Evolution of the PAN Lab on Digital Text Forensics

Paolo Rosso, Martin Potthast, Benno Stein, Efstathios Stamatatos, Francisco Rangel and Walter Daelemans

Abstract PAN is a networking initiative for digital text forensics, where researchers and practitioners study technologies for text analysis with regard to originality, authorship, and trustworthiness. The practical importance of such technologies is obvious for law enforcement, cyber-security, and marketing, yet the general public needs to be aware of their capabilities as well to make informed decisions about them. This is particularly true since almost all of these technologies are still in their infancy, and active research is required to push them forward. Hence PAN focuses on the evaluation of selected tasks from the digital text forensics in order to develop large-scale, standardized benchmarks, and to assess the state of the art. In this chapter we present the evolution of three shared tasks: plagiarism detection, author identification, and author profiling.

Paolo Rosso
PRHLT Research Center, Universitat Politècnica de València, Spain, e-mail: prosso@dsic.upv.es

Martin Potthast
Text Mining and Retrieval, Leipzig University, Germany, e-mail: martin.potthast@uni-leipzig.de

Benno Stein
Web Technology & Information Systems, Bauhaus-Universität Weimar, Germany, e-mail: benno.stein@uni-weimar.de

Efstathios Stamatatos
Dept. of Information and Communication Systems Engineering, University of the Aegean, Greece, e-mail: stamatatos@aegean.gr

Francisco Rangel
Autoritas Consulting, S.A., Spain & PRHLT Research Center, Universitat Politècnica de València, Spain, e-mail: francisco.rangel@autoritas.es

Walter Daelemans
CLiPS - Computational Linguistics Group, University of Antwerp, Belgium, e-mail: walter.daelemans@uantwerpen.be

1 Introduction

PAN[1] has become one of the main events for the digital text forensics community and it gathers a large audience of experts from information retrieval, natural language processing, and machine learning. The first two editions of PAN were organized in the form of workshops (2007-2008) at the conferences SIGIR 2007 and ECAI 2008 respectively. Since 2009, shared tasks have been organized at PAN, since 2010 Labs at CLEF, and since 2011 also at FIRE. At CLEF we have organized 31 shared tasks on authorship, originality, and trust: plagiarism detection (2010-2015), author identification (2011-2017), author profiling (2013-2017), Wikipedia vandalism detection (2010-2011), Wikipedia quality flaw detection (2012), sexual predator identification (2012), and author obfuscation (2016-2017). Each shared task had a considerable impact on its respective research field. Table 1 overviews key figures of the PAN Lab at CLEF in terms of registrations, runs / software, notebooks, attendees, and followers (Gollub et al, 2013; Potthast et al, 2014a; Stamatatos et al, 2015b; Rosso et al, 2016; Potthast et al, 2017). Since 2012 all of our shared tasks invite participants for *software* submissions instead of run submissions: more than 300 pieces of software have been submitted to PAN 2012 through PAN 2017, which have been repeatedly evaluated using the TIRA experimentation platform (Gollub et al, 2012a,b).

At FIRE[2] we organized 10 PAN shared tasks on text reuse / plagiarism detection in several languages (Arabic, Gujarati, Hindi, Persian) (Barrón-Cedeno et al, 2013; Gupta et al, 2012, 2013; Bensalem et al, 2015; Asghari et al, 2016), on source code texts (Flores et al, 2014, 2015), as well as on author profiling (Bengali, Hindi, Kannada, Malayalam, Russian[3], Tamil and Telegu[4]) also addressing novel research aspects such as personality recognition in source code (Rangel et al, 2016a).

In this chapter we will describe three of the shared tasks that we have organized at CLEF: plagiarism detection, author identification, and author profiling. The rest of this chapter is structured as follows. The next section is devoted to plagiarism detection: evolution of tasks, evaluation framework, and submitted approaches. Section three is on the evolution of tasks in author identification (closed/open set attribution, verification, clustering, diarization, and style breach detection) and the submitted approaches. Section four is on author profiling and its evolution (age, gender, personality, and language variety), background about the employed corpora, and the performance of the submitted approaches. The last section contains conclusions and discusses research aspects that we plan to address in the near future in the framework of the PAN Lab at CLEF.

---

[1] Initially, PAN stood for "Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection" http://pan.webis.de

[2] At CLEF 2010 in Padua, Carol Peters suggested cross-fertilization across evaluation forums.

[3] http://en.rusprofilinglab.ru/rusprofiling-at-pan

[4] http://nlp.amrita.edu:8080/INLI/Test.html

Table 1 Key figures of the PAN shared tasks at CLEF.

	2010	2011	2012	2013	2014	2015	2016	2017
Followers	151	181	232	286	302	333	337	347
Registrations	53	52	68	110	103	148	143	191
Runs/Software	27	27	48	58	57	54	37	34
Notebooks	22	22	34	47	36	52	29	30
Attendees	25	36	61	58	44	74	40	48

2 Plagiarism Detection

The first two editions of PAN were organized as workshops at the SIGIR 2007 and ECAI 2008 conferences. With the third edition at the SEPLN 2009 conference, PAN was organized for the first time as a (single) shared task: plagiarism detection. As it turned out, the time was ripe for this task, also evidenced by the comparably high number of first-time participants for a single shared task at that time.

Regardless of the fact that research on plagiarism detection was lacking (from algorithmic and conceptual perspectives) in various respects, the most pressing deficit probably was the missing evaluation and comparison of existing approaches. For decades, scientists published their findings individually, making up their own evaluation data, methodology, and performance measures ad hoc—often without consulting the relevant literature first. Comparisons between different related approaches were hardly ever conducted, so that an interested researcher entering the field had the problem of guessing which of the approaches reflected the state of the art or provided the strongest baseline. This shortcoming was addressed by our series of shared tasks: key contributions include corpora that have been created manually via crowdsourcing or semi-automatically, the implementation of a sophisticated evaluation setup including custom-built search engines, large-scale manual essay writing to simulate text reuse, and the first-time definition of suitable performance measures, incorporating the specifics of the task.

We want to point out that a technology that claims to detect plagiarism, in fact, does not. Instead, it detects evidence of text reuse, which may or may not be sufficient to judge whether an author plagiarized with a certain probability. While our task should have been called *text reuse detection* instead of plagiarism detection, we recognized that the misnomer was still justified: people both within and outside of academia search for "plagiarism detection", whereas hardly anyone is familiar with the term "text reuse detection". In light of this fact, we continue to use the former term for our task to ensure that interested people will find it in the future, while focusing our attention on text reuse detection as well as mentioning the connection in appropriate places.

From a public relations perspective the timing of this topic could not have been better. By 2011, when the shared task was in its third edition, we were fully prepared for the major plagiarism scandal that hit Germany in that year: it was discovered that the then-minister of defense, Karl Theodor zu Guttenberg, had plagiarized considerable parts of his dissertation, copying bits and pieces of text from more

than 200 sources to fill up his 400-page thesis, resulting in both the loss of his doctor's degree and his position. With the ensuing public outcry a number of other theses of famous politicians were checked and dozens more cases were found. These realistic cases of plagiarism provided us with an intriguing baseline to judge whether plagiarism detection technology was mature enough to detect such cases, as well as renewed the research interest in the task itself, which lasts to this day. Interestingly, dedicated plagiarism detection software hardly played a role in resolving these cases; the analyses were done manually, by up to hundreds of people who collaborated within Wikis to crowdsource their detection efforts. Also note in this regard, that there is a high chance that the collection of real cases is skewed towards ease of detection, while the difficult cases where plagiarizing authors carefully paraphrased the text they reused may have gone unnoticed. Aside from privacy issues this is another reason why these real cases can only serve as an additional source of data for evaluating plagiarism detectors. New corpora are needed to render the task of analysis more realistic beyond the detection of verbatim copy-paste operations.

2.1 Evolution of Tasks

The plagiarism detection task was organized for seven successive years, starting in 2009. In previous research (Stein et al, 2007), we interpreted plagiarism detection as a two-step retrieval process, which, given a suspicious document, consists of the tasks: (1) a source retrieval task, executed against a large collection of reference documents such as the web, followed by (2) a text alignment task, performed on the retrieved candidate sources against the suspicious documents with the objective of extracting plagiarized passages.[5]

In the first edition of our task, called external plagiarism detection, our goals were twofold (Potthast et al, 2009): (1) to create the first benchmark for plagiarism detection under the aforementioned retrieval process, consisting of suspicious documents with and without plagiarism, the former being drawn from a large-scale reference collection of documents obtained from the Project Gutenberg[6]; (2) to scale that setup to an—at that time—large-enough size so that participants would not just compare all pairs of documents to each other but to force them to do some sort of source retrieval within their approach. To ensure that the extraction of plagiarized passages from pairs of suspicious documents and retrieved candidate sources would be non-trivial, we applied so-called obfuscation strategies in order to emulate plagiarists attempting to hide their plagiarism by paraphrasing the reused texts. We implemented a number of automatic obfuscation strategies, which, for lack of a working paraphrasing model, ranged from random text operations to parts-of-speech-preserving word re-

[5] In the beginning, the two tasks were called "candidate retrieval" and "detailed comparison" respectively. Later on, as the importance of evaluating these tasks in isolation became clear, we found our initial choice of names to be too unspecific and decided to rename them for clarification as "source retrieval" and "text alignment".

[6] http://www.gutenberg.org

orderings. Although the automatic obfuscation strategies served as a good baseline for a bag-of-words-oriented plagiarism detector, the obfuscated passages obtained were still unreadable and hence lacked an appropriate semantics. To render the obfuscation step more realistic, we resorted to crowdsourcing the required paraphrases on Amazon's Mechanical Turk, which was still a rather new tool at the time. The resulting paraphrases were manually written, so that they served as a more realistic sample of human paraphrasing ability at passage level; still, the obfuscated passages were inserted at random into suspicious documents and could be spotted rather easily by human readers. Nevertheless, it turned out we were among the first to use crowdsourcing for paraphrasing acquisition, and the first to do so at passage level, so that we published a corresponding spin-off corpus, the Webis Crowd Paraphrase Corpus (Webis-CPC-11). In addition, we also provided an in-depth analysis of the corpus quality as well as machine learning technology that allowed for automatic quality assessment during paraphrase acquisition, severely reducing construction costs (Burrows et al, 2013). As became clear upon the review of the 14 approaches submitted, none of the participants actually implemented source retrieval but all of them went to great lengths to compare every document from the reference collection to each of the suspicious documents. In hindsight, the number of 41 000 documents was already too small to impose a source retrieval step. Significantly increasing the corpus size was still impossible for us since we had already exhausted the entire Project Gutenberg for our purposes. And, simply adding documents from a different source (and hence: different genre) would have been too easy to be recognized and undone. As a consequence, instead of treating plagiarism detection as an atomic task, we decided to evaluate source retrieval and text alignment in isolation. Within the next two iterations of the plagiarism detection task (Potthast et al, 2010a, 2011), the evaluation setup was refined with a focus on text alignment, while we started to build a new and independent evaluation setup specifically suited to source retrieval.

In addition to the task above, we invented and hosted the task of *intrinsic* plagiarism detection (Stein et al, 2011). The goal of this task is to identify plagiarized passages without exploiting an external document collection. I.e., tackling this task means finding evidence for writing style changes, which in turn may indicate that some text from another author has been copied into a suspicious document at hand. Although rather clear in its design, intrinsic plagiarism detection contains a number of considerable challenges; it was the foray of PAN into the field of writing style analysis and can be seen as the precursor of various authorship-related tasks that PAN hosts today. Similar to the external plagiarism detection task, the task was repeated three times in a row, refining its setup from year to year.

Starting in 2012 (Potthast et al, 2012a), a new evaluation setup for plagiarism detection was ready for use. This setup enabled (and still enables) us to evaluate source retrieval tasks in much more realistic settings, separating it from the evaluation of text alignment tasks. The setup was used for four successive years (Potthast et al, 2013a, 2014b, 2015; Hagen et al, 2015). While the text alignment task did not change much, for the source retrieval task a new search engine called ChatNoir (Potthast et al, 2012b) was built, which indexed the entire ClueWeb 2009 (ClueWeb09, 2009). Using this search engine, we compiled—via expert crowdsourcing—also a new corpus of

manually created plagiarism. In particular, more than 20 writers were recruited, where each writer was asked to write essays about some topic of her choice from the TREC ad hoc track, yielding a total of 300 essays. Each essay was supposed to be of 5 000 words length, and the research required to write the essay had to be conducted with ChatNoir's web interface, reusing text from the web pages found. Moreover, the writers were instructed to obfuscate the reused text passages in a way they deemed sufficient to successfully pass plagiarism detectors. Some writers spent significant effort to do so while others did not, resulting in a range of case difficulties. This corpus, called the Webis Text Reuse Corpus 2012 (Webis-TRC-12) (Potthast et al, 2013b), formed the basis for several spin-off research inquiries (e.g., analyzing the writing behavior of writers during search (Hagen et al, 2016)) as well as follow-up shared tasks on author diarization at PAN (Stamatatos et al, 2016; Tschuggnall et al, 2017). Participants of the task had to treat a given essay from the corpus as suspicious document and to use the ChatNoir API to retrieve all sources from which an essay's author reused text fragments. The queries and downloads of potential sources of the submitted approaches were meticulously logged to measure their performance in terms of retrieval effort and recall. To relieve participants from the task of also implementing text alignment technology, a "source oracle" was provided, which classified a downloaded document either as true or as false source.

2.2 Evaluation Framework

The evaluation framework that has been developed within the series of shared tasks on plagiarism detection had a strong impact on the community (Potthast et al, 2010b). It is employed to this day and helps to evaluate new algorithms, ensuring the comparability of new and historical evaluation results. The evaluation framework consists of three components: (1) a collection of corpora for text alignment and source retrieval, (2) a static, reproducible web search environment for source retrieval, and (3) tailored performance measures for both tasks.

Altogether 26 corpora have been constructed for our shared tasks. The collection includes corpora that were submitted to shared tasks that specifically invited their participants to submit not just software, but also data. Following the example of our shared tasks, spin-offs have been organized in subsequent years at other conferences, dedicated to specific languages not previously covered. Our corpus collection serves as a diverse resource, allowing for the evaluation of plagiarism detectors under many different scenarios, especially regarding the difficulty of the to-be-detected plagiarism cases.

The static web search environment is comprised of the web search engine ChatNoir, which indexes the ClueWeb 2009, the ClueWeb 2012, and (as of 2017) the CommonCrawl, delivering search results in milliseconds while using a state-of-the-art retrieval model and a standard user interface. The web search environment comes along with a framework that allows for browsing web pages from the aforementioned corpora as if being in the live web, while serving clicks on hyperlinks with

the version of the crawled page instead of the live version. The user behavior in the framework can be logged, allowing for reproducible, large-scale user studies, as well as evaluating various search-based approaches, including source retrieval.

Our text alignment measures consider the *granularity* of a detection result (i.e., they penalize a detector if it returns bits and pieces of a plagiarized passages instead of the reused passage as a whole) as well as formulas for precision and recall that discount multiple, overlapping detections for a given suspicious document. The measures can be combined into the so-called "PlagDet score", which allows for an absolute ranking among evaluated plagiarism detectors (Potthast et al, 2010b). The source retrieval measures are based on recall but consider the effort in terms of queries and downloads as well. Again, multiple detections of the same source documents are discounted when retrieving web pages that are duplicates of each other.

2.3 Submitted Approaches

Over the years, many approaches haven been submitted to the plagiarism detection task and its variants—too many to review all of them here. Table 2 shows the distribution of participants across tasks. A total of 74 approaches have been submitted to external plagiarism detection and its successor task text alignment, 26 approaches have been submitted to source retrieval, and 10 to intrinsic plagiarism detection. The approaches submitted in or after 2012 for text alignment, and in or after 2013 for source retrieval, have been archived in an operational state and are still available for re-evaluation within TIRA.

The approaches submitted to a task showed certain commonalities—a fact, which allowed us to discern and organize a general (task-specific) retrieval process, comprising a number of task-specific steps. Each step in turn can be operationalized in numerous ways, and, once the algorithmic pattern was revealed to participants, it guided their developments and allowed newcomers to catch up quickly without having to reinvent the wheel.

A text alignment approach generally is divided into three steps: (1) seeding, (2) extension, and (3) filtering. The names of these steps are borrowed from gene sequence alignment, a task in bioinformatics that relates to text alignment in that the problem structure is similar, albeit not the solution space. The seeding step takes as input two documents and outputs matches between them in terms of pairs of phrases (one from each document) for which a matching heuristic checks similarity in order to argue about equivalent semantics. A commonly used matching heuristic outputs all matching word 4-grams whose words have been synonym-normalized, stemmed, and sorted alphabetically. The heuristic thereby raises the matching probability of two word 4-grams, even if the author of a plagiarized document paraphrases a text resulting in new word ordering at phrase level. Another heuristic employs so-called stop word 8-grams, which are 8-grams consisting only of the stop words in order of appearance in a text (Stamatatos, 2011). Since plagiarists often focus

Table 2 An overview of author identification tasks at PAN evaluation campaigns.

Year	Tasks	Language	Submissions
2009	External plagiarism detection,	English, German, Spanish	10
	Intrinsic plagiarism detection	English	4
2010	External plagiarism detection,	English, German, Spanish	18
	Intrinsic plagiarism detection	English	2
2011	External plagiarism detection,	English	9
	Intrinsic plagiarism detection	English	4
2012	Text alignment	English	10
	Source retrieval	English	5
2013	Text alignment	English	9
	Source retrieval	English	9
2014	Text alignment	English	11
	Source retrieval	English	6
2015	Text alignment data submission	English, Farsi, Urdu, Chinese	8
	Source retrieval	English	5
	External Plagiarism Detection	Arabic	3
	Intrinsic Plagiarism Detection	Arabic	2
2016	Text Alignment	Persian	9
	Text Alignment data submission	Persian	5
	Source retrieval	English	1
2017	Text Alignment	Russian	1
	Source retrieval	English	1

on exchanging content words rather than function words, matching sequences of stop words are a telltale sign of reused text. The finesse of devising matching heuristics between two texts determines to a great extent how well a text alignment approach works, since matches that were not identified during seeding render the subsequent step of extension difficult if not impossible. In this regard, the more matching heuristics are employed simultaneously, the better. The extension step takes as input the matches obtained from seeding, and outputs the boundaries of pairs of passages (one from each document) that may have been copied and pasted by the original author before paraphrasing. When interpreting each match as a point in a two-dimensional plane spanned by the two documents' characters, clustering technology can be used to extend text regions with dense amounts of seeds towards larger passages for which a human would judge that they have obviously been reused in bulk. Finally, the filtering step implements some postprocessing to exclude results that seem implausible according to certain criteria; most participants, however, just remove detections that would otherwise harm their performance in terms of granularity.

A source retrieval approach is divided into four steps: (1) chunking, (2) keyphrase extraction, (3) query formulation, and (4) query and download scheduling. The chunking step takes as input a suspicious document and outputs possibly overlapping chunks that cover the text. Many chunking strategies have been devised, but it turned out that non-overlapping 150-word chunks are sufficient. Each chunk is used as input for the keyphrase extraction step, where the k keywords or phrases that best describe the contents of a chunk (e.g., according to a $tf \cdot idf$ ranking) are returned. The small chunk size renders the task of selecting the top k for $k < 20$ easier. In addition,

keyphrases from the entire document may be extracted to allow for querying the document's general topic. The query formulation step takes a set of keyphrases as input and returns queries consisting of at most 5 phrases, formulated by combining individual phrases and often started from the top-ranked keyphrase. In the query and download scheduling step, the queries are submitted to a search engine while trying to ensure that the most promising queries are submitted first, and the most promising search results are downloaded first to minimize the time to result. Here, queries comprising nouns were found to be most successful. The number of downloads per query may vary dependent on whether one wants to maximize the F-measure or just the recall. In the latter case, downloading more search results yields significant returns, whereas the likelihood of finding a second true positive detection after the first one in a given search result is small. I.e., the next query scheduled should be used after a true positive detection, whereas one may explore up to a hundred search results per query. The examples given for the aforementioned steps are, in fact, the ones followed by the most effective approach in terms of recall (0.89), dwarfing the best previously achieved recall (0.59) (Hagen et al, 2017). A number of additional heuristics render this approach comparable in terms of its effort to the previously best one while maintaining its recall.

3 Author Identification

Author identification aims at revealing the authors behind texts. It is an active research area (Stamatatos, 2009) associated with important applications mainly in the humanities (e.g., unmasking the authors of novels published anonymously or under aliases), forensics (e.g., identifying the author of harassing messages, linking proclamations of terrorist groups), and social media analytics (e.g., revealing multiple user accounts controlled by the same person, verifying the authenticity of posts). Author identification tasks can be either supervised (i.e., the training texts are labeled with authorship information) (Argamon and Juola, 2011; Juola and Stamatatos, 2013) or unsupervised (i.e., authorship information is either not available or not reliable) (Stamatatos et al, 2016; Tschuggnall et al, 2017).

What makes author identification challenging is that it deals with the personal style of authors. In contrast to other factors, like topic or sentiment, usually style is not associated with certain words and there is no consensus about its quantification. Moreover, it is especially hard to discover style markers (i.e, style-related quantifiable textual features) that remain unaffected in topic shifts or genre variations. Another crucial factor is text-length. For very long documents (e.g., novels), there are quite reliable methods (Koppel et al, 2007). However, when short or very short (e.g., tweets) texts are considered, it is much harder to retain high effectiveness.

Table 3 An overview of author identification tasks at PAN evaluation campaigns.

Year	Tasks	Genre	Language	Submissions
2011	Closed-set attribution, Open-set attribution, Verification	Emails	English	16
2012	Closed-set attribution, Open-set attribution, Clustering	Fiction	English	12
2013	Verification	Textbooks, Fiction, Newspaper articles	English, Greek, Spanish	18
2014	Verification	Essays, Reviews, Newspaper articles, Fiction	Dutch, English, Greek, Spanish	13
2015	Verification	Essays, Reviews, Newspaper articles, Fiction	Dutch, English, Greek, Spanish	18
2016	Clustering, Diarization	Reviews, Newspaper articles	Dutch, English, Greek	10
2017	Clustering, Style breach detection	Reviews, Newspaper articles	Dutch, English, Greek	9

3.1 Evolution of Tasks

A significant part of PAN activities is related to author identification. PAN evaluation campaigns since 2011 explored several tasks as summarized in Table 3 and described below:

- Closed-set attribution: Given a set of candidate authors and some texts unquestionably written by each one of them, the task is to find the most likely author among them for another text of disputed authorship.
- Open-set attribution: This is similar to the previous task. However, it is possible that none of the candidate authors is the author of the disputed text.
- Verification: Given a set of texts all written by the same author, the task is to examine whether another text is also written by that author.
- Clustering: Given a set of texts of unknown authorship, the task is to group them by authorship.
- Diarization: Given a text that may be written by multiple co-authors, the task is to identify the authorial components of each co-author.
- Style breach detection: Given a text that may be written by multiple co-authors, the task is to detect all borders where authors switch.

In the different editions over the years, variations of the main task are examined. For example, the closed-set attribution task in 2011 focused on a large pool of candidate authors while the 2012 edition examined a small set of candidate authors. The first editions of the author verification task assumed that all texts within a verification case are in the same thematic area and belong to the same genre while the 2015 edition examined more challenging cross-topic and cross-genre cases. The first edition of the clustering task (2016) considered full texts while the 2017 edition focused on paragraph-length texts.

It has to be underlined that, in most of the cases, previous work in these tasks was extremely limited (e.g., author verification, clustering, diarization). PAN campaigns attracted the attention of multiple research groups around the world and contributed to enrich the literature in those areas. For all of these tasks, new benchmark corpora were developed covering several natural languages and genres (as shown in Table 3) that became the standard in the field. Moreover, appropriate evaluation measures were proposed for each task taking into account both crisp answers and confidence scores (Stamatatos et al, 2014, 2016; Tschuggnall et al, 2017). Especially, for the author verification task, emphasis was given to the fact that some cases could be left unanswered since in the applications related with this task it is better not to give an answer rather than giving a wrong answer. It was also demonstrated that authorship clustering can also be seen as a retrieval problem (Stamatatos et al, 2016).

The first editions of PAN related to author identification (2011-2012) were quite ambitious attempting to explore multiple tasks simultaneously. It soon became apparent that it is much better if each task is examined separately and in consecutive campaigns so that research groups are more mature and can develop more sophisticated approaches. Another important conclusion was that it is better to study simple rather than complicated tasks. For example, author verification can be seen as a fundamental task in author identification since any other task can be transformed into a series of author verification cases. Focusing on author verification enables us to better estimate the state-of-the-art performance in this area since we have to worry about fewer parameters (e.g., the number of candidate authors and the distribution of texts over the authors are not so crucial factors in verification as they are in closed-set attribution). Another example of a complicated task is author diarization. This can be decomposed into simpler tasks like style breach detection and clustering of short texts (Tschuggnall et al, 2017).

Another important outcome of PAN campaigns was to highlight the fact that there are strong relationships between author identification tasks. For example, author verification relates not only to closed-set and open-set attribution but to clustering as well. The top-performing approach in author verification at PAN 2015 was also the winning method in the clustering task in PAN 2016 (Bagnall, 2016). Moreover, as already mentioned, authorship clustering is a basic building block in the author diarization task. The latter is also strongly related with the task of intrinsic plagiarism detection considered in the early editions of PAN (Potthast et al, 2009, 2010a, 2011).

3.2 Submitted Approaches

Most of the author identification tasks attracted a large number of participant teams from all around the world. The submitted methods explored several models regarding the extraction of stylometric measures from texts and the attribution model. This section reviews the most important novelties and conclusions that can be drawn.

In the first editions of author identification tasks at PAN, it seemed that approaches based on a rich set of stylometric features combining several kinds of measures, in-

Table 4 Distribution of PAN participants in the author verification task.

Verification model	PAN 2013	PAN 2014	PAN 2015
Intrinsic	13	10	11
Extrinsic	3	3	7
Eager	2	3	10
Lazy	14	10	8
Profile-based	4	1	4
Instance-based	11	12	12
Hybrid	1	0	2

cluding measures extracted by natural language processing tools (NLP), are the most promising ones (Argamon and Juola, 2011). However, in subsequent shared tasks most of the top-performing submissions were based on low-level features like character and word n-grams. Such simplistic and language-independent features when combined with sophisticated attribution models can provide very good results (Stamatatos et al, 2014, 2015a, 2016; Tschuggnall et al, 2017). A particularly interesting and very effective approach is to apply neural network language models in stylometry as demonstrated by the character-level recurrent neural network model that won top-ranked overall positions in PAN 2015 and PAN 2016 tasks (Bagnall, 2015, 2016). On the other hand, more sophisticated approaches exclusively based on syntactic analysis of texts by NLP tools were easily outperformed by simpler approaches. This can also be attributed to the fact that in most of the cases the NLP tools used by PAN participants were not specifically trained to handle the types of texts included in PAN corpora, therefore they provided quite noisy stylometric measures.

Certainly, the widest variety of methods submitted to PAN tasks refers to author verification. Table 4 shows the distribution of PAN participants per year according to several factors. Extrinsic verification models attempt to transform an author verification case from a one-class classification task to a binary classification task by considering a collection of texts written by other authors (with respect to the author in question). A typical representative of this paradigm is the *Impostors* method introduced by Koppel and Winter (Koppel and Winter, 2014). Nevertheless, intrinsic verification models focus on one-class classification. In all three relevant editions of PAN, extrinsic verification models won top-ranked positions (Juola and Stamatatos, 2013; Stamatatos et al, 2014, 2015a). However, the majority of PAN participants followed the intrinsic verification paradigm and only in the last edition of the shared task in PAN 2015 was there an increase of extrinsic models. A crucial open issue is how to find the most suitable set of external texts for a given verification case. The external documents used by relevant PAN submissions were downloaded from the World Wide Web with the help of a search engine and queries formed by texts of the training corpus (Seidman, 2013; Khonji and Iraqi, 2014).

Another important perspective is how to handle the training corpus. Eager methods attempt to build a binary classifier that learns to distinguish between positive (same-author) and negative (different-author) verification cases. Each verification

case is an instance of this binary classification task and a classifier is trained based on the training corpus. Conversely, lazy methods essentially avoid extracting any general model from the training corpus and make their decisions separately for each evaluation case. The number of eager methods submitted to PAN increased over the years. This is certainly associated with the volume of the provided training corpora. In early editions of this task (PAN 2013) the training corpus consisted of a few dozens of verification cases while in the last two editions (PAN 2014 and PAN 2015) there were hundreds of verification cases in the training corpus. However, eager methods heavily depend on the representativeness of the training corpus. One eager method, trained on PAN 2014 corpora and among the best-performing submissions in PAN 2014 (Fréry et al, 2014), was also applied to PAN 2015 corpora (as a baseline model) and practically failed (Stamatatos et al, 2015a).

Verification models can also be described according to the way they handle texts of known authorship. One approach is to concatenate them and extract a single representation (profile-based paradigm). Another approach is to extract a separate representation from each known text (instance-based paradigm). Yet another case is to combine these two paradigms (hybrid methods). The majority of PAN submissions consistently follow the instance-based paradigm including the top-performing ones in most of the cases.

An important conclusion extracted from PAN shared tasks in author verification was that it is possible to combine different verification models and provide a robust approach with enhanced performance. A simple ensemble model that was based on averaging the answers of all PAN participants achieved better results than any of the individual models in the PAN 2013 and PAN 2014 author verification tasks. In the corresponding task at PAN 2015, the ensemble of all submissions was outperformed by some individual models mainly due to the relatively low average results of many participants. It is important to underline that the author verification task at PAN 2015 focused on very challenging cross-topic and cross-genre cases. However, the submission that ranked second-best overall was also based on a heterogeneous ensemble that combined several base verification models (Moreau et al, 2015). Actually this approach was the most effective in the most challenging cross-genre corpus in Dutch (Stamatatos et al, 2015a). This clearly shows that heterogeneous ensembles is a promising approach and most suitable for challenging author verification tasks.

4 Author Profiling

Author profiling aims at identifying personal traits of an author on the basis of her/his writings. Traits such as gender, age, language variety, or personality are of high interest for areas such as marketing, forensics, or security. From the marketing viewpoint, to be able to identify personal traits from comments to blogs or reviews, may provide the companies with the possibility of better segmenting their audience, which is an important competitive advantage. From a forensic linguistics perspective one would like to be able to know the linguistic profile of the author of a harassing

text message (language used by a certain type of people) and identify a certain type of person (language as evidence). From a security point of view, these technologies may allow to profile and identify possible delinquents or even terrorists. Traditional investigations in computational linguistics (Argamon et al, 2003) and social psychology (Pennebaker, 2013) have been carried out mainly for English. Furthermore, pioneer researchers such as (Argamon et al, 2003) or (Holmes and Meyerhoff, 2003), focused on formal and well-written texts. Although with the rise of social media, researchers such as (Koppel et al, 2003) and (Schler et al, 2006) have moved their focus to blogs and fora.

Since 2013 we have been organizing the author profiling task at PAN with several objectives. We have covered different profiling aspects (age, gender, language variety, personality), languages (Arabic, Dutch, English, Italian, Portuguese), and genres (blogs, reviews, social media, and Twitter). The international interest in the shared task is made evident by the number of participants from a large number of countries (Table 5). Furthermore, many have been researchers that have investigated further the performance of their approaches on the corpora that were developed for the shared task. For example, the best performing team in the three first editions used a second order representation which relates documents with author profiles and subprofiles (e.g., males talking about video games) (López-Monroy et al, 2015). The authors of (Weren et al, 2014) investigated a high variety of different features on the PAN AP-2013 dataset and showed the contribution of information retrieval based features in age and gender identification. In this approach, the text to be identified was used as a query for a search engine. In (Maharjan et al, 2014), the authors used MapReduce to approach the task with 3 million n-gram based features. They improved the accuracy as well as reduced the processing time considerably. Finally, the EmoGraph graph-based approach (Rangel and Rosso, 2016) tried to capture how users convey verbal emotions in the morphosyntactic structure of the discourse. The sequence of grammatical categories is modeled as a graph which is enriched with topics, semantics of verbs, polarity, and emotions. They reported competitive results with the best performing systems at PAN 2013 and demonstrating its robustness against genres and languages at PAN 2014 (Rangel and Rosso, 2015).

In the following sections we describe the evolution of the tasks, how the corpora have been built and the main approaches used by the participants, all from the perspective of the lessons learned during the organization of this task.

4.1 Evolution of Tasks

In Table 5, a summary of the evolution of the author profiling tasks at PAN is shown.

Table 5 An overview of author profiling tasks at PAN evaluation campaigns.

Year	Tasks	Genres	Languages	Submissions
2013	Age, Gender	Social Media	English, Spanish	21
2014	Age, Gender	Social Media, Twitter Blogs, Reviews	English, Spanish	10
2015	Age, Gender, Personality	Twitter	English, Spanish Italian, Dutch	22
2016	Age, Gender	Cross-genre	English, Spanish	22
2017	Gender, Language variety	Twitter	English, Spanish Arabic, Portuguese	22

The first edition was organized in 2013 (Rangel et al, 2013) with the aim of investigating the age and gender identification in a social media realistic scenario. We collected thousands of social media posts in English and Spanish with a high variety of topics. With respect to age, we considered three classes following what was previously done in (Schler et al, 2006): 10s (13-17), 20s (23-27) and 30s (33-47). Furthermore, we wanted to test the robustness of the systems when dealing with fake age profiles such as sexual predators. Therefore, we included in the collection some texts from the previous year PAN shared task on sexual predator identification (Inches and Crestani, 2012).

In the second edition (Rangel et al, 2014), we extended the task to other genres besides social media. Concretely, we focused also on Twitter, blogs, and hotel reviews, in English and Spanish. We realized the difficulty of obtaining quality labeled data and proposed a methodology to annotate age and gender. In 2014, we opted for modeling age in a continuous way and considered the following classes: 18-24; 25-34; 35-49; 50-64; 65+. Finally, the Twitter subcorpus was constructed in cooperation with RepLab (Amigó et al, 2014) in order to address also the reputational perspective (e.g., profiling social media influencers, journalists, professionals, celebrities, among others).

In 2015 (Rangel et al, 2015), besides the focus on age and gender identification, we introduced the task of personality recognition in Twitter. We maintained the age ranges defined in 2014 (except ”50-64” and ”65+” that were merged to ”50-XX”) and, besides English and Spanish, we included also Dutch and Italian (only gender and personality recognition). The objective of the shared task organized in 2016 (Rangel et al, 2016b) was to investigate the robustness of the systems in a cross-genre evaluation. That is, training the systems in one genre and testing its performance in other genres. Concretely, we provided Twitter data for training in English, Spanish, and Dutch. The approaches were then tested on blogs and social media genres in English and Spanish, and essays and reviews in Dutch.

Finally, in 2017 (Rangel et al, 2017) we introduced two novelties: the language variety identification (together with the gender), and the Arabic and Portuguese languages (besides English and Spanish). This is the first time a task has been organized covering together gender and language variety identification, and we obtained interesting insights relating both profiling aspects. Furthermore, we addressed language variety from fine-grain and course-grain perspective where varieties that are close

geographically were grouped together (e.g. Canada and United States, Great Britain and Ireland, or New Zealand and Australia).

4.2 Corpora Development

The author profiling task organized at PAN has been focusing on social media texts. Our interest was to study how people use language in their daily lives. Thus, in 2013 we retrieved thousands of social media posts with a wide spectrum of topics. The ample diversity of topics made possible to go beyond standard cliches, for example, men writing about sports and women about shopping. Furthermore, people may use social media to talk about sex. Some users can also cross the line and commit sexual harassment. With the aim of investigating the robustness of the author profiling approaches in detecting possible predators, we included some texts from the previous year PAN task on sexual predator identification[7] (Inches and Crestani, 2012). With this configuration, a realistic scenario was provided to the participants:

- A large dataset (big data).
- High variety of topics.
- Sexual conversations vs. sexual predators.
- Possible fake users and automatic generated content (e.g., chatbots).

This realistic scenario, however, presented some problems from the research perspective. The annotation (age and gender) was made on the basis of what the users self-reported, and they could have lied. Due to that, it was difficult to analyze errors: has the system failed or has it actually detected a fake profile? Therefore, we introduced a methodology to annotate data (and to not trust what users say). In the next subsections, we briefly describe this methodology for each trait.

Gender annotation based on dictionary and photos review

Depending on the genre, the annotation of the gender was based on different methods. In the case of blogs or reviews, the starting point were lists of well-known users (e.g., celebrities or politicians on the one hand, colleagues or students on the other). Furthermore in case of Twitter, we took advantage of meta-information to label the profiles in two steps:

- Firstly, the user name was searched in a dictionary of proper nouns. Users with ambiguous names were discarded.
- Secondly, each profile photograph was visually reviewed in order to ensure the right gender. Users with ambiguous photography (e.g., non-personal photos) were discarded.

[7] Texts from predators and adult-adult sexual conversations have been segmented into the corresponding age and gender groups.

Age annotation based on LinkedIn profiles

LinkedIn[8] is a professional network where people, among other things, can detail their resume. We looked for public LinkedIn profiles which share a personal blog URL or a Twitter account. We verified that the blog or the Twitter account existed, it was written in one of the languages we were interested in, and it was updated only by one person and this person was easily identifiable (we discarded organizational accounts). We looked for age information in the LinkedIn profile (in some cases the birth date is published). When this information was not available, we looked for the degree starting date in the education section. Following the information of Table 6, we figured out the age range. We discarded users whose education dates were not clear. To ensure the quality of the annotation, this process was done by two independent annotators and a third one decided in case of disagreement.

Table 6 Age range by degree starting date for data collected in the year 2014.

Degree starting date	Age group
2006-. . . .	18-24
1997-2006	25-34
1982-1996	35-49
1967-1981	50-64
.. . . -1966	65+

Personality traits annotation based on BFI-10 online test

Personality may be defined along five traits using the Five Factor Theory (Costa and McCrae, 2008), which is the most widely accepted model in psychology. The five traits are: openness to experience (O), conscientiousness (C), extraversion (E), agreeableness (A), and emotional stability / neuroticism (N). Personality traits, as well as users' gender and age, were self-assessed with the BFI-10 online test[9] (Rammstedt and John, 2007) and reported as scores normalized between -0.5 and +0.5.

The personality test consists of ten statements such as "I am a reserved person", "I have few artistic interests", or "I am sociable". The user has to valuate how much she/he agrees with each statement. Furthermore, she/he is asked for the age, gender, and Twitter account. This allowed us to retrieve the user's timeline and associate it with the profile aspects.

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

[9] We have created a web page with the BFI-10 test (http://mypersonality.autoritas.net) and promoted it in social media such as Twitter and Facebook

Language variety annotation based on geographical retrieval

A language variety is the specific form of a language that is shared by a group of people depending on their regional, social, or contextual situation. Taking advantage of Twitter geographical retrieval, we can obtain users who share a location and a language, and hence, a common language variety. To annotate users with their corresponding language variety, we have followed the following steps:

- Firstly, we decided which languages and language varieties will be part of the dataset. We selected four languages (Arabic, English, Portuguese and Spanish), and the varieties were selected following previous investigations (e.g. the selection of Arabic varieties followed (Sadat et al, 2014) as shown in Table 7).
- Varieties have been linked to geographical regions. For each language variety, the countries where this variety is used have been selected. Then, the capital cities (sometimes also the most populated cities) have been identified.
- Given the geographical coordinates of the capital cities, we have retrieved all the tweets generated in a radius around these coordinates (generally 15 kilometers).
- Unique authors who wrote the retrieved tweets have been identified. Their entire timeline was then retrieved. Tweets written in other languages or retweets have been removed.
- Users whose tweets were not geotagged in the corresponding coordinates, or whose location did not coincide with the corresponding capital city have been removed. This avoids the inclusion of users who wrote when temporarily being in a particular place (e.g., tourists or temporary workers).

Table 7 Language varieties.

Arabic	English	Portuguese	Spanish
Egypt	Australia	Brazil	Argentina
Gulf	Canada	Portugal	Chile
Levantine	Great Britain		Colombia
Maghrebi	Ireland		Mexico
	New Zealand		Peru
	United States		Spain
			Venezuela

Although according to the Oxford English Dictionary the definition of dialect refers to ”a variety of a language that is a characteristic of a particular group of the language’s speakers” and ”a language that is socially subordinated to a regional or national standard language”, the main criticism is that people from the same region are likely to talk about the same local topics. This may allow shallow topic-based methods to achieve competitive results. However, the obtained results showed that the best results could not be achieved only with topic-based features since they did not capture other linguistic patterns that are even more common such as differences in used characters (e.g., in English organise/organize), parts-of-speech sequences

(e.g., in Portuguese quero quixar-me/quero-me queixar *I want to complain*), or even words that appear only in some varieties (e.g., in Arabic, the words طاريك (*your remembrance*), شعاد (*but what about*) and لقياك (*meeting you*) are only used in the Gulf variety.)

4.3 Submitted Approaches

Following Pennebaker investigations (Pennebaker, 2013), most participants have combined different kinds of style-based features such as frequencies of punctuation marks, capital letters, quotations, and so on, together with Part-of-Speech tags or genre-specific features such as HTML-based features as image URLs, links, Twitter hashtags, or user mentions. Other stylistic markers such as the use of slang, contractions, or character flooding have been used as well.

Different content-based features have also been used: Latent Semantic Analysis, bag-of-words (weighted by frequency and tf-idf), dictionary-based words, topic-based words, entropy-based words, class-dependent words, named entities, etc. With respect to emotional features, some participants have extracted emotions, appraisal, admiration, positive/negative emoticons, positive/negative words, emojis, and sentiment words. Resources such as LIWC[10] have been widely used.

Language models based on different kinds of n-gram models (e.g., word, character) have been widely used in all the editions, obtaining competitive results, although almost always combined with other kinds of features. Other features such as readability indices (e.g., Flesch-Kinkaid, Gunning fog, SMOG, Coleman-Liau), information retrieval (the text to be identified was used as a query for a search engine), or collocations have been used by some participants. Finally, in recent years, especially in 2017, deep learning approaches have been widely used, mainly based on distributed representations such as word and character embeddings.

With respect to classification algorithms and their evolution, most of the participants have approached the task with traditional machine learning algorithms such as Logistic Regression, Support Vector Machines, Naive Bayes, BayesNet, or Random Forest. There have also been participants who approached the task with distance-based methods. It is difficult to highlight the best algorithms due to the combination of them by participants, but in most cases the best performing teams used Support Vector Machines.

As previously said, deep learning methods have been widely used: Recurrent Neural Networks and Convolutional Neural Networks with configurations of attention mechanism, max-pooling layer, or fully-connected layer. Although these deep learning approaches obtained good results, they did not achieve the best ones.

In Table 8, best results at PAN per trait and language (accuracy for age, gender and language variety, RMSE for personality) were achieved in Twitter. Best results were obtained in 2015 in age, personality and gender in Dutch, English, Italian,

[10] https://liwc.wpengine.com

and Spanish, in 2017 in language variety and gender identification in Arabic and Portuguese.

Table 8 Best results at PAN.

Trait	Arabic	Dutch	English	Italian	Portuguese	Spanish
Age	-	-	0.8380	-	-	0.7955
Gender	0.8031	0.9688	0.8592	0.8611	0.8700	0.9659
Language variety	0.8313	-	0.8988	-	0.9838	0.9621
Personality	-	0.0563	0.1442	0.1044	-	0.1235

5 Conclusions

The shared tasks of PAN are designed both to measure the technical state of the art and to foster the development of new approaches for important problems in the field of digital text forensics. The shared task principle seems to be ideally suited for this endeavor; in particular, it attracts different research groups from different fields, which all have their own view and solution approach to tackle such kinds of "ill-posed" or fuzzy problems. The fuzziness of most of the PAN shared task problems has several causes: the complexity of language, the complexity of features to describe language phenomena, the complex distribution of the phenomena over text registers, or the missing theory about corpus size and robust feature quantification, to mention a few.

We, at PAN, address this challenging research situation by evolving our shared tasks. Stated differently, we are looking for the "right" question that we want to ask the research community. The three strands of task evolutions presented in this chapter reflect this. However, the evolution must be driven carefully: we cannot completely re-model all tasks with each new PAN edition since (1) it may become too complicated for us to put together all pieces of the puzzle, and (2) we depend on the expertise that has been built up among the researchers of the participating teams, and we cannot require them to acquire and operationalize effective expertise from scratch each year. Hence we try to evolve the tasks in such a way that, on the one hand, they remain closely connected to the nature of the problem and, on the other hand, their variation brings enough insights to further develop the field. In this regard, we will continue the research on author identification, author profiling, or author obfuscation—although from different perspectives: cross-domain authorship attribution, style change detection, or multimodal author profiling (age and gender).

PAN has become a reference point in the digital text forensics community. Multiple shared tasks attracted a large number of participants and motivated research teams all over the world to start conducting research in this area. The corpora developed in the framework of PAN shared tasks have become standard benchmark datasets used in any subsequent study. Certainly, PAN copora are far from ideal and sometimes they may suffer from low volumes of data, noise, or lack of realism.

Therefore, maximizing the performance on those specific datasets should not be seen as panacea for the research community.

In addition, it is very important that PAN promotes reproducibility issues by requiring software submissions and encouraging participants to also provide their open-source code. All gathered approaches can be viewed as a library of tools, the largest in this area, available to replicate evaluation results and be applied to future corpora. The mere existence of this library enables the study of new tasks, like author obfuscation. PAN welcomes any other scientific use of this collection of software.

Acknowledgements The first author acknowledges the SomEMBED TIN2015-71147-C2-1-P MINECO research project. The work on the author profiling data in Arabic was made possible by NPRP grant #9-175-1-033 from the Qatar National Research Fund (a member of Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

Amigó E, Carrillo-de-Albornoz J, Chugur I, Corujo A, Gonzalo J, Meij E, de Rijke M, Spina D (2014) Overview of RepLab 2014: author profiling and reputation dimensions for online reputation management. In: Proceedings of the Fifth International Conference of the CLEF Initiative

Argamon S, Juola P (2011) Overview of the international authorship identification competition at PAN-2011. In: CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands

Argamon S, Koppel M, Fine J, Shimoni AR (2003) Gender, genre, and writing style in formal written texts. TEXT 23:321–346

Asghari H, Mohtaj S, Fatemi O, Faili H, Rosso P, Potthast M (2016) Algorithms and corpora for persian plagiarism detection: Overview of pan at fire 2016. In: Notebook Papers of FIRE 2016, FIRE-2016, Kolkata, India, December 7-10, CEUR Workshop Proceedings. CEUR-WS.org, vol 1737, pp 135–144

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

Bagnall D (2016) Authorship Clustering Using Multi-headed Recurrent Neural Networks. In: Balog K, Cappellato L, Ferro N, Macdonald C (eds) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, CEUR-WS.org

Barrón-Cedeno A, Rosso P, Devi SL, Clough P, Stevenson M (2013) Pan@fire: Overview of the cross-language !ndian text re-use detection competition. In: Notebook Papers of FIRE 2011, FIRE-2011, Mumbai, India, December 2-4

Bensalem I, Boukhalfa I, Rosso P, Abouenour L, Darwish K, Chikhi S (2015) Overview of the araplagdet pan@ fire2015 shared task on arabic plagiarism detection. In: Notebook Papers of FIRE 2015, FIRE-2015, Gandhinagar, India, December 4-6, CEUR Workshop Proceedings. CEUR-WS.org, vol 1587, pp 111–122

Burrows S, Potthast M, Stein B (2013) Paraphrase Acquisition via Crowdsourcing and Machine Learning. Transactions on Intelligent Systems and Technology (ACM TIST) 4(3):43:1–43:21, DOI http://dx.doi.org/10.1145/2483669.2483676

ClueWeb09 (2009) The ClueWeb09 Dataset, 2009. URL http://lemurproject.org/clueweb09/

Costa PT, McCrae RR (2008) The revised neo personality inventory (neo-pi-r). The SAGE handbook of personality theory and assessment 2:179–198

Flores E, Rosso P, Moreno L, Villatoro-Tello E (2014) Pan@fire: Overview of soco track on the detection of source code re-use. In: Notebook Papers of FIRE 2014, FIRE-2014, Bangalore, India, December 5-7

Flores E, Barrón-Cedeño A, Moreno L, Rosso P (2015) Pan@fire: Overview of cl-soco track on the detection of cross-language source code re-use 1587:1–5

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

Gollub T, Stein B, Burrows S (2012a) Ousting ivory tower research: towards a web framework for providing experiments as a service. In: Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval, ACM, pp 1125–1126

Gollub T, Stein B, Burrows S, Hoppe D (2012b) Tira: Configuring, executing, and disseminating information retrieval experiments. In: Database and expert systems applications (DEXA), 2012 23rd international workshop on, IEEE, pp 151–155

Gollub T, Potthast M, Beyer A, Busse M, Rangel F, Rosso P, Stamatatos E, Stein B (2013) Recent trends in digital text forensics and its evaluation: Plagiarism detection, author identification, and author profiling. In: 4th Int. Conf. of CLEF on Information Access Evaluation meets Multilinguality, Multimodality, and Visualization, CLEF 2013, Springer-Verlag, LNCS(8138), pp 53–58

Gupta P, Clough P, Rosso P, Stevenson M (2012) Pan@fire: Overview of the cross-language !ndian news story search (cl!nss) track. In: Notebook Papers of FIRE 2012, FIRE-2012, Kolkata, India, December 17-19

Gupta P, Clough P, Rosso P, Stevenson M, Banchs RE (2013) Pan@fire: Overview of the cross-language !ndian news story search (cl!nss) track. In: Notebook Papers of FIRE 2013, FIRE-2013, Delhi, India, December 4-6

Hagen M, Potthast M, Stein B (2015) Source Retrieval for Plagiarism Detection from Large Web Corpora: Recent Approaches. In: Working Notes Papers of the CLEF 2015 Evaluation Labs, CLEF and CEUR-WS.org, CEUR Workshop Proceedings, URL http://www.clef-initiative.eu/publication/working-notes

Hagen M, Potthast M, Völske M, Gomoll J, Stein B (2016) How Writers Search: Analyzing the Search and Writing Logs of Non-fictional Essays. In: Kelly D, Capra R, Belkin N, Teevan J, Vakkari P (eds) Proceedings of the 1st ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR 16), ACM, pp 193–202, DOI http://dx.doi.org/10.1145/2854946.2854969

Hagen M, Potthast M, Adineh P, Fatehifar E, Stein B (2017) Source Retrieval for Web-Scale Text Reuse Detection. In: Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM 17), ACM

Holmes J, Meyerhoff M (2003) The Handbook of Language and Gender. Blackwell Handbooks in Linguistics, Wiley

Inches G, Crestani F (2012) Overview of the International Sexual Predator Identification Competition at PAN-2012. In: Forner P, Karlgren J, Womser-Hacker C (eds) CLEF 2012 Evaluation Labs and Workshop – Working Notes Papers, 17-20 September, Rome, Italy

Juola P, Stamatatos E (2013) Overview of the author identification task at PAN 2013. In: Working Notes for CLEF 2013 Conference

Khonji M, Iraqi Y (2014) A slightly-modified gi-based author-verifier with lots of features (asgalf). In: CLEF 2014 Labs and Workshops, Notebook Papers, CLEF and CEUR-WS.org

Koppel M, Winter Y (2014) Determining if two documents are written by the same author. Journal of the American Society for Information Science and Technology 65(1):178–187

Koppel M, Argamon S, Shimoni AR (2003) Automatically categorizing written texts by author gender

Koppel M, Schler J, Bonchek-Dokow E (2007) Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research 8:1261–1276

López-Monroy AP, Montes-y Gómez M, Escalante HJ, Villaseñor-Pineda L, Stamatatos E (2015) Discriminative subprofile-specific representations for author profiling in social media. Knowledge-Based Systems 89:134–147

Maharjan S, Shrestha P, Solorio T, Hasan R (2014) A straightforward author profiling approach in mapreduce. In: Advances in Artificial Intelligence. Iberamia, pp 95–107

Moreau E, Jayapal A, Lynch G, Vogel C (2015) Author Verification: Basic Stacked Generalization Applied To Predictions from a Set of Heterogeneous Learners. In: Cappellato L, Ferro N, Gareth J, San Juan E (eds) Working Notes Papers of the CLEF 2015 Evaluation Labs

Pennebaker JW (2013) The Secret Life of Pronouns: What Our Words Say About Us. Bloomsbury USA

Potthast M, Stein B, Eiselt A, Barrón-Cedeño A, Rosso P (2009) Overview of the 1st International Competition on Plagiarism Detection. In: Stein B, Rosso P, Stamatatos E, Koppel M, Agirre E (eds) SEPLN 09 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), CEUR-WS.org, pp 1–9, URL http://ceur-ws.org/Vol-502

Potthast M, Barrón-Cedeño A, Eiselt A, Stein B, Rosso P (2010a) Overview of the 2nd International Competition on Plagiarism Detection. In: Braschler M, Harman D, Pianta E (eds) Working Notes Papers of the CLEF 2010 Evaluation Labs, URL http://www.clef-initiative.eu/publication/working-notes

Potthast M, Stein B, Barrón-Cedeño A, Rosso P (2010b) An Evaluation Framework for Plagiarism Detection. In: Huang CR, Jurafsky D (eds) 23rd International Conference on Computational Linguistics (COLING 10), Association for Computational Linguistics, Stroudsburg, Pennsylvania, pp 997–1005

Potthast M, Eiselt A, Barrón-Cedeño A, Stein B, Rosso P (2011) Overview of the 3rd International Competition on Plagiarism Detection. In: Petras V, Forner P, Clough P (eds) Working Notes Papers of the CLEF 2011 Evaluation Labs, URL http://www.clef-initiative.eu/publication/working-notes

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 (2012a) Overview of the 4th International Competition on Plagiarism Detection. In: Forner P, Karlgren J, Womser-Hacker C (eds) Working Notes Papers of the CLEF 2012 Evaluation Labs, URL http://www.clef-initiative.eu/publication/working-notes

Potthast M, Hagen M, Stein B, Graßegger J, Michel M, Tippmann M, Welsch C (2012b) ChatNoir: A Search Engine for the ClueWeb09 Corpus. In: Hersh B, Callan J, Maarek Y, Sanderson M (eds) 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12), ACM, p 1004, DOI http://dx.doi.org/10.1145/2348283.2348429

Potthast M, Gollub T, Hagen M, Tippmann M, Kiesel J, Rosso P, Stamatatos E, Stein B (2013a) Overview of the 5th International Competition on Plagiarism Detection. In: Forner P, Navigli R, Tufis D (eds) Working Notes Papers of the CLEF 2013 Evaluation Labs, URL http://www.clef-initiative.eu/publication/working-notes

Potthast M, Hagen M, Völske M, Stein B (2013b) Crowdsourcing Interaction Logs to Understand Text Reuse from the Web. In: Fung P, Poesio M (eds) Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 13), Association for Computational Linguistics, pp 1212–1221, URL http://www.aclweb.org/anthology/P13-1119

Potthast M, Gollub T, Rangel F, Rosso P, Stamatatos E, Stein B (2014a) Improving the reproducibility of pan's shared tasks: Plagiarism detection, author identification, and author profiling. In: 5th Int. Conf. of CLEF on Information Access Evaluation meets Multilinguality, Multimodality, and Interaction, CLEF 2014, Springer-Verlag, LNCS(8685), pp 268–299

Potthast M, Hagen M, Beyer A, Busse M, Tippmann M, Rosso P, Stein B (2014b) Overview of the 6th International Competition on Plagiarism Detection. In: Cappellato L, Ferro N, Halvey M, Kraaij W (eds) Working Notes Papers of the CLEF 2014 Evaluation Labs, CLEF and CEUR-WS.org, CEUR Workshop Proceedings, URL http://www.clef-initiative.eu/publication/working-notes

Potthast M, Göring S, Rosso P, Stein B (2015) Towards Data Submissions for Shared Tasks: First Experiences for the Task of Text Alignment. In: Working Notes Papers of the CLEF 2015 Evaluation Labs, CLEF and CEUR-WS.org, CEUR Workshop Proceedings, URL http://www.clef-initiative.eu/publication/working-notes

Potthast M, Rangel F, Tschuggnall M, Stamatatos E, Rosso P, Stein B (2017) Overview of pan'17: Author identification, author profiling, and author obfuscation. In: 8th Int. Conf. of CLEF

on Experimental IR Meets Multilinguality, Multimodality, and Visualization, CLEF 2017, Springer-Verlag, LNCS(10456), pp 275–290

Rammstedt B, John O (2007) 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, pp 203–212

Rangel F, Rosso P (2015) On the multilingual and genre robustness of emographs for author profiling in social media. In: 6th international conference of CLEF on experimental IR meets multilinguality, multimodality, and interaction, Springer-Verlag, LNCS(9283), pp 274–280

Rangel F, Rosso P (2016) On the impact of emotions on author profiling. Information processing & management 52(1):73–92

Rangel F, Rosso P, Moshe Koppel M, Stamatatos E, Inches G (2013) Overview of the author profiling task at pan 2013. In: Forner P., Navigli R., Tufis D. (Eds.), CLEF 2013 labs and workshops, notebook papers. CEUR-WS.org, vol. 1179

Rangel F, Rosso P, Chugur I, Potthast M, Trenkmann M, Stein B, Verhoeven B, Daelemans W (2014) Overview of the 2nd author profiling task at pan 2014. In: Cappellato L., Ferro N., Halvey M., Kraaij W. (Eds.) CLEF 2014 labs and workshops, notebook papers. CEUR-WS.org, vol. 1180

Rangel F, Rosso P, Potthast M, Stein B, Daelemans W (2015) Overview of the 3rd author profiling task at pan 2015. In: Cappellato L., Ferro N., Jones G., San Juan E. (Eds.) CLEF 2015 labs and workshops, notebook papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1391

Rangel F, González F, Restrepo F, Montes M, Rosso P (2016a) Pan at fire: Overview of the pr-soco track on personality recognition in source code. Notebook Papers of FIRE 2016, FIRE-2016, Kolkata, India, December 7-10, CEUR Workshop Proceedings CEUR-WSorg 1737:1–5

Rangel F, Rosso P, Verhoeven B, Daelemans W, Potthast M, Stein B (2016b) Overview of the 4th author profiling task at PAN 2016: cross-genre evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs, CLEF and CEUR-WS.org, CEUR Workshop Proceedings

Rangel F, Rosso P, Potthast M, Stein B (2017) Overview of the 5th author profiling task at pan 2017: Gender and language variety identification in twitter. Working Notes Papers of the CLEF

Rosso P, Rangel F, Potthast M, Stamatatos E, Tschuggnall M, Stein B (2016) Overview of the pan'2016 - new challenges for authorship analysis: Cross-genre profiling, clustering, diarization, and obfuscation. In: 7th Int. Conf. of CLEF on Experimental IR Meets Multilinguality, Multimodality, and Interaction, CLEF 2016, Springer-Verlag, LNCS(9822), pp 332–350

Sadat F, Kazemi F, Farzindar A (2014) Automatic identification of arabic language varieties and dialects in social media. Proceedings of SocialNLP p 22

Schler J, Koppel M, Argamon S, Pennebaker JW (2006) Effects of age and gender on blogging. In: AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs, AAAI, pp 199–205

Seidman S (2013) Authorship Verification Using the Impostors Method. In: Forner P, Navigli R, Tufis D (eds) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers

Stamatatos E (2009) A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology 60:538–556

Stamatatos E (2011) Plagiarism detection using stopword n-grams. J Am Soc Inf Sci Technol 62(12):2512–2527, DOI 10.1002/asi.21630, URL http://dx.doi.org/10.1002/asi.21630

Stamatatos E, Daelemans W, Verhoeven B, Stein B, Potthast M, Juola P, Sánchez-Pérez MA, Barrón-Cedeño A (2014) Overview of the author identification task at PAN 2014. In: Working Notes for CLEF 2014 Conference, pp 877–897

Stamatatos E, Daelemans W, Verhoeven B, Juola P, López-López A, Potthast M, Stein B (2015a) Overview of the author identification task at PAN 2015. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum

Stamatatos E, Potthast M, Rangel F, Rosso P, Stein B (2015b) Overview of the pan/clef 2015 evaluation lab. In: 6th Int. Conf. of CLEF on Experimental IR meets Multilinguality, Multimodality, and Interaction, CLEF 2015, Springer-Verlag, LNCS(9283), pp 518–538

Stamatatos E, Tschuggnall M, Verhoeven B, Daelemans W, Specht G, Stein B, Potthast M (2016) Clustering by Authorship Within and Across Documents. In: Working Notes Papers of the

CLEF 2016 Evaluation Labs, CLEF and CEUR-WS.org, CEUR Workshop Proceedings, vol 1609, URL http://ceur-ws.org/Vol-1609/

Stein B, Meyer zu Eißen S, Potthast M (2007) Strategies for Retrieving Plagiarized Documents. In: Clarke C, Fuhr N, Kando N, Kraaij W, de Vries A (eds) 30th International ACM Conference on Research and Development in Information Retrieval (SIGIR 07), ACM, New York, pp 825–826, DOI http://dx.doi.org/10.1145/1277741.1277928

Stein B, Lipka N, Prettenhofer P (2011) Intrinsic Plagiarism Analysis. Language Resources and Evaluation (LRE) 45(1):63–82, DOI http://dx.doi.org/10.1007/s10579-010-9115-y

Tschuggnall M, Stamatatos E, Verhoeven B, Daelemans W, Specht G, Stein B, Potthast M (2017) Overview of the Author Identification Task at PAN-2017: Style Breach Detection and Author Clustering. In: Working Notes Papers of the CLEF 2017 Evaluation Labs, CLEF and CEUR-WS.org, CEUR Workshop Proceedings

Weren E, Kauer A, Mizusaki L, Moreira V, de Oliveira P, Wives L (2014) Examining multiple features for author profiling. In: Journal of Information and Data Management, pp 266–279