ISSN 1870-4069

Authorship Verification: A Review of Recent Advances

Efstathios Stamatatos

University of the Aegean, Karlovassi, Greece

stamatatos@aegean.gr

Abstract. Authorship verification attempts to decide whether the author of a given set of texts is also the author of a disputed text. In comparison to closed-set and open-set attribution, the most popular tasks in relevant literature, the verification setting has some important advantages. First, it is more general since any attribution problem can be decomposed into a series of verification cases. Then, certain factors that affect the performance of closed-set and open-set attribution, like the candidate set size and the distribution of training texts over the candidate authors have limited impact on authorship verification. It is, therefore, more feasible to estimate the error rate of authorship attribution technology, needed in the framework of forensic applications, when focusing on the verification setting. Recently, there has been increasing interest for authorship verification, mainly due to the PAN shared tasks organized in 2013, 2014, and 2015. Multiple methods were developed and tested in new benchmark corpora covering several languages and genres. This paper presents a review of recent advances in this field focusing on the evaluation results of PAN shared tasks. Moreover, it discusses successes, failures, and open issues.

Keywords. Authorship analysis, authorship verification, text categorization.

1 Introduction

Authorship attribution is the line of research dealing with the quantification of writing style in texts and revealing the identity of their authors using computational methods [23, 58]. Applications closely related with this area are mainly from the humanities (e.g., revealing the author of novels published anonymously, verifying the authorship of literary works, etc.) [24, 34, 64] and forensics (e.g., discovering authorship links between proclamations of different terrorist groups, resolving copyright disputes, revealing multiple aliases of the same user in social media, verifying the authorship of suicide notes, etc) [1, 36, 65].

Authorship attribution can be seen as a single-label multi-class text categorization task. In each authorship attribution case, a candidate set (i.e., suspects) and samples of their writing are given. Then, the task is to find the most likely candidate based on the similarities of their personal style with the text under investigation. Other factors, like topic and genre of text or sentiment polarity should not affect this procedure. However, this is particularly challenging, since it is not yet possible to extract stylometric measures that are only determined by the personal style of author and are immune to changes in topic or genre [59].

The setting examined in the majority of the published studies refers to *closed-set attribution*, where a well-defined set of suspects is given and one of them is necessarily the true author of the disputed text [11, 47, 50, 53, 56]. This scenario matches the requirements of many forensic cases (i.e., traditionally solved by forensic linguists) where police investigators are able to provide a list of suspects based on the assumption that they are the only ones with access to certain resources, having knowledge of certain facts, etc. An alternative, more robust, setting is *open-set attribution* where it is possible the true author not to be included in the list of likely suspects [31, 54]. This is more appropriate in cases where it is not possible to rule out any likely author (e.g., a post in social media could be written by anyone).

A special case of open-set attribution is *authorship verification* where the candidate set is singleton [33, 35, 45]. In other words, given a set of texts by the same author, the task is to determine whether a text under investigation is by that author or not. This is essentially a one-class classification problem since the negative class is chaotic (i.e., all texts by all other authors).

Until recently, there were limited research studies dealing with authorship verification either exclusively [10, 16, 20, 33, 43] or in parallel with closed-set attribution [38, 62]. The recent influential studies of Koppel [32, 35] highlighting the significance of verification as a fundamental problem in authorship attribution and, mainly, a series of PAN shared tasks organized in 2013, 2014, and 2015 radically increased interest and research teams working in this area [25, 61, 60]. PAN evaluation campaigns provided benchmark corpora covering several natural languages and genres as well as an experimentation and evaluation framework to assess the performance of multiple verification methods. Since 2013, significant progress has been reported and multiple studies improved state-of-the-are methods, provided a better understanding of their strengths and weaknesses [6, 7, 9, 18, 22, 30, 45], and highlighted their applications in humanities and forensics [48, 63, 64]. This paper presents a review of recent advances in this field, focusing on the evaluation results of PAN shared tasks.

In the remaining of this paper, Section 2 discusses the advantages of verification setting over closed-set and multi-class open-set attribution. Then, Section 3 presents an overview of PAN shared tasks in authorship verification. Sections 4 and 5 review recent methods focusing on the stylometric features and the properties of the verification models they use, respectively. Section 6 briefly presents main evaluation results of PAN shared tasks and, finally, Section 7 summarizes main conclusions and discusses open issues.

2 Verification vs. Attribution

Authorship verification is a fundamental problem in authorship attribution since any problem, either a closed-set or open-set case, can be decomposed into a set of verification problems. However, it is quite challenging in comparison to both closed-set and open-set attribution since a verification model should estimate whether the disputed text is *similar enough* with respect to the given texts by a certain author while an attribution model should estimate who the *most similar* candidate author is.

As already mentioned, authorship attribution is associated with significant forensic applications. However, it is questionable whether it can be used as evidence in court. Certainly, this technology can be used by investigators to guide their focus on specific suspects and then collect other admissible evidence (e.g., DNA samples) to be presented in court. In United States federal courts, the *Daubert* standard that determines the admissibility of scientific expert testimony requires the estimation of the error rate of a scientific method [49]. Although it is possible to estimate the error rate of specific forensic methods, like DNA analysis [28], how could the error rate of authorship attribution be determined? Certainly, there are several factors that affect the performance of an attribution model, including the number of candidate authors, the distribution of training texts over the candidate authors, the length of text samples, and whether the texts under investigation match in genre and topic, not to mention factors like style ageing (when the personal style of someone changes over time). This is not unusual in forensic science, since the accuracy of many technologies used to provide forensic evidence is affected by specific factors. For example, fingerprint matching performance deteriorates in the case of latent fingerprint identification [12] while speaker recognition accuracy is affected by the duration of audio samples, the number of samples, cross-channel conditions, voice ageing etc. [8]

The estimation of error rate of authorship attribution technology is certainly more feasible if we adopt the verification setting. An inherent problem in closed-set and open-set attribution is that the performance of the attribution model deteriorates by increasing the size of candidate set [31, 38]. On the other hand, in authorship verification the candidate set is always singleton and therefore the error rate is easier to be estimated. Another crucial issue in closed-set and multi-class open-set attribution is that the performance of attribution models depends on the distribution of training texts over the candidate authors. The so called *class imbalance* problem causes attribution models to prefer majority authors in their predictions [57]. However, in a forensic case, the fact that many text samples are available for a certain candidate author should not make that suspect the most likely author of texts under investigation. Authorship verification is more robust to class imbalance since each candidate author is examined separately.

3 PAN Evaluation Campaigns

PAN is a series of shared tasks in digital text forensics [1]. Since 2009 several authorship analysis tasks have been organized including closed-set and open-set attribution, author profiling, author clustering and author obfuscation. In three consecutive editions of PAN (2013, 2014, and 2015) a shared task in authorship verification was organized and attracted the participation of multiple research teams (18 in 2013, 13 in 2014 and 18 in 2015). PAN organizers built new benchmark corpora covering several languages (English, Dutch, Greek, and Spanish) and genres (essays, novels, reviews, newspaper articles) and provided an online experimentation framework for software submissions and evaluation [15]. [2]

[1] http://pan.webis.de
[2] http://www.tira.io/

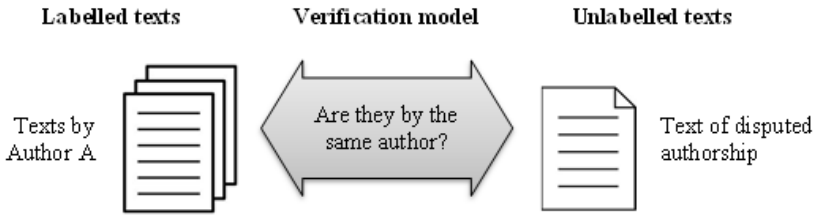

Fig. 1. An authorship verification problem as defined in PAN shared tasks.

The definition of the verification task is as follows: *Given a small set of documents by the same author, is an additional (out-of-set) document also by that author?*. This definition is different than the one adopted by Koppel et al. [35] where they attempt to determine whether two documents are by the same author. The latter can be seen as an unsupervised task, where all documents are unlabelled, in terms of authorship. On the other hand, the PAN definition corresponds to a semi-supervised task where some documents are labelled by authorship (the documents by a certain author). The verification process as it is used in PAN tasks is demonstrated in Figure 1.

Each PAN corpus comprises a set of verification problems and within each problem a set of labelled (or *known*) documents (all by the same author) and exactly one unlabelled (or *unknown*) document are given. PAN participants should provide a binary answer (the unknown document is (not) by the same author) and a score in [0,1] indicating the probability of a positive answer (0 means it is certain that the unknown and known documents are not by the same author and 1 means the opposite). In case the verification method founds a specific problem too hard to solve, it is possible to leave it unanswered by providing a score value of exactly 0.5.[3] Evaluation of submissions is performed based on two measures: the Area Under the ROC curve (AUROC) and $c@1$, that is a modification of accuracy that takes into account the problems left unanswered [44]. The final ranking of participants is provided by the product of AUROC and $c@1$. [4]

An overview of the PAN corpora for authorship verification can be seen in Table 1. [5] The training part of each corpus was given to participants in order to develop and fine-tune their approaches while the evaluation corpus was released after the final deadline of submissions. It is important to notice that each corpus, either in training or evaluation set, is balanced with respect to the distribution of positive and negative verification problems. In other words, the prior probability of a positive (or negative) answer is 0.5. This is a general condition that can be applied to any real authorship verification case where there is no additional evidence that favours positive (or negative) answers.

[3] In PAN 2014 and 2015 editions the binary answers are omitted. Any score value greater than 0.5 corresponds to a positive answer and any score lower than 0.5 corresponds to a negative answer.

[4] In PAN 2013 two separate rankings were produced, one based on AUROC and another based on F_1 score

[5] All corpora can be downloaded from http://pan.webis.de/

Table 1. Authorship verification corpora used in PAN shared tasks.

	Corpus	Training problems	Evaluation problems	Mean labelled texts / problem	Mean text length (words)
PAN 2013	English (Textbooks)	10	30	3.98	1058
	Greek (Articles)	20	30	5.16	1823
	Spanish (Editorials+Fiction)	5	25	3.07	849
PAN 2014	Dutch (Essays)	96	96	1.89	405
	Dutch (Reviews)	100	100	1.02	114
	English (Essays)	200	200	2.62	841
	English (Novels)	100	200	1.00	5115
	Greek (Articles)	100	100	2.77	1470
	Spanish (Articles)	100	100	5.00	1129
PAN 2015	Dutch (Cross-genre)	100	165	1.75	357
	English (Cross-topic)	100	500	1.00	508
	Greek (Cross-topic)	100	100	2.87	717
	Spanish (Mixed)	100	100	4.00	950

However, when such evidence exists, this parameter should be taken into account in the evaluation process.

In PAN 2013 and PAN 2014 tasks, all documents within a verification problem are in the same language, belong to the same genre, and there are thematic similarities. This means that the disputed text and the known texts have certain similarities. In PAN 2015 the only valid assumption is that all documents within a problem are in the same language. The disputed and the known documents may belong to different genres and their themes can be quite distant. This makes the latter edition of PAN very challenging since it is well known that genre and topic affect stylometric measures considerably. On the other hand, the assumption that all texts should have thematic similarities and belong to the same genre is not realistic since in many forensic cases this is certainly not possible (e.g., imagine the case of verifying the authenticity of a suicide note).

In general, the contribution of PAN shared tasks to the progress of authorship verification research is undoubted. PAN attracted the attention of multiple research teams in this task and provided benchmark corpora that became the standard in this field. Moreover, alternative verification methods were systematically compared and the state-of-the-art performance was estimated. It is also important that based on the fact that PAN required software submissions, a library of verification models is now available and can be used in future evaluations on new corpora as well as in the framework of new tasks, for example *author obfuscation* [46].

On the other hand, there are certain weaknesses of PAN tasks. The volume of some of the provided benchmarks is limited (e.g., the Spanish part of the PAN 2013 corpus). Evaluation results and associated conclusions are corpus-specific due to the lack of homogeneity in corpora properties (i.e., number of problems, known documents per problem, words per document). In addition, the quality of some submissions is questionable since they might be based on naive methods and hasty software imple-

mentations while some of the notebook papers do not provide a detailed description of their approach.

4 Stylometric Analysis

In authorship attribution research there is a wide variety of measures that attempt to capture nuances of the personal style of the authors [58]. Most of the measures proposed in the relevant literature correspond to lexical features, e.g., function word frequencies, word *n*-grams, word-length distribution, vocabulary richness measures, etc. Another effective type of features operates on the character level, e.g., character *n*-grams, punctuation mark frequencies, etc. Such features are language-independent and capture intra-word and inter-word information. More sophisticated measures require the analysis of texts by natural language processing tools and then syntactic (e.g., part-of-speech frequencies, rewrite rule frequencies, syntactic *n*-grams, etc.) or semantic features (e.g., semantic dependencies, use of synonyms, etc.) can be extracted. These higher-level features are usually noisy due to errors performed by NLP tools but usually they are useful complement of other, lower-level and more powerful features. Finally, in case all texts share some properties, for instance, they belong to the same genre (e.g., e-mails), they are about a certain thematic area (e.g., computer sales), they are in a specific format (e.g., html), it is possible and very effective to define application-specific measures for that particular domain.

An early study in authorship verification showed that sophisticated syntactic features can enhance the performance of simple lexical features, however, the gain in performance is not significant [16]. Most of the verification approaches submitted to PAN are based on low-level (lexical and character) features and avoid the use of syntactic, semantic, or application-specific features. One reason for that is that measures like character *n*-grams or very frequent word frequencies can practically be applied to any natural language with minimal requirements for text pre-processing. The use of NLP tools by PAN participants was limited to POS tagging and full syntactic parsing. This language-dependent analysis sometimes could not be applied to all languages covered by PAN corpora [60, 61]. Certainly, existing NLP tools, most probably not specifically trained for the texts under investigation, are expected to provide quite noisy stylometric measures.

Another common practice is to combine several features in an attempt to compromise for the weaknesses of a specific feature type. It is also remarkable that in some cases the proposed methods had to select the most suitable feature set for a given collection of verification problems [55]. That way, the features that seem to be more effective (using the training corpus) for a particular language, genre, or topic are selected.

5 Verification Models

The verification model decides whether the disputed text and the known texts are by the same author based on the degree of similarity in terms of the stylometric representation of texts. There are two main categories of verification models:

- *Intrinsic models*: They provide a decision based only on the analysis of the texts found in a given verification problem (the provided known and unknown texts of a certain problem). They exclude the use of external texts by other authors. Therefore, intrinsic models handle the authorship verification problem as a one-class classification case avoiding the use of any external, either labelled or unlabelled, data. Typical examples of intrinsic models are described by Jankowska et al. [22], Potha and Stamatatos [45], Layton [37], Halvani et al. [18], and Bartoli et al. [4]. Intrinsic models are usually faster since they are limited in the analysis of the known and unknown texts. Moreover, they are more robust since their performance does not depend on external factors.
- *Extrinsic models*: They use external documents by other authors to estimate if the similarity of the disputed texts with the known texts is significant enough. Such models actually transform the authorship verification problem to a binary classification case where the known texts form the positive class and the external documents form the negative class. Typical examples of extrinsic models are described by Koppel et al. [35], Seidman [55], Bagnall [2], Veenman & Li [66], and Kocher & Savoy [30]. Extrinsic models are usually more effective than intrinsic ones since a binary classification problem is easier to handle in comparison to a one-class problem. One crucial issue with this kind of methods is the use of an appropriate set of external documents. It is very important to use external documents that belong to the same genre with the ones under investigation [35].

From another point of view, the verification models can be distinguished by the type of learning they use.

- *Eager learning models*: They attempt to extract a general model of authorship verification based on the training corpus. Each verification problem is seen as an instance, either positive (when the known and unknown texts are by the same authors) or negative (in the opposite case). The set of instances of the training corpus is used to train a binary classifier which can then be used to guess the most likely class of any given verification case. Typical examples of this category are described by Frery et al. [13], Bartoli, et al. [4], Pacheco et al. [42], Hürlimann et al. [19], and Brocardo et al. [7]. Such models are effective only when the training corpus is representative of the verification cases that we are going to solve. Their effectiveness and complexity depend on the size and characteristics of the training corpus. Moreover, they can take advantage of powerful supervised learning algorithms, like SVM, neural networks, etc. and they are usually very fast in application phase.
- *Lazy learning models*: They handle every verification case separately. During the training phase they are practically resting. Once a verification case is available in the application phase, they perform all necessary kinds of analysis to estimate their answer. Typical examples of lazy learning models are described by Koppel et al. [35], Khonji and Iraqi [27], Bagnall [2], Jankowska et al. [22], Potha and Stamatatos [45], and Halvani et al. [18]. Such models require higher time cost in the application phase in comparison to eager learning models. However, a big strength is that they do not depend too much on the properties of the training corpus.

Finally, another distinguishing characteristic of verification models refers to the way they handle the labelled examples (known documents by the same author) within each verification problem.

- *Profile-based models*: They concatenate all known documents and then compare the concatenated text with the disputed text. Essentially, they attempt to capture the stylistic properties of the author by discarding any differences between the provided texts. Typical examples of profile-based models are described by Potha and Stamatatos [45], Kocher and Savoy [29], Pacheco et al. [42], Halvani et al. [30], and Kocher & Savoy [30]. A significant strength of such methods is that when text length is limited, by concatenating all available labelled texts they increase the robustness of stylometric representation. On the other hand, concatenated text may have a quite distant representation with respect to its constituent texts especially when the topic and genre of these texts do not match.
- *Instance-based models*: They handle each labelled text separately and compare it with the disputed text. Such models consider each text as a separate instance of author's style. When multiple labelled texts are available, they combine the answers to provide the final decision. Typical examples of this category are described by Seidman [55], Jankowska et al. [22], Moreau et al. [41], Brocardo et al. [7], and Castro-Castro et al. [9]. Another variation is to first concatenate all labelled texts and then split the resulting text into samples of equal size [5, 17]. Instance-based models are better able to exploit significant differences among labelled texts given that they can effectively handle the set of answers (one for each labelled text). On the other hand, they are affected by text length limitations.

It is also notable that some approaches attempt to combine profile-based and instance-based paradigms by first analysing each labelled text separately and then combining the extracted representations of all labelled texts [26, 51]. Such *hybrid* methods practically fail to combine the strengths of the two paradigms.

Table 2 shows the distribution of PAN participants over the types of verification models as defined above. [6] It is clear that the majority of PAN participants follow an intrinsic, lazy, and instance-based methodology. Eager learning method began to be popular in late editions of PAN when the size of the training corpus allowed the development of relatively effective models [13]. Moreover, extrinsic models gain popularity over the years based on the excellent results achieved by Seidman [55], Khonji and Iraqi [27], and Bagnall [2].

6 PAN Results

Analytical evaluation results of PAN participants in benchmark corpora including tests of statistical significance are provided in [25, 61, 60]. Table 3 shows the best results achieved by PAN participants for each corpus and the average performance of all PAN

[6] PAN shared tasks in authorship verification received 18 submissions in 2013, 13 submissions in 2014 and 18 submissions in 2015. All but two (in 2013) research teams also submitted a notebook describing their method.

Table 2. Distribution of PAN participants over the verification model categories (defined in Section 5).

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

participants. It is clear that factors like language and genre do not affect the performance of verification models significantly. For instance, the results of Dutch essays are very high while the performance on another corpus in the same language (Dutch reviews) or in the same genre (English essays) are relatively low. Other factors, like the number of labelled texts per verification problem or text length (see Table 1) are certainly significant. In general, when there is a low number of labelled texts (1 or 2) of limited text length (less than 500 words), the performance of verification models worsens.

On the other hand, it is not always possible to explain high or low performance of verification models on a specific corpus based on the quantitative properties of corpus exclusively. There are other qualitative properties that are more useful. For instance, the English novels corpus consists of parts of novels on a specific subgenre of horror fiction that is characterized by an unusual vocabulary and extremely florid prose. This makes similarities between different authors to seem more significant than in normal prose. Verification results for that corpus are poor despite the relatively high text length of its documents.

It should be underlined that the performance of verification models is not heavily affected when the texts within a verification model do not match in genre and thematic area, as it happens in PAN 2015 corpora. Although the average performance of PAN participants is relatively low for those corpora, there were certain submissions capable of reaching impressively high results in these challenging cases [2, 4, 41].

A summary of characteristics of the best-performing systems in the 3 editions of PAN can be seen in Table 4. All submissions that achieved the best performance result (either $c@1$ or AUROC) in any of the PAN evaluation corpora are presented. For each submission the properties of its verification model as well as its requirements for elaborate analysis to extract stylometric features are described. As can be seen, most of the best-performing models use only low-level stylometric measures, like character and word *n*-grams. Only a few methods require more sophisticated analysis like POS tagging, or topic modeling (e.g., LSA, LDA). With respect to the verification model properties, extrinsic models, although a minority in PAN participants (see Table 2), are well represented in best-performing submissions and, actually, all three PAN overall winner submissions for 2013, 2014, and 2015 shared tasks belong to this category [2,

Table 3. Evaluation results (best and average performance in terms of $c@1$ and AUROC) of PAN participants on authorship verification corpora.

		c@1		AUROC	
	Corpus	**Best**	**Average**	**Best**	**Average**
PAN 2013	English (Textbooks)	0.80	0.66	0.84	0.61
	Greek (Articles)	0.83	0.52	0.82	0.60
	Spanish (Editorials+Fiction)	0.84	0.59	0.93	0.67
PAN 2014	Dutch (Essays)	0.91	0.75	0.93	0.76
	Dutch (Reviews)	0.69	0.55	0.76	0.59
	English (Essays)	0.71	0.58	0.72	0.60
	English (Novels)	0.72	0.57	0.75	0.61
	Greek (Articles)	0.81	0.60	0.89	0.67
	Spanish (Articles)	0.78	0.68	0.90	0.71
PAN 2015	Dutch (Cross-genre)	0.77	0.55	0.83	0.60
	English (Cross-topic)	0.76	0.56	0.81	0.62
	Greek (Cross-topic)	0.85	0.54	0.89	0.67
	Spanish (Mixed)	0.83	0.59	0.93	0.66
	Average	0.79	0.60	0.85	0.64

27, 55]. It is also notable that none of the best-performing methods adopts the profile-based paradigm.

An important conclusion extracted from PAN shared tasks is that it is possible to combine different verification models and provide a robust approach with enhanced performance. PAN organizers report the results of a heterogeneous ensemble that combines that answers of all participants (by averaging the scores in each verification problem) and in many cases the performance of this ensemble is better than or competitive with the best-performing PAN participant [25, 61, 60]. Figures 2, 3, and 4 depict illustrative examples for three PAN corpora: the English essays corpus and the Spanish articles corpus from PAN 2014 as well as the Greek articles corpus from PAN 2015, respectively. In more detail, ROC curves on the evaluation parts of these corpora are shown for two methods: (i) the best-performing PAN model for that particular corpus and (ii) the ensemble combining answers by all PAN participants. As can be seen, in the case of English essays, the performance of the ensemble is better than the best individual participant in almost the whole ROC space. Concerning the Spanish articles corpus, the picture is more complicated since both the best PAN participant and the ensemble are competitive and each one of them is the best choice in a certain area of ROC space. When false positives have higher cost the best PAN participant is more effective while when the false negatives are more important the ensemble is a better choice. Finally, when examining the Greek articles corpus, the best PAN participant clearly outperforms the ensemble except in the case the cost of false negatives is extremely high.

In general, the performance of the ensemble in PAN 2015 corpora was lower in comparison to PAN 2014 corpora [61, 60]. This can be partially explained by the consid-

Table 4. Brief description of best-performing models in PAN shared tasks.

PAN participant	Verification model	Elaborate stylometric analysis
Bagnall et al. 2015 [2]	extrinsic, lazy, instance-based	none
Bartoli et al. 2015 [4]	intrinsic, eager, instance-based	POS tagging
Frery et al. 2014 [13]	intrinsic, eager, instance-based	none
Ghaeini et al. 2013 [14]	intrinsic, lazy, instance-based	POS tagging
Halvani 2013 [17]	intrinsic, lazy, instance-based	none
Jankowska et al. 2013 [21]	intrinsic, lazy, instance-based	none
Khonji & Iraqi 2014 [27]	extrinsic, lazy, instance-based	none
Mayor et al. 2014 [39]	extrinsic, lazy, instance-based	none
Modaresi & Gross 2014 [40]	intrinsic, eager, instance-based	none
Moreau et al. 2015 [41]	extrinsic, eager, instance-based	LDA, POS tagging
Satyam et al. 2014 [52]	intrinsic, lazy, instance-based	LSA
Seidman 2013 [55]	extrinsic, lazy, instance-based	none
Veenman & Li 2013 [66]	extrinsic, lazy, instance-based	none

erably low performance scores of several participants in PAN 2015 corpora. Certainly, more sophisticated models for combining different methods can provide better results. So far, there is limited research regarding the optimal way to combine heterogeneous verification models [41].

7 Discussion

Since 2013 there has been a significant progress in authorship verification research mainly due to PAN evaluation campaigns that focused on this task. There are multiple research teams around the world that conduct research in this area and multiple methods and variations of them are nowadays available. Based on benchmark corpora in several languages and genres, produced in the framework of PAN shared tasks, systematic evaluation of proposed methods has been performed.

Certainly, there are several factors that affect the performance of verification methods as can be seen in the results of Table 3 in combination with the properties of each corpus (Table 1). However, these factors are less than the ones that are considered in both closed-set and multi-class open-set attribution, since the candidate set size and the distribution of texts over the candidate authors have limited effect in authorship verification.

The average performance of best systems in all PAN corpora (see Table 3) indicates that the error rate of state-of-the-art methods in authorship verification is around 20%. [7] Although this is too high in comparison to the most effective technologies used to provide forensic evidence (e.g., the error rate of DNA analysis is less than 1% [28]), it is comparable to other technologies that analyse noisy data, like latent fingerprint matching [12] or speaker identification [8]. The relatively higher AUROC scores indicate that the verification models are able to rank answers more effectively and they

[7] An average $c@1$ value of 0.79 indicates an accuracy of around 80%.

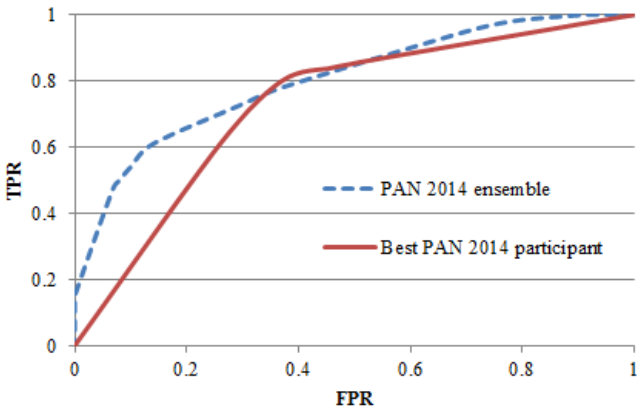

Fig. 2. ROC curves on the evaluation corpus of English essays in PAN 2014 of the best-performing participant for that particular corpus [13] and the ensemble of all PAN 2014 participants.

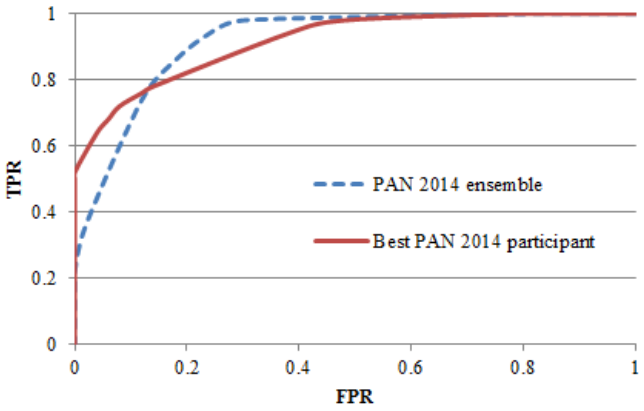

Fig. 3. ROC curves on the evaluation corpus of Spanish articles in PAN 2014 of the best-performing participant for that particular corpus [27] and the ensemble of all PAN 2014 participants.

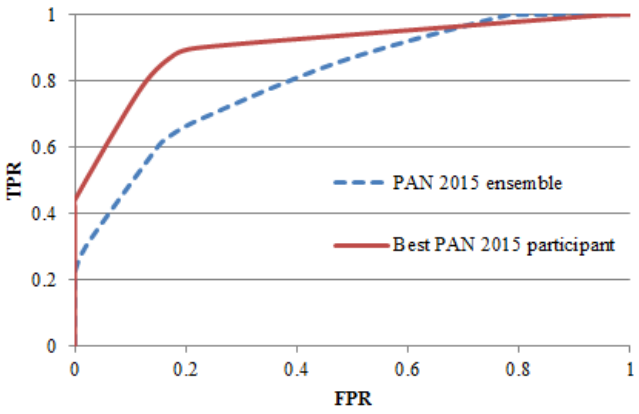

Fig. 4. ROC curves on the evaluation part of Greek cross-topic corpus in PAN 2015 of the best-performing participant for that particular corpus [2] and the ensemble of all PAN 2015 participants.

need to be further improved in transforming this ranking to binary answers. Taking into account the prior probability of positive answers is crucial in this direction [2]. All PAN corpora consider an equal prior probability for positive and negative instances. However, this assumption is not true in all application scenarios and more extensive evaluation experiments are needed by controlling this parameter.

The most effective models in PAN evaluation campaigns follow the extrinsic approach where the verification problem is transformed to a binary classification task by considering external documents by other authors [2, 27, 55]. An inherent problem of such methods is the appropriate selection of external documents for a given verification case. So far, existing approaches do not use sophisticated methods to select external documents and they just use texts from the training corpus or texts found my search engines based on queries extracted from specific seed documents [55, 66]. Considerable improvement can be expected if the most suitable set of external documents is found for a given verification problem [35].

Another important conclusion is that verification methods that apply eager learning can be very effective when the training corpus is large enough and comprises similar cases with the evaluation problems [13, 41]. On the other hand, if the training corpus is not representative of the difficulties found in evaluation set, eager learning models fail [60]. In practice, this means that if we want to apply such models in forensic applications, for any given verification problem, we should prepare an appropriate training corpus with cases of similar characteristics. In case there are certain suspects and labelled texts by them, it is possible to build such a corpus. It remains to be seen whether general-purpose corpora covering specific languages and genres can be useful in this respect.

PAN evaluation campaigns demonstrated that combining heterogeneous verification models is a very effective choice [25, 61]. Heterogeneous ensemble achieve consistently

high performance in most corpora. Challenging cases where the genre or topic of texts within a verification problem do not match can be handled by more sophisticated ensemble models [41] that select the most appropriate models for each verification problem separately. The existence of a library of verification methods makes this research direction very promising.

Authorship verification tasks at PAN provided the necessary background to explore other relevant tasks. Based on the implementations of verification methods submitted to PAN shared tasks, another task focusing on author obfuscation (i.e., attempting to modify the style of a document so that a verification method does not recognize its author) was recently organized [46]. In another recent PAN shared task in author clustering (grouping documents by authorship) a variation of an authorship verification model was the best-performing participant [3]. All these indicate that verification is a fundamental task in authorship attribution and if we are able to deal with verification effectively it is possible to solve practically any case. There is a lot of room for improvement towards this direction.

References

1. Abbasi, A., Chen, H.: Applying authorship analysis to extremist-group web forum messages. Intelligent Systems, IEEE 20(5), 67–75 (2005)
2. Bagnall, D.: Author Identification using multi-headed Recurrent Neural Networks. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
3. Bagnall, D.: 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 (2016)
4. Bartoli, A., Dagri, A., Lorenzo, A.D., Medvet, E., Tarlao, F.: An Author Verification Approach Based on Differential Features. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
5. Bobicev, V.: Authorship Detection with PPM. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
6. Boukhaled, M.A., Ganascia, J.G.: Probabilistic anomaly detection method for authorship verification. In: Besacier, L., Dediu, A.H., Martín-Vide, C. (eds.) Proceedings of Statistical Language and Speech Processing: Second International Conference. pp. 211–219. Springer International Publishing (2014)
7. Brocardo, M.L., Traore, I., Woungang, I.: Authorship verification of e-mail and tweet messages applied for continuous authentication. J. Comput. Syst. Sci. 81(8), 1429–1440 (2015)
8. Campbell, J.P., Shen, W., Campbell, W.M., Schwartz, R., Bonastre, J.F., Matrouf, D.: Forensic speaker recognition. IEEE Signal Processing Magazine 26(2), 95–103 (2009)
9. Castro-Castro, D., Arcia, Y.A., Brioso, M.P., Guillena, R.M.: Authorship verification, average similarity analysis. In: Recent Advances in Natural Language Processing. pp. 84–90 (2015)
10. Escalante, H.J., y Gómez, M.M., Pineda, L.V.: Particle swarm model selection for authorship verification. In: Proceedingsof the 14th Iberoamerican Conference on Pattern Recognition. pp. 563–570 (2009)
11. Escalante, H.J., Solorio, T., Montes-y-Gómez, M.: Local histograms of character n-grams for authorship attribution. In: Proc. of the 49th Annual Meeting of the Association for Computational Linguistics. pp. 288–298 (2011)

12. Feng, J., Jain, A.K., Nandakumar, K.: Fingerprint matching. Computer 43, 36–44 (2010)
13. Fréry, J., Largeron, C., Juganaru-Mathieu, M.: UJM at clef in author identification. In: CLEF 2014 Labs and Workshops, Notebook Papers. CLEF and CEUR-WS.org (2014)
14. Ghaeini, M.: Intrinsic Author Identification Using ModifiedWeighted KNN. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
15. Gollub, T., Potthast, M., Beyer, A., Busse, M., Pardo, F.M.R., Rosso, P., Stamatatos, E., Stein, B.: Recent trends in digital text forensics and its evaluation - plagiarism detection, author identification, and author profiling. In: Proceedings of the 4th International Conference of the CLEF Initiative. pp. 282–302 (2013)
16. van Halteren, H.: Linguistic profiling for author recognition and verification. In: Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics. ACL (2004)
17. Halvani, O., Steinebach, M., Zimmermann, R.: Authorship Verification via k-Nearest Neighbor Estimation. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
18. Halvani, O., Winter, C., Pflug, A.: Authorship verification for different languages, genres and topics. Digital Investigation 16, S33 – S43 (2016)
19. Hürlimann, M., Weck, B., van den Berg, E., Šuster, S., Nissim, M.: GLAD: Groningen Lightweight Authorship Detection. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2015)
20. Iqbal, F., Khan, L.A., Fung, B.C.M., Debbabi, M.: e-mail authorship verification for forensic investigation. In: Proceedings of the 2010 ACM Symposium on Applied Computing. pp. 1591–1598. ACM (2010)
21. Jankowska, M., Keselj, V., Milios, E.: Proximity based one-class classification with Common N-Gram dissimilarity for authorship verification task. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
22. Jankowska, M., Milios, E.E., Keselj, V.: Author verification using common n-gram profiles of text documents. In: Proceedings of COLING, 25th International Conference on Computational Linguistics. pp. 387–397 (2014)
23. Juola, P.: Authorship Attribution. Foundations and Trends in Information Retrieval 1, 234–334 (2008)
24. Juola, P.: How a computer program helped reveal J. K. Rowling as author of A Cuckoo's Calling. Scientific American (2013)
25. Juola, P., Stamatatos, E.: Overview of the author identification task at PAN 2013. In: Working Notes for CLEF 2013 Conference (2013)
26. Kern, R.: Grammar Checker Features for Author Identification and Author Profiling. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
27. Khonji, M., Iraqi, Y.: 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 (2014)
28. Kloosterman, A., Sjerps, M., Quak, A.: Error rates in forensic DNA analysis: Definition, numbers, impact and communication. Forensic Science International: Genetics 12, 77 – 85 (2014)
29. Kocher, M., Savoy, J.: UniNE at CLEF 2015: Author Identification. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
30. Kocher, M., Savoy, J.: A simple and efficient algorithm for authorship verification. Journal of the Association for Information Science and Technology (2016)

31. Koppel, M., Schler, J., Argamon, S.: Authorship Attribution in the Wild. Language Resources and Evaluation 45, 83–94 (2011)
32. Koppel, M., Schler, J., Argamon, S., Winter, Y.: The fundamental problem of authorship attribution. English Studies 93(3), 284–291 (2012)
33. Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research 8, 1261–1276 (2007)
34. Koppel, M., Seidman, S.: Automatically identifying pseudepigraphic texts. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 1449–1454 (2013)
35. 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)
36. Lambers, M., Veenman, C.: Forensic authorship attribution using compression distances to prototypes. In: Geradts, Z., Franke, K., Veenman, C. (eds.) Computational Forensics, Lecture Notes in Computer Science, vol. 5718, pp. 13–24. Springer Berlin Heidelberg (2009)
37. Layton, R.: A simple Local n-gram Ensemble for Authorship Verification. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2014)
38. Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. In: Proceedings of the Twenty-Second International Conference on Computational Linguistics (COLING 2008). pp. 513–520 (2008)
39. Mayor, C., Gutierrez, J., Toledo, A., Martinez, R., Ledesma, P., Fuentes, G., , Meza, I.: A Single Author Style Representation for the Author Verification Task. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2014)
40. Modaresi, P., Gross, P.: A Language Independent Author Verifier Using Fuzzy C-Means Clustering. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2014)
41. Moreau, E., Jayapal, A., Lynch, G., Vogel, C.: Author Verification: Basic Stacked Generalization Applied To Predictions from a Set of Heterogeneous Learners. In: Cappellato, L., Ferro, N., Gareth, J., San Juan, E. (eds.) Working Notes Papers of the CLEF 2015 Evaluation Labs (2015)
42. Pacheco, M., Fernandes, K., Porco, A.: Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2015)
43. Pavelec, D., Oliveira, L.S., Justino, E., Batista, L.V.: Using conjunctions and adverbs for author verification. Journal of Universal Computer Science 14(18), 2967–2981 (2008)
44. Peñas, A., Rodrigo, A.: A simple measure to assess non-response. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 1415–1424. ACL (2011)
45. Potha, N., Stamatatos, E.: A profile-based method for authorship verification. In: Artificial Intelligence: Methods and Applications - Proceedings of the 8th Hellenic Conference on AI, SETN. pp. 313–326 (2014)
46. Potthast, M., Hagen, M., Stein, B.: Author Obfuscation: Attacking the State of the Art in Authorship Verification. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CLEF and CEUR-WS.org (2016)
47. Qian, T., Liu, B., Chen, L., Peng, Z.: Tri-training for authorship attribution with limited training data. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL. pp. 345–351 (2014)

48. Roffo, G., Cristani, M., Bazzani, L., Minh, H.Q., Murino, V.: Trusting skype: Learning the way people chat for fast user recognition and verification. In: Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on. pp. 748–754 (2013)
49. Saks, M.J., Koehler, J.J.: The coming paradigm shift in forensic identification science. Science 309(5736), 892–895 (2005)
50. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL. pp. 93–102 (2015)
51. Sari, Y., Stevenson, M.: A Machine Learning-based Intrinsic Method for Cross-topic and Cross-genre Authorship Verification. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2015)
52. Satyam, Anand, Dawn, A., , Saha, S.: Statistical Analysis Approach to Author Identification Using Latent Semantic Analysis. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2014)
53. Savoy, J.: Authorship attribution based on a probabilistic topic model. Information Processing and Management 49(1), 341–354 (2013)
54. Schaalje, G.B., Blades, N.J., Funai, T.: An open-set size-adjusted bayesian classifier for authorship attribution. Journal of the American Society for Information Science and Technology 64(9), 1815–1825 (2013)
55. Seidman, S.: Authorship Verification Using the Impostors Method. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
56. Seroussi, Y., Zukerman, I., Bohnert, F.: Authorship attribution with topic models. Computational Linguistics 40(2), 269–310 (2014)
57. Stamatatos, E.: Author identification: Using text sampling to handle the class imbalance problem. Information Processing and Management 44(2), 790 – 799 (2008)
58. Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology 60, 538–556 (2009)
59. Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy 21, 421–439 (2013)
60. 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 of CLEF 2015 - Conference and Labs of the Evaluation forum (2015)
61. Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., Sánchez-Pérez, M.A., Barrón-Cedeño, A.: Overview of the author identification task at PAN 2014. In: Working Notes for CLEF 2014 Conference. pp. 877–897 (2014)
62. Stamatatos, E., Fakotakis, N., Kokkinakis, G.: Automatic text categorization in terms of genre and author. Computational Linguistics 26(4), 471–495 (2000)
63. Stolerman, A.: Authorship Verification. Ph.D. thesis, Drexel University (2015)
64. Stover, J.A., Winter, Y., Koppel, M., Kestemont, M.: Computational authorship verification method attributes a new work to a major 2nd century african author. Journal of the American Society for Information Science and Technology 67(1), 239–242 (2016)
65. Sun, J., Yang, Z., Liu, S., Wang, P.: Applying stylometric analysis techniques to counter anonymity in cyberspace. Journal of Networks 7(2), 259–266 (2012)
66. Veenman, C., Li, Z.: Authorship Verification with Compression Features. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)