321–346 (2003)
3. Bagnall, D.: Author Identification Using Multi-headed Recurrent Neural Networks. In: CLEF 2015 Working Notes. CEUR. (2015)
4. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating Gender on Twitter. In: Proceedings of EMNLP '11. ACL. (2011)
5. Burrows, S., Potthast, M., Stein, B.: Paraphrase Acquisition via Crowdsourcing and Machine Learning. ACM TIST 4(3), 43:1–43:21 (2013)
6. 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. (2014)
7. Celli, F., Lepri, B., Biel, J.I., Gatica-Perez, D., Riccardi, G., Pianesi, F.: The Workshop on Computational Personality Recognition 2014. In: Proceedings of ACM MM'14. (2014)
8. Celli, F., Pianesi, F., Stillwell, D., Kosinski, M.: Workshop on Computational Personality Recognition: Shared Task. In: Proceedings of WCPR at ICWSM 2013 (2013)
9. Celli, F., Polonio, L.: Relationships Between Personality and Interactions in Facebook. In: Social Networking: Recent Trends, Emerging Issues and Future Outlook. Nova Science Publishers, Inc (2013)
10. Chaski, C.E.: Who's at the Keyboard: Authorship Attribution in Digital Evidence Invesigations. International Journal of Digital Evidence 4 (2005)
11. Chittaranjan, G., Blom, J., Gatica-Perez, D.: Mining Large-scale Smartphone Data for Personality Studies. Personal and Ubiquitous Computing 17(3), 433–450 (2013)
12. Fréry, J., Largeron, C., Juganaru-Mathieu, M.: UJM at clef in author identification. In: CLEF 2014 Labs and Workshops, Notebook Papers. CEUR. (2014)
13. Gollub, T., Potthast, M., Beyer, A., Busse, M., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Recent Trends in Digital Text Forensics and its Evaluation. In: Proceedings of CLEF 2013. Springer. (2013)
14. Gollub, T., Stein, B., Burrows, S.: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Proceedings of SIGIR 12. ACM (2012)
15. Hagen, M., Potthast, M., Stein, B.: Source Retrieval for Plagiarism Detection from Large Web Corpora: Recent Approaches. In: CLEF 2015 Working Notes. CEUR. (2015)
16. van Halteren, H.: Linguistic Profiling for Author Recognition and Verification. In: Proceedings of ACL 04. ACL. (2004)
17. Holmes, J., Meyerhoff, M.: The Handbook of Language and Gender. Blackwell Handbooks in Linguistics, Wiley (2003)
18. Jankowska, M., Keselj, V., Milios, E.: CNG Text Classification for Authorship Profiling Task—Notebook for PAN at CLEF 2013. In: CLEF 2013 Working Notes. CEUR. (2013)
19. Juola, P.: Authorship Attribution. Foundations and Trends in Information Retrieval 1, 234–334 (2008)
20. Juola, P.: How a Computer Program Helped Reveal J.K. Rowling as Author of A Cuckoo's Calling. Scientific American (2013)
21. Juola, P., Stamatatos, E.: Overview of the Author Identification Task at PAN-2013. In: CLEF 2013 Working Notes. CEUR. (2013)
22. Kalimeri, K., Lepri, B., Pianesi, F.: Going Beyond Traits: Multimodal Classification of Personality States in the Wild. In: Proceedings of ICMI 13. ACM. (2013)
23. Koppel, M., Argamon, S., Shimoni, A.R.: Automatically Categorizing Written Texts by Author Gender. Literary and Linguistic Computing 17(4). (2002)
24. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring Differentiability: Unmasking Pseudonymous Authors. J. Mach. Learn. Res. 8, 1261–1276 (2007)

25. 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)
26. Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., Graepel, T.: Manifestations of User Personality in Website Choice and Behaviour on Online Social Networks. Machine Learning. (2013)
27. López-Monroy, A.P., y Gómez, M.M., Jair-Escalante, H., nor Pineda, L.V.: Using Intra-Profile Information for Author Profiling—Notebook for PAN at CLEF 2014. In: CLEF 2014 Working Notes. CEUR. (2014)
28. Lopez-Monroy, A.P., Montes-Y-Gomez, M., Escalante, H.J., Villasenor-Pineda, L., Villatoro-Tello, E.: INAOE's Participation at PAN'13: Author Profiling Task—Notebook for PAN at CLEF 2013. In: CLEF 2013 Working Notes. CEUR. (2013)
29. Luyckx, K., Daelemans, W.: Authorship Attribution and Verification with many Authors and Limited Data. In: Proceedings of COLING 08. (2008)
30. Maharjan, S., Shrestha, P., Solorio, T., Hasan, R.: A Straightforward Author Profiling Approach in MapReduce. In: Advances in Artificial Intelligence. Iberamia. (2014)
31. Mairesse, F., Walker, M.A., Mehl, M.R., Moore, R.K.: Using Linguistic Cues for the Automatic Recognition of Personality in Conversation and Text. Journal of Artificial Intelligence Research 30(1), 457–500 (2007)
32. Meyer zu Eißen, S., Stein, B.: Intrinsic Plagiarism Detection. In: Proceedings of ECIR 06. LNCS, vol. 3936. Springer. (2006)
33. Mohammadi, G., Vinciarelli, A.: Automatic personality perception: Prediction of Trait Attribution Based on Prosodic Features. Affective Computing, IEEE Transactions on 3(3), 273–284 (2012)
34. Moreau, E., Jayapal, A., Lynch, G., Vogel, C.: Author Verification: Basic Stacked Generalization Applied to Predictions from a set of Heterogeneous Learners. In: CLEF 2015 Working Notes. CEUR. (2015)
35. Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T.: "How old do you think I am?"; A Study of Language and Age in Twitter. Proceedings of ICWSM 13. AAAI. (2013)
36. Oberlander, J., Nowson, S.: Whose Thumb is it Anyway?: Classifying Author Personality from Weblog Text. In: Proceedings of COLING 06. ACL. (2006)
37. Peñas, A., Rodrigo, A.: A Simple Measure to Assess Non-response. In: Proceedings of HLT '11. ACL. (2011)
38. Pennebaker, J.W., Mehl, M.R., Niederhoffer, K.G.: Psychological Aspects of Natural Language Use: Our Words, Our Selves. Annual Review of Psychology 54(1), 547–577 (2003)
39. Potthast, M., Barrón-Cedeño, A., Eiselt, A., Stein, B., Rosso, P.: Overview of the 2nd International Competition on Plagiarism Detection. In: CLEF 2010 Working Notes. CEUR. (2010).
40. Potthast, M., Barrón-Cedeño, A., Stein, B., Rosso, P.: Cross-Language Plagiarism Detection. Language Resources and Evaluation (LRE) 45, 45–62 (2011)
41. Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection. In: CLEF 2011 Working Notes (2011).
42. Potthast, M., Gollub, T., Hagen, M., Graßegger, J., Kiesel, J., Michel, M., Oberländer, A., Tippmann, M., Barrón-Cedeño, A., Gupta, P., Rosso, P., Stein, B.: Overview of the 4th International Competition on Plagiarism Detection. In: CLEF 2012 Working Notes. CEUR. (2012)
43. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 5th International Competition on Plagiarism Detection. In: CLEF 2013 Working Notes. CEUR. (2013).

44. 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: Proceedings of CLEF 14. Springer. (2014)
45. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B.: Overview of the 6th International Competition on Plagiarism Detection. In: CLEF 2014 Working Notes. CEUR. (2014)
46. Potthast, M., Göring, S., Rosso, P., Stein, B.: Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment. In: CLEF 2015 Working Notes. CEUR. (2015)
47. Potthast, M., Hagen, M., Stein, B., Graßegger, J., Michel, M., Tippmann, M., Welsch, C.: ChatNoir: A Search Engine for the ClueWeb09 Corpus. In: Proceedings of SIGIR 12. ACM. (2012)
48. Potthast, M., Hagen, M., Völske, M., Stein, B.: Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: Proceedings of ACL 13. ACL. (2013)
49. Potthast, M., Stein, B., Barrón-Cedeño, A., Rosso, P.: An Evaluation Framework for Plagiarism Detection. In: Proceedings of COLING 10. ACL. (2010)
50. Potthast, M., Stein, B., Eiselt, A., Barrón-Cedeño, A., Rosso, P.: Overview of the 1st International Competition on Plagiarism Detection. In: Proceedings of PAN at SEPLN 09. CEUR. (2009)
51. Quercia, D., Lambiotte, R., Stillwell, D., Kosinski, M., Crowcroft, J.: The Personality of Popular Facebook Users. In: Proceedings of CSCW 12. ACM. (2012)
52. Rammstedt, B., John, O.: Measuring Personality in One Minute or Less: A 10 Item Short Version of the Big Five Inventory in English and German. In: Journal of Research in Personality. (2007)
53. Rangel, F., Rosso, P.: On the Impact of Emotions on Author Profiling. In: Information Processing & Management, Special Issue on Emotion and Sentiment in Social and Expressive Media (In Press) (2014)
54. Rangel, F., Rosso, P., Celli, F., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: CLEF 2015 Working Notes. CEUR. (2015)
55. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd Author Profiling Task at PAN 2014. In: CLEF 2014 Working Notes. CEUR. (2014)
56. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013—Notebook for PAN at CLEF 2013. In: CLEF 2013 Working Notes. CEUR. (2013)
57. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all Character N-grams are Created Equal: A Study in Authorship Attribution. In: Proceedings of NAACL 15. ACL. (2015)
58. Sapkota, U., Solorio, T., Montes-y-Gómez, M., Bethard, S., Rosso, P.: Cross-topic Authorship Attribution: Will Out-of-topic Data Help? In: Proceedings of COLING 14. (2014)
59. Schler, J., Koppel, M., Argamon, S., Pennebaker, J.W.: Effects of Age and Gender on Blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs. AAAI (2006)
60. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PloS one 8(9), 773–791 (2013)
61. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology 60, 538–556 (2009)

62. Stamatatos, E.: On the Robustness of Authorship Attribution Based on Character N-gram Features. Journal of Law and Policy 21, 421–439 (2013)
63. Stamatatos, E., Daelemans, W., Verhoeven, B., Juola, P., López-López, A., Potthast, M., Stein, B.: Overview of the Author Identification Task at PAN-2015. In: Working Notes Papers of the CLEF 2015 Evaluation Labs. CEUR. (2015)
64. 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: CLEF 2014 Working Notes. CEUR. (2014)
65. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic Text Categorization in Terms of Genre and Author. Comput. Linguist. 26(4), 471–495 (2000)
66. Stein, B., Lipka, N., Prettenhofer, P.: Intrinsic Plagiarism Analysis. Language Resources and Evaluation (LRE) 45, 63–82 (2011)
67. Stein, B., Meyer zu Eißen, S.: Near Similarity Search and Plagiarism Analysis. In: Proceedings of GFKL 05. Springer. (2006)
68. Sushant, S.A., Argamon, S., Dhawle, S., Pennebaker, J.W.: Lexical Predictors of Personality Type. In: In Proceedings of Joint Interface/CSNA 2005.
69. 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 LREC 2014. ACL. (2014)
70. Weren, E., Kauer, A., Mizusaki, L., Moreira, V., de Oliveira, P., Wives, L.: Examining Multiple Features for Author Profiling. In: Journal of Information and Data Management. (2014)
71. Zhang, C., Zhang, P.: Predicting gender from blog posts. Tech. rep., Technical Report. University of Massachusetts Amherst, USA (2010)