Assessing the Applicability of Authorship Verification Methods

Oren Halvani*
Christian Winter
Lukas Graner
Fraunhofer Institute for Secure Information Technology SIT, Darmstadt, Germany
<FirstName>.<LastName>@sit.fraunhofer.de

Abstract

Authorship verification (AV) is a research subject in the field of digital text forensics that concerns itself with the question, whether two documents have been written by the same person. During the past two decades, an increasing number of proposed AV approaches can be observed. However, a closer look at the respective studies reveals that the underlying characteristics of these methods are rarely addressed, which raises doubts regarding their applicability in real forensic settings. The objective of this paper is to fill this gap by proposing clear criteria and properties that aim to improve the characterization of existing and future AV approaches. Based on these properties, we conduct three experiments using 12 existing AV approaches, including the current state of the art. The examined methods were trained, optimized and evaluated on three self-compiled corpora, where each corpus focuses on a different aspect of applicability. Our results indicate that part of the methods are able to cope with very challenging verification cases such as 250 characters long informal chat conversations (72.7% accuracy) or cases in which two scientific documents were written at different times with an average difference of 15.6 years (> 75% accuracy). However, we also identified that all involved methods are prone to cross-topic verification cases.

CCS Concepts

• **Applied computing → Computer forensics**; • **Computing methodologies → Machine learning**.

ACM Reference Format:
Oren Halvani, Christian Winter, and Lukas Graner. 2019. Assessing the Applicability of Authorship Verification Methods. In *Proceedings of the 14th International Conference on Availability, Reliability and Security (ARES 2019) (ARES '19), August 26–29, 2019, Canterbury, United Kingdom.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3339252.3340508

1 Introduction

Digital text forensics aims at examining the originality and credibility of information in electronic documents and, in this regard, at extracting and analyzing information about the authors of the respective texts [30]. Among the most important tasks of this field are **authorship attribution** (AA) and **authorship verification** (AV), where the former deals with the problem of identifying the most likely author of a document $\mathcal{D}_\mathcal{U}$ with unknown authorship, given a set of texts of candidate authors. AV, on the other hand, focuses on the question whether $\mathcal{D}_\mathcal{U}$ was in fact written by a known author $\mathcal{A}$, where only a set of reference texts $\mathbb{D}_\mathcal{A}$ of this author is given. Both disciplines are strongly related to each other, as any AA problem can be broken down into a series of AV problems [27]. Breaking down an AA problem into multiple AV problems is especially important in such scenarios, where the presence of the true author of $\mathcal{D}_\mathcal{U}$ in the candidate set cannot be guaranteed.

In the past two decades, researchers from different fields including linguistics, psychology, computer science and mathematics proposed numerous techniques and concepts that aim to solve the AV task. Probably due to the interdisciplinary nature of this research field, AV approaches were becoming more and more diverse, as can be seen in the respective literature. In 2013, for example, Veenman and Li [42] presented an AV method based on compression, which has its roots in the field of information theory. In 2015, Bagnall [2] introduced the first deep learning approach that makes use of language modeling, an important key concept in statistical natural language processing. In 2017, Castañeda and Calvo [13] proposed an AV method that applies a semantic space model through *Latent Dirichlet Allocation*, a generative statistical model used in information retrieval and computational linguistics.

Despite the increasing number of AV approaches, a closer look at the respective studies reveals that only minor attention is paid to their underlying characteristics such as reliability and robustness. These, however, must be taken into account before AV methods can be applied in real forensic settings. The objective of this paper is to fill this gap and to propose important properties and criteria that are not only intended to characterize AV methods, but also allow their assessment in a more systematic manner. By this, we hope to contribute to the further development of this young[1] research field. Based on the proposed properties, we investigate the applicability of 12 existing AV approaches on three self-compiled corpora, where each corpus involves a specific challenge.

The rest of this paper is structured as follows. Section 2 discusses the related work that served as an inspiration for our analysis. Section 3 comprises the proposed criteria and properties to characterize AV methods. Section 4 describes the methodology, consisting of the used corpora, examined AV methods, selected performance measures and experiments. Finally, Section 5 concludes the work and outlines future work.

*Corresponding author.

ARES '19, August 26–29, 2019, Canterbury, United Kingdom

[1] According to the literature [38], Stamatatos et al. were the first researchers who discussed AV in the context of natural language texts in 2000 [39]. AV, therefore, can be seen as a young field in contrast to AA, which dates back to the 19th century [14].

2 Related Work

Over the years, researchers in the field of authorship analysis identified a number of challenges and limitations regarding existing studies and approaches. Azarbonyad et al. [1], for example, focused on the questions if the writing styles of authors of short texts change over time and how this affects AA. To answer these questions, the authors proposed an AA approach based on time-aware language models that incorporate the temporal changes of the writing style of authors. In one of our experiments, we focus on a similar question, namely, whether it is possible to recognize the writing style of authors, despite of large time spans between their documents. However, there are several differences between our experiment and the study of Azarbonyad et al. First, the authors consider an AA task, where one anonymous document $\mathcal{D}_{\mathcal{U}}$ has to be attributed to one of n possible candidate authors, while we focus on an AV task, where $\mathcal{D}_{\mathcal{U}}$ is compared against one document $\mathcal{D}_{\mathcal{A}}$ of a known author. Second, the authors focus on texts with informal language (emails and tweets) in their study, while in our experiment we consider documents written in a formal language (scientific works). Third, Azarbonyad et al. analyzed texts with a time span of four years, while in our experiment the average time span is 15.6 years. Fourth, in contrast to the approach of the authors, none of the 12 examined AV approaches in our experiment considers a special handling of temporal stylistic changes.

In recent years, the new research field **author obfuscation** (AO) evolved, which concerns itself with the task to fool AA or AV methods in a way that the true author cannot be correctly recognized anymore. To achieve this, AO approaches which, according to Gröndahl and Asokan [7] can be divided into manual, computer-assisted and automatic types, perform a variety of modifications on the texts. These include simple synonym replacements, rule-based substitutions or word order permutations. In 2016, Potthast et al. [29] presented the first large-scale evaluation of three AO approaches that aim to attack 44 AV methods, which were submitted to the PAN-AV competitions during 2013-2015 [19, 37, 38]. One of their findings was that even basic AO approaches have a significant impact on many AV methods. More precisely, the best-performing AO approach was able to flip on average $\approx 47\%$ of an authorship verifier's decisions towards choosing N ("different author"), while in fact Y ("same author") was correct [29]. In contrast to Potthast et al., we do not focus on AO to measure the robustness of AV methods. Instead, we investigate in one experiment the question how trained AV models behave, if the lengths of the questioned documents are getting shorter and shorter. To our best knowledge, this question has not been addressed in previous authorship verification studies.

3 Characteristics of Authorship Verification

Before we can assess the applicability of AV methods, it is important to understand their fundamental characteristics. Due to the increasing number of proposed AV approaches in the last two decades, the need arose to develop a systematization including the conception, implementation and evaluation of authorship verification methods. In regard to this, only a few attempts have been made so far. In 2004, for example, Koppel and Schler [22] described for the first time the connection between AV and unary classification, also known as **one-class classification**. In 2008, Stein et al. [40] compiled an overview of important algorithmic building blocks for AV where, among other things, they also formulated three AV problems as decision problems. In 2009, Stamatatos [34] coined the phrases *profile-* and *instance-based* approaches that initially were used in the field of AA, but later found their way also into AV. In 2013 and 2014, Stamatatos et al. [19, 26] introduced the terms *intrinsic-* and *extrinsic* models that aim to further distinguish between AV methods. However, a closer look at previous attempts to characterize authorship verification approaches reveals a number of misunderstandings, for instance, when it comes to draw the borders between their underlying classification models. In the following subsections, we clarify these misunderstandings, where we redefine previous definitions and propose new properties that enable a better comparison between AV methods.

3.1 Reliability (Determinism)

Reliability is a fundamental property any AV method must fulfill in order to be applicable in real-world forensic settings. However, since there is no consistent concept nor a uniform definition of the term "**reliability**" in the context of authorship verification according to the screened literature, we decided to reuse a definition from applied statistics, and adapt it carefully to AV.

In his standard reference[2] book, Bollen [4] gives a clear description for this term: "*Reliability is the consistency of measurement*" and provides a simple example to illustrate its meaning: At time t_1 we ask a large number of persons the same question Q and record their responses. Afterwards, we remove their memory of the dialogue. At time t_2 we ask them again the same question Q and record their responses again. "*The reliability is the consistency of the responses across individuals for the two time periods. To the extent that all individuals are consistent, the measure is reliable*" [4]. This example deals with the consistency of the measured objects as a factor for the reliability of measurements. In the case of authorship verification, the analyzed objects are static data, and hence these cannot be a source of inconsistency. However, the measurement system itself can behave inconsistently and hence unreliable. This aspect can be described as **intra-rater reliability**.

Reliability in authorship verification is satisfied, if an AV method always generates the same prediction $\alpha \in \{\texttt{Y}, \texttt{N}\}$ for the same input $\rho = (\mathcal{D}_{\mathcal{U}}, \mathcal{D}_{\mathcal{A}})$, or in other words, if the method behaves **deterministically**. Several AV approaches, including [8–11, 15–18, 26] fall into this category. In contrast, if an AV method behaves **non-deterministically** such that two different predictions for ρ are possible, the method can be rated as unreliable. Many AV approaches, including [2, 3, 12, 13, 21, 22, 24, 27, 33] belong to this category, since they involve randomness (e. g., weight initialization, feature subsampling, chunk generation or impostor selection), which might distort the evaluation, as every run on a test corpus very likely leads to different results. Under lab conditions, results of non-deterministic AV methods can (and should) be counteracted by averaging multiple runs. However, it remains highly questionable if such methods are generally applicable in realistic forensic cases, where the prediction α regarding a verification case ρ might sometimes result in Y and sometimes in N.

[2] According to *Google Scholar*, Bollen's book was cited more than 30,000 times making it a standard reference across different research fields.

3.2 Optimizability

Another important property of an AV method is optimizability. We define an AV method as **optimizable**, if it is designed in such a way that it offers adjustable hyperparameters that can be tuned against a training/validation corpus, given an optimization method such as grid or random search. Hyperparameters might be, for instance, the selected distance/similarity function, the number of layers and neurons in a neural network or the choice of a kernel method. The majority of existing AV approaches in the literature (for example, [6, 13, 15, 16, 18, 22, 23, 26]) belong to this category. On the other hand, if a published AV approach involves hyperparameters that have been entirely fixed such that there is no further possibility to improve its performance from outside (without deviating from the definitions in the publication of the method), the method is considered to be **non-optimizable**. Non-optimizable AV methods are preferable in forensic settings as, here, the existence of a training/validation corpus is not always self-evident. Among the proposed AV approaches in the respective literature, we identified only a small fraction [8, 21, 42] that fall into this category.

3.3 Model Category

From a machine learning point of view, authorship verification represents a unary classification problem [15, 22, 26, 28, 40]. Yet, in the literature, it can be observed that sometimes AV is treated as a unary [17, 18, 24, 26] and sometimes as a binary classification task [15, 21, 23, 42]. We define the way an AV approach is modeled by the phrase **model category**. However, before explaining this in more detail, we wish to recall what unary/one-class classification exactly represents. For this, we list the following verbatim quotes, which characterize one-class classification, as can be seen, almost identically (emphasis by us):

- *"In one-class classification it is assumed that only information of one of the classes, the target class, is available. This means that* ***just example objects of the target class*** *can be used and that* ***no information*** *about the other class* ***of outlier objects is present****."* [41]
- *"One-class classification (OCC) [. . .] consists in making a description of a target class of objects and in detecting whether a new object resembles this class or not. [. . .] The OCC model is developed* ***using target class samples only****."* [31]
- *"In one-class classification framework, an object is classified as belonging or not belonging to a target class, while* ***only*** *sample* ***examples*** *of objects* ***from the target class*** *are available during the training phase."* [17]

Note that in the context of authorship verification, target class refers to the known author $\mathcal{A}$ such that for a document $\mathcal{D}_\mathcal{U}$ of an unknown author $\mathcal{U}$ the task is to verify whether $\mathcal{U} = \mathcal{A}$ holds. One of the most important requirements of any existing AV method is a **decision criterion**, which aims to accept or reject a questioned authorship. A decision criterion can be expressed through a simple scalar threshold θ or a more complex model $\theta_\mathcal{M}$ such as a hyperplane in a high-dimensional feature space. As a consequence of the above statements, the determination of θ or $\theta_\mathcal{M}$ has to be performed solely on the basis of $\mathbb{D}_\mathcal{A}$, otherwise the AV method cannot be considered to be unary. However, our conducted literature research regarding existing AV approaches revealed that there are uncertainties how to precisely draw the borders between unary and binary AV methods (for instance, [5, 26, 28]). Nonetheless, few attempts have been made to distinguish both categories from another perspective. Potha and Stamatatos [28], for example, categorize AV methods as either **intrinsic** or **extrinsic** (emphasis by us):

(1) *"Intrinsic verification models view it [i. e., the verification task] as a one-class classification task and are based exclusively on analysing the similarity between [$\mathbb{D}_\mathcal{A}$] and [$\mathcal{D}_\mathcal{U}$]. [. . .] Such methods [. . .]* ***do not require any external resources****."* [28]

(2) *"On the other hand, extrinsic verification models attempt to transform the verification task to a pair classification task by* ***considering external documents*** *to be used as* ***samples of the negative class****."* [28]

While we agree with statement (2), the former statement (1) is unsatisfactory, as intrinsic verification models are **not necessarily unary**. For example, the AV approach GLAD proposed by Hürlimann et al. [15] directly contradicts statement (1). Here, the authors *"decided to cast the problem as a* ***binary classification task*** *where class values are* Y *[$\mathcal{A} = \mathcal{U}$] and* N *[$\mathcal{A} \neq \mathcal{U}$]. [. . .] We do* ***not introduce any negative examples*** *by means of external documents, thus adhering to an* ***intrinsic approach****."* [15].

A misconception similar to statement (1) can be observed in the paper of Jankowska et al. [16], who introduced the so-called CNG approach claimed to be a one-class classification method. CNG is intrinsic in that way that it considers only $\mathbb{D}_\mathcal{A}$ when deciding a problem ρ. However, the decision criterion, which is a threshold θ, is determined on a set of verification problems, labeled either as Y or N. This incorporates "external resources" for defining the decision criterion, and it constitutes an implementation of binary classification between Y and N in analogy to the statement of Hürlimann et al. [15] mentioned above. Thus, CNG is in conflict with the unary definition mentioned above. In a subsequent paper [17], however, Jankowska et al. refined their approach and introduced a modification, where θ was determined solely on the basis of $\mathbb{D}_\mathcal{A}$. Thus, the modified approach can be considered as a true unary AV method, according to the quoted definitions for unary classification.

In 2004, Koppel and Schler [22] presented the Unmasking approach which, according to the authors, represents a unary AV method. However, if we take a closer look at the learning process of Unmasking, we can see[3] that it is based on a binary SVM classifier that consumes feature vectors (derived from *"degradation curves"*) labeled as Y ("same author") or N ("different author"). Unmasking, therefore, cannot be considered to be unary as the decision is not solely based on the documents within $\mathbb{D}_\mathcal{A}$, in analogy to the CNG approach of Jankowska et al. [16] discussed above.

It should be highlighted again that the aforementioned three approaches are **binary-intrinsic** since their decision criteria θ or $\theta_\mathcal{M}$ was determined on a set of problems labeled in a binary manner (Y and N) while after training, the verification is performed in an intrinsic manner, meaning that $\mathbb{D}_\mathcal{A}$ and $\mathcal{D}_\mathcal{U}$ are compared against θ or $\theta_\mathcal{M}$ but not against documents within other verification problems (cf. Figure 1). A crucial aspect, which might have lead to misperceptions regarding the model category of these approaches in the past, is the fact that two different class domains are involved. On the one hand, there is the **class domain of authors**, where the task

[3] See the intuitive illustration provided in [3, Figure 1].

is to distinguish $\mathcal{A}$ and $\neg\mathcal{A}$. On the other hand, there is the *elevated* or *lifted* **domain of verification problem classes**, which are Y and N. The training phase of binary-intrinsic approaches is used for learning to distinguish these two classes, and the verification task can be understood as putting the verification problem as a whole into class Y or class N, whereby the class domain of authors fades from the spotlight (cf. Figure 1).

Besides unary and binary-intrinsic methods, there is a third category of approaches, namely **binary-extrinsic** AV approaches (for example, [2, 12, 20, 21, 23, 27, 42]). These methods use external documents during a potentially existing training phase and – more importantly – during testing. In these approaches, the decision between $\mathcal{A}$ and $\neg\mathcal{A}$ is put into the focus, where the external documents aim to construct the counter class $\neg\mathcal{A}$.

Based on the above observations, we conclude that the key requirement for judging the model category of an AV method depends solely on the aspect **how** its decision criterion θ or $\theta_{\mathcal{M}}$ is determined (cf. Figure 1):

(1) An AV method is **unary** if and only if its decision criterion θ or $\theta_{\mathcal{M}}$ is determined solely on the basis of the target class $\mathcal{A}$ during testing. As a consequence, an AV method cannot be considered to be unary if documents not belonging to $\mathcal{A}$ are used to define θ or $\theta_{\mathcal{M}}$.
(2) An AV method is **binary-intrinsic** if its decision criterion θ or $\theta_{\mathcal{M}}$ is determined on a training corpus comprising verification problems labeled either as Y or N (in other words documents of several authors). However, once the training is completed, a binary-intrinsic method has no access to external documents anymore such that the decision regarding the authorship of $\mathcal{D}_{\mathcal{U}}$ is made on the basis of the reference data of $\mathcal{A}$ as well as θ or $\theta_{\mathcal{M}}$.
(3) An AV method is **binary-extrinsic** if its decision criterion θ or $\theta_{\mathcal{M}}$ is determined during testing on the basis of external documents that represent the outlier class $\neg\mathcal{A}$.

Note that optimizable AV methods such as [9, 17] are not excluded to be unary. Provided that θ or $\theta_{\mathcal{M}}$ is not subject of the optimization procedure, the model category remains unary. The reason for this is obvious; **Hyperparameters** might influence the resulting performance of unary AV methods. The **decision criterion** itself, however, remains unchanged.

3.4 Implications

Each model category has its own implications regarding prerequisites, evaluability, and applicability.

3.4.1 Unary AV Methods: One advantage of unary AV methods is that they do not require a specific document collection strategy to construct the counter class $\neg\mathcal{A}$, which reduces their complexity. On the downside, the choice of the underlying machine learning model of a unary AV approach is restricted to one-class classification algorithms or unsupervised learning techniques, given a suitable decision criterion.

However, a far more important implication of unary AV approaches concerns their **performance assessment**. Since unary classification (not necessarily AV) approaches depend on a fixed decision criterion θ or $\theta_{\mathcal{M}}$, performance measures such as the area under the ROC curve (AUC) are meaningless. Recall that ROC analysis is used for evaluating classifiers, where the decision threshold is not finally fixed. ROC analysis requires that the classifier generates scores, which are comparable across classification problem instances. The ROC curve and the area under this curve is then computed by considering all possible discrimination thresholds for these scores. While unary AV approaches might produce such scores, introducing a **variable** θ would change the semantics of these approaches. Since unary AV approaches have a fixed decision criterion, they provide only a single point in the ROC space. To assess the performance of a unary AV method, it is, therefore, mandatory to consider the confusion matrix that leads to this point in the ROC space.

Another implication is that unary AV methods are necessarily instance-based and, thus, require a set $\mathbb{D}_{\mathcal{A}} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots\}$ of multiple documents of the known author $\mathcal{A}$. If only one reference document is available ($\mathbb{D}_{\mathcal{A}} = \{\mathcal{D}_{\mathcal{A}}\}$), this document must be artificially turned into multiple samples from the author. In general, unary classification methods need multiple samples from the target class since it is not possible to determine a *relative* closeness to that class based on only one sample.

3.4.2 Binary AV Methods: On the plus side, binary-intrinsic or extrinsic AV methods benefit from the fact that we can choose among a variety of **binary**[4] and ***n*-ary**[5] classification models. However, if we consider designing a binary-intrinsic AV method, it should not be overlooked that the involved classifier will learn nothing about individual authors, but only similarities or differences that hold in general for Y and N verification problems [23].

If, on the other hand, the choice falls on a binary-extrinsic method, a strategy has to be considered for collecting representative documents for the outlier class $\neg\mathcal{A}$. Several existing methods such as [23, 27, 42] rely on search engines for retrieving appropriate documents, but these search engines might refuse their service if a specified quota is exhausted. Additionally, the retrieved documents render these methods inherently **non-deterministic**. Moreover, such methods cause relatively **high runtimes** [19, 38]. Using search engines also requires an active Internet connection, which might not be available or allowed in specific scenarios. But even if we can access the Internet to retrieve documents, there is **no guarantee** that the true author is not among them. With these points in mind, the **applicability** of binary-extrinsic methods in real-world cases, i. e., in real forensic settings, remains highly questionable.

4 Methodology

In the following, we introduce our three self-compiled corpora, where each corpus represents a different challenge. Next, we describe which authorship verification approaches we considered for the experiments and classify each AV method according to the properties introduced in Section 3. Afterwards, we explain which performance measures were selected with respect to the conclusion made in Section 3.4.1. Finally, we describe our experiments, present the results and highlight a number of observations.

[4] For example: Support vector machines, logistic regression or perceptrons.
[5] For example: Naive Bayes, random forests or a variety of neural networks.

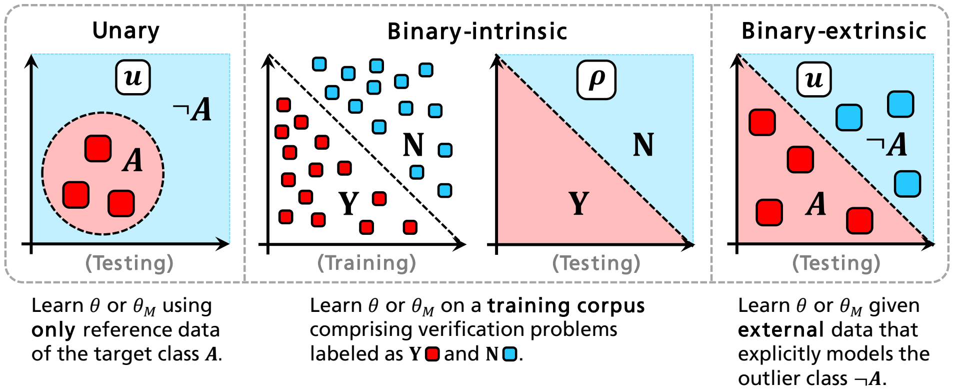


Figure 1: The three possible model categories of authorship verification approaches. Here, $\mathcal{U}$ refers to the instance (for example, a document or a feature vector) of the unknown author. $\mathcal{A}$ is the target class (known author) and $\neg\mathcal{A}$ the outlier class (any other possible author). In the binary-intrinsic case, ρ denotes the verification problem (subject of classification), and Y and N denote the regions of the problem feature space where, according to a training corpus, the authorship holds or not.

Corpus C	$\lvert C \rvert$	$\lvert \mathbb{D}_{\mathcal{A}} \rvert$	avg_len($\mathcal{D}_{\mathcal{U}}$)	avg_len($\mathcal{D}_{\mathcal{A}}$)
C_{DBLP} (train)	32	1	6,771	5,491
C_{DBLP} (test)	48	1	7,650	5,714
C_{Perv} (train)	440	2	8,157	7,268
C_{Perv} (test)	660	2	9,611	8,692
C_{Reddit} (train)	40	1	7,011	6,909
C_{Reddit} (test)	60	1	6,974	6,990

Table 1: All training and testing corpora used in our experiments. Here, $\lvert C \rvert$ denotes the number of verification problems in each corpus C and $\lvert \mathbb{D}_{\mathcal{A}} \rvert$ the number of the known documents. The average character length of the unknown document $\mathcal{D}_{\mathcal{U}}$ and the known document $\mathcal{D}_{\mathcal{A}}$ (concatenation of all known documents $\mathbb{D}_{\mathcal{A}}$) is denoted by avg_len($\mathcal{D}_{\mathcal{U}}$) and avg_len($\mathcal{D}_{\mathcal{A}}$), respectively.

4.1 Corpora

A serious challenge in the field of AV is the lack of publicly available (and suitable) corpora, which are required to train and evaluate AV methods. Among the few publicly available corpora are those that were released by the organizers of the well-known PAN-AV competitions[6] [19, 37, 38]. In regard to our experiments, however, we cannot use these corpora, due to the absence of relevant metadata such as the precise time spans where the documents have been written as well as the topic category of the texts. Therefore, we decided to compile our own corpora based on English documents, which we crawled from different publicly accessible sources. In what follows, we describe our three constructed corpora, which are listed together with their statistics in Table 1. Note that all corpora are balanced such that verification cases with **matching** (Y) and **non-matching** (N) authorships are evenly distributed.

4.1.1 DBLP Corpus As a first corpus, we compiled C_{DBLP} that represents a collection of 80 excerpts from scientific works including papers, dissertations, book chapters and technical reports, which we have chosen from the well-known *Digital Bibliography & Library Project* (DBLP) platform[7]. Overall, the documents[8] were written by 40 researchers, where for each author $\mathcal{A}$, there are exactly two documents. Given the 80 documents, we constructed for each author $\mathcal{A}$ two verification problems ρ_1 (a Y-case) and ρ_2 (an N-case). For ρ_1 we set $\mathcal{A}$'s first document as $\mathcal{D}_{\mathcal{A}}$ and the second document as $\mathcal{D}_{\mathcal{U}}$. For ρ_2 we reuse $\mathcal{D}_{\mathcal{A}}$ from ρ_1 as the known document and selected a text from another (random) author as the unknown document. The result of this procedure is a set of 80 verification problems, which we split into a training and test set based on a 40/60% ratio. Where possible, we tried to **restrict** the content of each text to the **abstract** and **conclusion** of the original work. However, since in many cases these sections were too short, we also considered other parts of the original works such as introduction or discussion sections. To ensure that the extracted text portions are appropriate for the AV task, each original work was **preprocessed manually**. More precisely, we removed tables, formulas, citations, quotes and sentences that include non-language content such as mathematical constructs or specific names of researchers, systems or algorithms. The average time span between both documents of an author is 15.6 years. The minimum and maximum time span are 6 and 40 years, respectively. Besides the temporal aspect of C_{DBLP}, another challenge of this corpus is the formal (scientific) language, where the usage of stylistic devices[9] is more restricted, in contrast to other genres such as novels or poems.

4.1.2 Perverted Justice Corpus As a second corpus, we compiled C_{Perv}, which represents a collection of 1,645 chat conversations of 550 sex offenders crawled from the *Perverted-Justice* portal[10]. The chat conversations stem from a variety of sources including emails and instant messengers (e. g., *MSN*, *AOL* or *Yahoo*), where for each conversation, we ensured that only chat lines from the offender were extracted. We applied the same problem construction procedure as for the corpus C_{DBLP}, which resulted in 1,100 verification

[6] https://pan.webis.de

[7] https://dblp.uni-trier.de

[8] Note that each document is single-authored.

[9] For example, repetitions, metaphors, rhetorical questions, oxymorons, etc.

[10] http://www.perverted-justice.com

AV Method	Model Categ.	Optimizability	Determinism
MOCC [9]	unary	determin.	optimizable
OCCAV [8]	unary	determin.	non-optimiz.
COAV [10]	binary-intr.	determin.	optimizable
AVeer [11]	binary-intr.	determin.	optimizable
GLAD [15]	binary-intr.	determin.	optimizable
DistAV [18]	binary-intr.	determin.	optimizable
Unmasking [22]	binary-intr.	non-determin.	optimizable
Caravel [2]	binary-extr.	non-determin.	optimizable
GenIM [33]	binary-extr.	non-determin.	optimizable
ImpGI [27]	binary-extr.	non-determin.	optimizable
SPATIUM [21]	binary-extr.	non-determin.	non-optimiz.
NNCD [42]	binary-extr.	determin.	non-optimiz.

Table 2: All 12 AV methods, classified according to their properties.

problems that again were split into a training and test set given a 40/60% ratio. In contrast to the corpus $\mathcal{C}_{\text{DBLP}}$, we only performed slight preprocessing. Essentially, we removed user names, timestamps, URLs, multiple blanks as well as annotations that were not part of the original conversations from all chat lines. Moreover, we did not normalize words (for example, shorten words such as *"nooooo"* to *"no"*) as we believe that these represent important style markers. Furthermore, we did not remove newlines between the chat lines, as the positions of specific words might play an important role regarding the individual's writing style.

4.1.3 Reddit Corpus As a third corpus, we compiled $\mathcal{C}_{\text{Reddit}}$, which is a collection of 200 aggregated postings crawled from the *Reddit* platform[11]. Overall, the postings were written by 100 Reddit users and stem from a variety of subreddits. In order to construct the Y-cases, we selected exactly two postings from disjoint subreddits for each user such that both the known and unknown document $\mathcal{D}_{\mathcal{A}}$ and $\mathcal{D}_{\mathcal{U}}$ differ in their topic. Regarding the N-cases, we applied the opposite strategy such that $\mathcal{D}_{\mathcal{A}}$ and $\mathcal{D}_{\mathcal{U}}$ belong to the same topic. The rationale behind this is to figure out to which extent AV methods can be fooled in cases, where the topic matches but not the authorship and vice versa. Since for this specific corpus we have to control the topics of the documents, we did not perform the same procedure applied for $\mathcal{C}_{\text{DBLP}}$ and $\mathcal{C}_{\text{Perv}}$ to construct the training and test sets. Instead, we used for the resulting 100 verification problems a 40/60% hold-out split, where both training and test set are entirely disjoint.

4.2 Examined Authorship Verification Methods

As a basis for our experiments, we reimplemented 12 existing AV approaches, which have shown their potentials in the previous PAN-AV competitions [19, 37] as well as in a number of AV studies. The methods are listed in Table 2 together with their classifications regarding the AV characteristics, which we proposed in Section 3. All (optimizable) AV methods were tuned regarding their hyperparameters, according to the original procedure mentioned in the respective paper. However, in the case of the binary-extrinsic methods (GenIM, ImpGI and NNCD) we had to use an alternative impostors generation strategy in our reimplementations, due to technical problems. In the respective papers, the authors used search engine queries to generate the impostor documents, which are needed to model the counter class $\neg\mathcal{A}$. Regarding our reimplementations, we used the documents from the static corpora (similarly to the idea of Kocher and Savoy [21]) to generate the impostors in the following manner: Let $\mathcal{C} = \{\rho_1, \rho_2, \ldots, \rho_n\}$ denote a corpus with n verification problems. For each $\rho_i = (\mathcal{D}_{\mathcal{U}_i}, \mathbb{D}_{\mathcal{A}_i})$ we choose all unknown documents $\mathcal{D}_{\mathcal{U}_j}$ in $\mathcal{C}$ with $i \neq j$ and append them the impostor set $\mathbb{U}$. Here, it should be highlighted that both GenIM and ImpGI consider the number of impostors as a **hyperparameter** such that the resulting impostor set is a subset of $\mathbb{U}$. In contrast to this, NNCD considers all $\mathcal{U}_j \in \mathbb{U}$ as possible impostors. This fact plays an important role in the later experiments, where we compare the AV approaches to each other. Although our strategy is not flexible like using a search engine, it has one advantage that, here, it is assumed that the true author of an unknown document is not among the impostors, since in our corpora the user/author names are known[12] beforehand.

4.3 Performance Measures

According to our extensive literature research, numerous measures (e. g., Accuracy, F_1, c@1, AUC, AUC@1, κ or EER) have been used so far to assess the performance of AV methods. In regard to our experiments, we decided to use c@1 and AUC for several reasons. First, Accuracy, F_1 and κ are not applicable in cases where AV methods leave verification problems unanswered, which concerns some of our examined AV approaches. Second, using AUC alone is meaningless for non-optimizable AV methods, as explained in Section 3.4.1. Third, both have been used in the PAN-AV competitions [37, 38]. Note that we also list the confusion matrix outcomes.

4.4 Experiments

Overall, we focus on three experiments, which are based on the corpora introduced in Section 4.1:

(1) The Effect of Stylistic Variation Across Large Time Spans
(2) The Effect of Topical Influence
(3) The Effect of Limited Text Length

In the following each experiment is described in detail.

4.4.1 The Effect of Stylistic Variation Across Large Time Spans: In this experiment, we seek to answer the question if the writing style of an author $\mathcal{A}$ can be recognized, given a large time span between two documents of $\mathcal{A}$. The motivation behind this experiment is based on the statement of Olsson [25] that language acquisition is a **continuous process**, which is not only acquired, but also **can be lost**. Therefore, an important question that arises here is, if the writing style of a person remains "stable" across a large time span, given the fact that language in each individual's life is never "fixed" [25]. Regarding this experiment, we used the $\mathcal{C}_{\text{DBLP}}$ corpus. The results of the 12 examined AV methods are listed in Table 3, where it can be seen that the majority of the examined AV

[11] https://www.reddit.com

[12] However, it might be possible that behind multiple user names there is only one person (in other words, we cannot guarantee: one user = one account).

AV Method	c@1	AUC	TP	FN	FP	TN	UP
Caravel	**0.792**	0.905	19	5	5	19	0
COAV	**0.750**	0.802	18	6	6	18	0
NNCD	0.729	0.858	14	10	3	21	0
SPATIUM	0.717	0.788	15	5	5	14	9
GLAD	0.708	**0.821**	14	10	4	20	0
ImpGI	0.708	0.816	22	2	12	12	0
GenIM	0.703	0.768	14	7	5	16	6
Unmasking	0.688	0.747	19	5	10	14	0
AVeer	0.667	0.781	17	7	9	15	0
DistAV	0.583	0.681	7	17	3	21	0
MOCC	0.542	0.655	7	17	5	19	0
OCCAV	0.521	0.750	1	23	0	24	0

Table 3: Evaluation results for the test corpus $\mathcal{C}_{\text{DBLP}}$ in terms of c@1 and AUC. TP, FN, FP and TN represent the four confusion matrix outcomes, while UP denotes the number of unanswered verification problems. Note that AUC scores for the non-optimizable and unary AV methods are grayed out.

AV Method	c@1	AUC	TP	FN	FP	TN	UP
GenIM	**0.533**	0.521	22	8	20	10	0
MOCC	0.517	0.537	8	22	7	23	0
AVeer	0.517	0.492	18	12	17	13	0
NNCD	0.517	0.561	13	17	12	18	0
Unmasking	0.508	**0.559**	15	14	15	15	1
ImpGI	0.500	0.456	30	0	30	0	0
OCCAV	0.483	0.428	0	30	1	29	0
COAV	0.450	0.532	18	12	21	9	0
SPATIUM	0.432	0.487	14	9	16	7	14
Caravel	0.426	0.451	13	10	12	14	11
DistAV	0.417	0.434	3	27	8	22	0
GLAD	0.350	0.340	8	22	17	13	0

Table 4: Evaluation results for the test corpus $\mathcal{C}_{\text{Reddit}}$.

methods yield useful recognition results with a maximum value of 0.792 in terms of c@1. With the exception of the binary-intrinsic approach COAV, the remaining top performing methods belong to the binary-extrinsic category. This category of AV methods has also been superior in the PAN-AV competitions [19, 37, 38], where they outperformed binary-intrinsic and unary approaches three times in a row (2013–2015).

The top performing approaches Caravel, COAV and NNCD deserve closer attention. All three are based on character-level language models that capture low-level features similar to character n-grams, which have been shown in numerous AA and AV studies (for instance, [24, 35]) to be highly effective and robust. In [3, 10], it has been shown that Caravel and COAV were also the two top-performing approaches, where in [10] they were evaluated on the PAN-2015 AV corpus [37], while in [3] they were applied[13] on texts obtained from *Project Gutenberg*. Although both approaches perform similarly, they differ in the way how the decision criterion θ is determined. While COAV requires a training corpus to learn θ, Caravel assumes that the given test corpus (which provides the **impostors**) is balanced. Given this assumption, Caravel first computes similarity scores for all verification problems in the corpus and then sets θ to the median of all similarities (cf. Figure 3). Thus, from a machine learning perspective, there is some undue training on the test set. Moreover, the applicability of Caravel in realistic scenarios is questionable, as a forensic case is *not* part of a corpus where the Y/N-distribution is **known beforehand**.

Another interesting observation can be made regarding COAV, NNCD and OCCAV. Although all three differ regarding their model category, they use the same underlying compression algorithm (PPMd) that is responsible for generating the language model. While the former two approaches perform similarly well, OCCAV achieves a poor c@1 score ($\approx$ 0.5). An obvious explanation for this is a wrongly calibrated threshold θ, as can be seen from the confusion matrix, where almost all answers are N-predictions. Regarding the NNCD approach, one should consider that $\mathcal{D}_{\mathcal{U}}$ is compared against $\mathcal{D}_{\mathcal{A}}$ as well as $n-1$ impostors within a corpus comprised of n verification problems. Therefore, a Y-result is correct with relatively high certainty (i. e., the method has high precision compared to other approaches with a similar c@1 score), as NNCD decided that author $\mathcal{A}$ fits best to $\mathcal{D}_{\mathcal{U}}$ among n candidates. In contrast to Caravel, NNCD only retrieves the impostors from the given corpus, but it does not exploit background knowledge about the distribution of problems in the corpus.

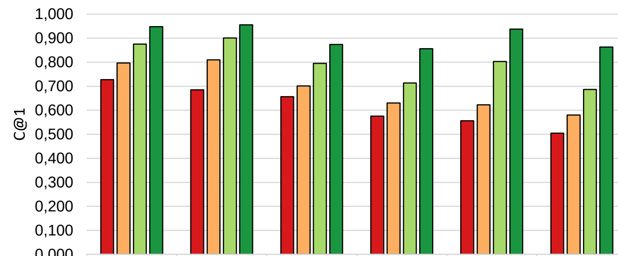
Overall, the results indicate that it is possible to recognize writing styles across large time spans. To gain more insights regarding the question which features led to the correct predictions, we inspected the AVeer method. Although the method achieved only average results, it benefits from the fact that it can be interpreted easily, as it relies on a simple distance function, a fixed threshold θ and predefined feature categories such as function words. Regarding the correctly recognized Y-cases, we noticed that conjunctive adverbs such as *"hence"*, *"therefore"* or *"moreover"* contributed mostly to AVeer's correct predictions. However, a more in-depth analysis is required in future work to figure out whether the decisions of the remaining methods are also primarily affected by these features.

4.4.2 The Effect of Topical Influence: In this experiment, we investigate the question if the writing style of authors can be recognized under the influence of topical bias. In real-world scenarios, the topic of the documents within a verification problem ρ is not always known beforehand, which can lead to a serious challenge regarding the recognition of the writing style. Imagine, for example, that ρ consists of a known and unknown document $\mathcal{D}_{\mathcal{A}}$ and $\mathcal{D}_{\mathcal{U}}$ that are written by the same author ($\mathcal{A} = \mathcal{U}$) while at the same time differ regarding their topic. In such a case, an AV method that it focusing "too much" on the topic (for example on specific nouns or phrases) will likely predict a different authorship ($\mathcal{A} \neq \mathcal{U}$). On the other hand, when $\mathcal{D}_{\mathcal{A}}$ and $\mathcal{D}_{\mathcal{U}}$ match regarding their topic, while being written by different authors, a topically biased AV method might erroneously predict $\mathcal{A} = \mathcal{U}$. In the following we show to which extent these assumptions hold. As a data basis for this experiment, we used the $\mathcal{C}_{\text{Reddit}}$ corpus introduced in Section 4.1.3. The results regarding the 12 AV methods are given in Table 4, where it can be

[13] Note that the implementation in [3] differs from the one used in this paper.

AV Method	c@1	AUC	TP	FN	FP	TN	UP
NNCD	**0.991**	0.999	325	5	1	329	0
Caravel	0.973	**0.995**	319	9	9	319	4
GLAD	0.970	**0.995**	317	13	7	323	0
AVeer	0.924	0.975	298	32	18	312	0
COAV	0.923	0.980	303	27	24	306	0
Unmasking	0.918	0.975	314	16	38	292	0
SPATIUM	0.892	0.971	312	10	53	235	50
ImpGI	0.886	0.974	322	8	67	263	0
GenIM	0.885	0.973	323	7	69	261	0
MOCC	0.853	0.960	246	84	13	317	0
DistAV	0.847	0.913	282	48	53	277	0
OCCAV	0.798	0.992	197	133	0	330	0

Table 5: Evaluation results for the test corpus C_{Perv}.



	GLAD	Caravel	Unmasking	AVeer	NNCD	COAV
250 char.	0,727	0,685	0,656	0,576	0,556	0,505
500 char.	0,797	0,810	0,702	0,630	0,623	0,580
1000 char.	0,876	0,901	0,795	0,714	0,803	0,686
2000 char.	0,948	0,956	0,874	0,856	0,938	0,864

Figure 2: Evaluation results for the four versions of the test corpus C_{Perv} in terms of c@1.

AV Method	TP	FN	FP	TN	Total (Y/N/UP)
GLAD	203	127	53	277	(256/404/0)
Caravel	225	103	104	226	(329/329/2)
Unmasking	158	169	56	272	(214/441/5)
AVeer	56	274	6	324	(62/598/0)
NNCD	40	290	3	327	(43/617/0)
COAV	328	2	325	5	(653/7/0)

Table 6: Confusion matrix outcomes for the 250 characters version of the test corpus C_{Perv}.

seen that our assumptions hold. All examined AV methods (with no exception) are fooled by the topical bias in the corpus. Here, the highest achieved results in terms of c@1 and AUC are very close to random guessing. A closer look at the confusion matrix outcomes reveals that some methods, for example ImpGI and OCCAV, perform almost entirely inverse to each other, where the former predicts nothing but Y and the latter nothing but N (except 1 Y). Moreover, we can assume that the lower c@1 is, the stronger is the focus of the respective AV method on the topic of the documents. Overall, the results of this experiment suggest that none of the examined AV methods is robust against topical influence.

4.4.3 The Effect of Limited Text Length: In our third experiment, we investigate the question how text lengths affect the results of the examined AV methods. The motivation behind this experiment is based on the observation of Stamatatos et al. [37] that text length is an important issue, which has not been thoroughly studied within authorship verification research. To address this issue, we make use of the C_{Perv} corpus introduced in Section 4.1.2. The corpus is suitable for this purpose, as it comprises a large number of verification problems, where more than 90% of all documents have sufficient text lengths (≥ 2,000 characters). This allows a stepwise truncation and by this to analyze the effect between the text lengths and the recognition results. However, before considering this, we first focus on the results (shown in Table 5) after applying all 12 AV methods on the original test corpus. As can be seen in Table 5, almost all approaches perform very well with c@1 scores up to 0.991. Although these results are quite impressive, it should be noted that a large fraction of the documents comprises thousands of words. Thus, the methods can learn precise representations based on a large variety of features, which in turn enable a good determination of (dis)similarities between known/unknown documents. To investigate this issue in more detail, we constructed four versions of the test corpus and equalized the unknown document lengths to 250, 500, 1000, and 2000 characters. Then, we applied the top performing AV methods with a c@1 value > 0.9 on the four corpora. Here, we reused the same models and hyperparameters (including the decision criteria θ and $\theta_{\mathcal{M}}$) that were determined on the training corpus. The intention behind this was to observe the robustness of the trained AV models, given the fact that during training they were confronted with longer documents. The results are illustrated in Figure 2, where it can be observed that GLAD yields the most stable results across the four corpora versions, where even for the corpus with the 250 characters long unknown documents, it achieves a c@1 score of 0.727. Surprisingly, Unmasking performs similarly well, despite of the fact that the method has been designed for longer texts i. e., book chunks of **at least 500 words** [22]. Sanderson and Guenter also point out that the Unmasking approach is less useful when dealing with relatively short texts [32]. However, our results show a different picture, at least for this corpus.

One explanation of the resilience of GLAD across the varying text lengths might be due to its decision model $\theta_{\mathcal{M}}$ (an SVM with a linear kernel) that withstands the absence of missing features caused by the truncation of the documents, in contrast to the distance-based approaches AVeer, NNCD and COAV, where the decision criterion θ is reflected by a simple scalar. Table 6 lists the confusion matrix outcomes of the six AV methods regarding the 250 characters version of C_{Perv}. Here, it can be seen that the underlying SVM model of GLAD and Unmasking is able to regulate its Y/N-predictions, in contrast to the three distance-based methods, where the majority of predictions fall either on the Y- or on the N-side. To gain a better picture regarding the stability of the decision criteria θ and $\theta_{\mathcal{M}}$ of the methods, we decided to take a closer look on the ROC curves (cf. Figure 3) generated by GLAD, Caravel and COAV for the four corpora versions, where a number of interesting observations can be made. When focusing on AUC, it turns out that all three methods perform very similar to each other, whereas big discrepancies between GLAD and COAV can be observed regarding c@1. When we consider the current and maximum achievable results (depicted by the circles and triangles, respectively) it becomes apparent that GLAD's model behaves stable, while the one of COAV becomes increasingly vulnerable the more the documents are shortened. When looking at the ROC curve of Caravel, it can be clearly seen that the actual and maximum achievable results are very close to each other. This is not surprising, due to the fact that Caravel's threshold always lies at the median point of the ROC curve, provided that the given corpus is balanced. While inspecting the 250 characters long documents in more detail, we identified that they share similar vocabularies consisting of chat abbreviations such as *"lol"* (laughing out loud) or *"k"* (ok), smileys and specific obscene words. Therefore, we assume that the verification results of the examined methods are mainly caused by the similar vocabularies between the texts.

5 Conclusion and Future Work

We highlighted the problem that underlying characteristics of authorship verification approaches have not been paid much attention in the past research and that these affect the applicability of the methods in real forensic settings. Then, we proposed several properties that enable a better characterization and by this a better comparison between AV methods. Among others, we explained that the performance measure AUC is meaningless in regard to **unary** or specific **non-optimizable** AV methods, which involve a fixed decision criterion (for example, NNCD). Additionally, we mentioned that determinism must be fulfilled such that an AV method can be rated as reliable. Moreover, we clarified a number of misunderstandings in previous research works and proposed three clear criteria that allow to classify the model category of an AV method, which in turn influences its design and the way how it should be evaluated. In regard to binary-extrinsic AV approaches, we explained which challenges exist and how they affect their applicability.

In an experimental setup, we applied 12 existing AV methods on three self-compiled corpora, where the intention behind each corpus was to focus on a different aspect of the methods applicability. Our findings regarding the examined approaches can be summarized as follows: Despite of the good performance of the five AV methods GenIM, ImpGI, Unmasking, Caravel and SPATIUM, none of them can be truly considered as reliable and therefore applicable in real forensic cases. The reason for this is not only the non-deterministic behavior of the methods but also their dependence (excepting Unmasking) on an impostor corpus. Here, it must be guaranteed that the true author is not among the candidates, but also that the impostor documents are suitable such that the AV task not inadvertently degenerates from style to topic classification. In particular, the applicability of the Caravel approach remains highly questionable, as it requires a corpus where the information regarding Y/N-distribution is known beforehand in order to set the threshold. In regard to the two examined unary AV approaches MOCC and OCCAV, we observed that these perform poorly on all three corpora in comparison to the binary-intrinsic and binary-extrinsic methods. Most likely, this is caused by the wrong threshold setting, as both tend to generate more N-predictions. From the remaining approaches, GLAD and COAV seem to be a good choice for realistic scenarios. However, the former has been shown to be more robust in regard to varying text lengths given a fixed model, while the latter requires a retraining of the model (note that both performed almost equal in terms of AUC). Our hypothesis, which we leave open for future work, is that AV methods relying on a complex model $\theta_{\mathcal{M}}$ are more robust than methods based on a scalar-threshold θ. Lastly, we wish to underline that all examined approaches failed in the cross-topic experiment. One possibility to counteract this is to apply text distortion techniques (for instance, [36]) in order to control the topic influence in the documents.

As one next step, we will compile additional and larger corpora to investigate the question whether the evaluation results of this paper hold more generally. Furthermore, we will address the important question how the results of AV methods can be *interpreted* in a more systematic manner, which will further influence the practicability of AV methods besides the proposed properties.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research (BMBF) under the project "DORIAN" (Scrutinise and thwart disinformation).

References

[1] Hosein Azarbonyad, Mostafa Dehghani, Maarten Marx, and Jaap Kamps. 2015. Time-Aware Authorship Attribution for Short Text Streams. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 727–730.

[2] Douglas Bagnall. 2015. Author Identification Using Multi-headed Recurrent Neural Networks. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*

[3] Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. 2019. Generalizing Unmasking for Short Texts. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 654–659.

[4] Kenneth A. Bollen. 1989. *Structural Equations with Latent Variables.* Wiley.

[5] Mohamed Amine Boukhaled and Jean-Gabriel Ganascia. 2014. *Probabilistic Anomaly Detection Method for Authorship Verification.* Springer International Publishing, Cham, 211–219.

[6] Daniel Castro Castro, Yaritza Adame Arcia, María Pelaez Brioso, and Rafael Muñoz Guillena. 2015. Authorship Verification, Average Similarity Analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*. INCOMA Ltd. Shoumen, BULGARIA, 84–90.

[7] Tommi Gröndahl and N. Asokan. 2019. Text Analysis in Adversarial Settings: Does Deception Leave a Stylistic Trace? *CoRR* abs/1902.08939 (2019). arXiv:1902.08939

[8] Oren Halvani, Lukas Graner, and Inna Vogel. 2018. Authorship Verification in the Absence of Explicit Features and Thresholds. In *Advances in Information Retrieval*, Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury (Eds.). Springer International Publishing, 454–465.

[9] Oren Halvani and Martin Steinebach. 2014. An Efficient Intrinsic Authorship Verification Scheme Based on Ensemble Learning. In *Ninth International Conference on Availability, Reliability and Security, ARES 2014, Fribourg, Switzerland, September 8-12, 2014.* Washington, DC, USA, 571–578.

[10] Oren Halvani, Christian Winter, and Lukas Graner. 2017. On the Usefulness of Compression Models for Authorship Verification. In *Proceedings of the 12th International Conference on Availability, Reliability and Security (ARES '17)*. ACM, New York, NY, USA, Article 54, 10 pages.

[11] Oren Halvani, Christian Winter, and Anika Pflug. 2016. Authorship Verification for Different Languages, Genres and Topics. *Digit. Investig.* 16, S (March 2016), S33–S43.

[12] Josué Gerardo Gutiérrez Hernández, José Casillas, Paola Ledesma, Gibran Fuentes Pineda, and Iván Vladimir Meza Ruíz. 2015. Homotopy Based Classification for Author Verification Task: Notebook for PAN at CLEF 2015. In *Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum, Toulouse, France, September 8-11, 2015.*

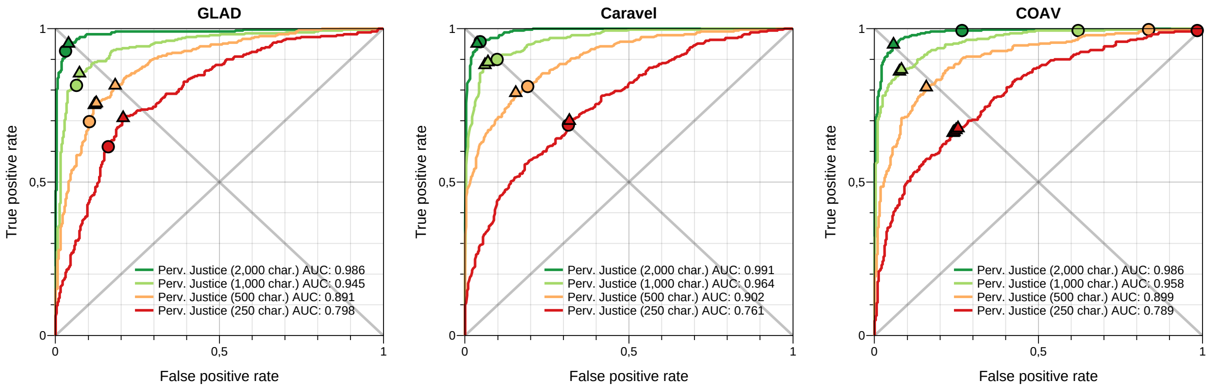


Figure 3: ROC curves for GLAD, Caravel and COAV (applied on the four corpora versions of C_{Perv}). The circles and triangles depict the current and maximum achievable c@1 values on the corpus, respectively. Note that Caravel's thresholds always lie along the EER-line.

[13] Ángel Hernández-Castañeda and Hiram Calvo. 2017. Author Verification Using a Semantic Space Model. *Computación y Sistemas* 21, 2 (2017).

[14] David I. Holmes. 1998. The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing* 13, 3 (1998), 111–117.

[15] Manuela Hürlimann, Benno Weck, Esther von den Berg, Simon Šuster, and Malvina Nissim. 2015. GLAD: Groningen Lightweight Authorship Detection. In *Working Notes of CLEF 2015 – Conference and Labs of the Evaluation forum, Toulouse, France, September 8–11, 2015*. 12.

[16] Magdalena Jankowska, Vlado Keselj, and Evangelos E. Milios. 2013. Proximity Based One-class Classification with Common N-Gram Dissimilarity for Authorship Verification Task Notebook for PAN at CLEF 2013. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*.

[17] Magdalena Jankowska, Evangelos E. Milios, and Vlado Keselj. 2014. Author Verification Using Common N-Gram Profiles of Text Documents. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, Jan Hajic and Junichi Tsujii (Eds.). ACL, 387–397.

[18] John Noecker Jr and Michael Ryan. 2012. Distractorless Authorship Verification. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)* (23-25), Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Istanbul, Turkey.

[19] Patrick Juola and Efstathios Stamatatos. 2013. Overview of the Author Identification Task at PAN 2013. In *Working Notes for CLEF 2013 Conference, Valencia, Spain, September 23-26, 2013*. 20.

[20] Mahmoud Khonji and Youssef Iraqi. 2014. A Slightly-Modified GI-Based Author-Verifier with Lots of Features (ASGALF). In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014*. 977–983.

[21] Mirco Kocher and Jacques Savoy. 2015. UniNE at CLEF 2015 Author Identification: Notebook for PAN at CLEF 2015. In *CLEF (Working Notes) (CEUR Workshop Proceedings)*, Vol. 1391. CEUR-WS.org.

[22] Moshe Koppel and Jonathan Schler. 2004. Authorship Verification as a One-Class Classification Problem. In *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004 (ACM International Conference Proceeding Series)*, Carla E. Brodley (Ed.), Vol. 69. ACM.

[23] Moshe Koppel and Yaron Winter. 2014. Determining if Two Documents are Written by the Same Author. *JASIST* 65, 1 (2014), 178–187.

[24] Tempestt J. Neal, Kalaivani Sundararajan, and Damon L. Woodard. 2018. Exploiting Linguistic Style as a Cognitive Biometric for Continuous Verification. In *2018 International Conference on Biometrics, ICB 2018, Gold Coast, Australia, February 20-23, 2018*. IEEE, 270–276.

[25] J. Olsson. 2008. *Forensic Linguistics: Second Edition: An Introduction To Language, Crime and the Law*. Bloomsbury Academic.

[26] Nektaria Potha and Efstathios Stamatatos. 2014. A Profile-Based Method for Authorship Verification. In *Artificial Intelligence: Methods and Applications: 8th Hellenic Conference on AI, SETN 2014, Ioannina, Greece, May 15–17, 2014. Proceedings*. Springer International Publishing, 313–326.

[27] Nektaria Potha and Efstathios Stamatatos. 2017. An Improved Impostors Method for Authorship Verification. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings*. 138–144.

[28] Nektaria Potha and Efstathios Stamatatos. 2018. Intrinsic Author Verification Using Topic Modeling. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence, SETN 2018, Patras, Greece, July 09-12, 2018*. ACM, 20:1–20:7.

[29] Martin Potthast, Matthias Hagen, and Benno Stein. 2016. Author Obfuscation: Attacking the State of the Art in Authorship Verification. In *Working Notes Papers of the CLEF 2016 Evaluation Labs (CEUR Workshop Proceedings)*, Vol. 1609. CLEF and CEUR-WS.org, 716–749.

[30] Martin Potthast, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2019. A Decade of Shared Tasks in Digital Text Forensics at PAN. In *Advances in Information Retrieval*, Leif Azzopardi, Benno Stein, Norbert Fuhr, Philipp Mayr, Claudia Hauff, and Djoerd Hiemstra (Eds.). Springer International Publishing, Cham, 291–300.

[31] Oxana Ye. Rodionova, Paolo Oliveri, and Alexey L. Pomerantsev. 2016. Rigorous and Compliant Approaches to One-Class Classification. *Chemometrics and Intelligent Laboratory Systems* 159 (2016), 89 – 96.

[32] Conrad Sanderson and Simon Guenter. 2006. Short Text Authorship Attribution via Sequence Kernels, Markov Chains and Author Unmasking: An Investigation. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP '06)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 482–491.

[33] Shachar Seidman. 2013. Authorship Verification Using the Impostors Method Notebook for PAN at CLEF 2013. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23-26, 2013*.

[34] Efstathios Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. *J. Am. Soc. Inf. Sci. Technol.* 60, 3 (March 2009), 538–556.

[35] Efstathios Stamatatos. 2013. On the Robustness of Authorship Attribution Based on Character N-Gram Features. *Journal of Law and Policy* 21 (01 2013), 421–439.

[36] Efstathios Stamatatos. 2017. Authorship Attribution Using Text Distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for the Computational Linguistics, EACL 2017, April 3-7, 2017, Valencia, Spain*. The Association for Computer Linguistics.

[37] Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2015. Overview of the Author Identification Task at PAN 2015. In *Working Notes of CLEF 2015 – Conference and Labs of the Evaluation forum, Toulouse, France, September 8–11, 2015*. 17.

[38] Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Benno Stein, Martin Potthast, Patrick Juola, Miguel A. Sánchez-Pérez, and Alberto Barrón-Cedeño. 2014. Overview of the Author Identification Task at PAN 2014. In *Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15–18, 2014*. 877–897.

[39] Efstathios Stamatatos, Nikos Fakotakis, and George K. Kokkinakis. 2000. Automatic Text Categorization in Terms of Genre and Author. *Computational Linguistics* 26, 4 (2000), 471–495.

[40] Benno Stein, Nedim Lipka, and Sven Meyer zu Eissen. 2008. Meta Analysis within Authorship Verification. In *19th International Workshop on Database and Expert Systems Applications (DEXA 2008), 1-5 September 2008, Turin, Italy*. IEEE Computer Society, 34–39.

[41] David Martinus Johannes Tax. 2001. *One-Class Classification: Concept Learning In the Absence of Counter-Examples*. Ph.D. Dissertation. Delft University of Technology.

[42] Cor J. Veenman and Zhenshi Li. 2013. Authorship Verification with Compression Features. In *Working Notes for CLEF 2013 Conference , Valencia, Spain, September 23–26, 2013*. 6.