# Author Obfuscation: Attacking the State of the Art in Authorship Verification[*]

Martin Potthast, Matthias Hagen, and Benno Stein

Bauhaus-Universität Weimar
<first name>.<last name>@uni-weimar.de

**Abstract** We report on the first large-scale evaluation of author obfuscation approaches built to attack authorship verification approaches: the impact of 3 obfuscators on the performance of a total of 44 authorship verification approaches has been measured and analyzed. The best-performing obfuscator successfully impacts the decision-making process of the authorship verifiers on average in about 47% of the cases, causing them to misjudge a given pair of documents as having been written by "different authors" when in fact they would have decided otherwise if one of them had not been automatically obfuscated. The evaluated obfuscators have been submitted to a shared task on author obfuscation that we organized at the PAN 2016 lab on digital text forensics. We contribute further by surveying the literature on author obfuscation, by collecting and organizing evaluation methodology for this domain, and by introducing performance measures tailored to measuring the impact of author obfuscation on authorship verification.

## 1 Introduction

The development of author identification technology has reached a point at which it can be carefully employed in the wild to resolve cases of unknown or disputed authorship. For a recent example, a state-of-the-art forensic software played a role in breaking the anonymity of J. K. Rowling, who published her book "The Cuckoo's Calling" under the pseudonym Robert Gailbraith in order to "liberate" herself from the pressure of stardom, caused by her success with the Harry Potter series.[1] Moreover, forensic author identification software is part of the toolbox of forensic linguists, who employ the technology on a regular basis to support their testimony in court as expert witnesses in cases where the authenticity of a piece of writing is important. Despite their successful application, none of the existing approaches has been shown to work flawless—a fact that is rooted in the complexity and the ill-posedness of the problem. All approaches have a likelihood of returning false decisions under certain circumstances, but the circumstances under which they do are barely understood. It is hence particularly interesting to analyze whether and how these circumstances can be controlled, since any form of control over the outcome of an author identification software bears the risk of misuse.

In fiction, a number of examples can be found where authors tried to remain anonymous, and where they, overtly or covertly, tried to imitate the writing style of others. In

[*] A summary of this report has been published as part of [92]

[1] http://languagelog.ldc.upenn.edu/nll/?p=5315

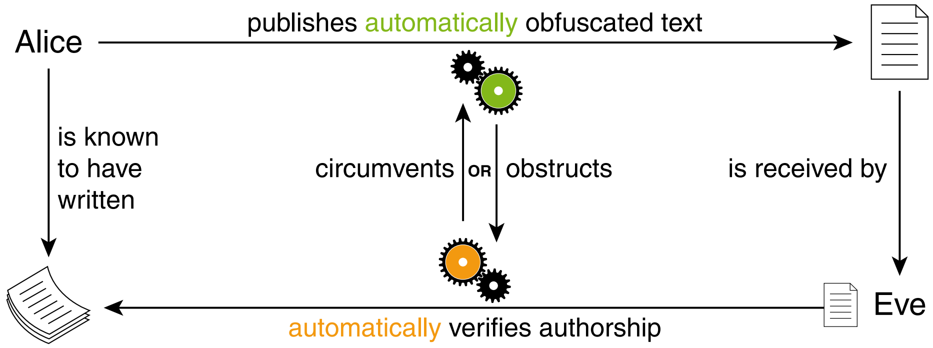

Alice
publishes automatically obfuscated text
is known
to have
written
circumvents
OR
obstructs
is received by
Eve
automatically verifies authorship

**Figure 1.** Schematic view of the opposed tasks of author masking and authorship verification.

fact, style imitation is a well-known learning technique in writing courses. But the question of whether humans are ultimately capable of controlling their own writing style so as to fool experts into believing they have not written a given piece of text, or even that someone else has, is difficult to answer based on observation alone: are the known cases more or less all there is, or are they just the tip of the iceberg (i.e., examples of unskilled attempts)? And, if the "expert" to be fooled is not a human but an author identification software, the rules are changed entirely. The fact that software is used to assist author identification increases the attack surface of investigations to spoil the decision-making process of the software. This is troublesome since the human operator of such a software may be ignorant of its flaws, and biased toward taking the software's output at face value instead of treating it with caution. After all, being convinced of the quality of a software is a necessary precondition to employing it to solve a problem.

At PAN 2016, we organized for the first time a shared task on author obfuscation to begin exploring the potential vulnerabilities of author identification technology. A number of interesting subtasks related to author obfuscation can be identified, from which we have selected that of author masking. This task is built on top of the task of authorship verification, a subtask of author identification, which was organized at PAN 2013 through PAN 2015 [47, 97, 98] (see Figure 1 for an illustration):

| **Authorship verification:** | | **Author masking:** |
|---|---|---|
| Given two documents,<br>decide whether they have been<br>written by the same author. | vs. | Given two documents by the same author,<br>paraphrase the designated one so that the<br>author cannot be verified anymore. |

The two tasks are diametrically opposed to each other: the success of a certain approach for one of these tasks depends on its "immunity" against the most effective approaches for the other. The two tasks are also entangled, since the development of a new approach for one of them should build upon the capabilities of existing approaches for the other. However, compared to authorship verification, author obfuscation in general (and author masking in particular) received little attention to date. A reason for this may be rooted in the fact that author masking requires (automatic) paraphrasing as a subtask, which poses a high barrier of entry to newcomers. To facilitate future research on both tasks, we contribute the following analyses and building blocks:

1. First-time large-scale evaluation of 44 state-of-the-art authorship verification approaches attacked by 3 author obfuscation approaches in 4 evaluation settings. This evaluation allows for judging both the feasibility of author obfuscation as well as the vulnerability of the state of the art in authorship verification. To cut a long story short, it turns out that even basic author obfuscation approaches have significant impact on many authorship verification approaches.
2. Proposal of performance measures to quantify the impact that obfuscation has on authorship analysis technology.
3. Survey of related work on author obfuscation, and a systematic review and organization of evaluation methodology for author obfuscation. In particular, we identify the three performance dimensions *safety*, *soundness*, and *sensibleness*, wherein an author obfuscation approach should excel before being considered fit for practical use; we detail how obfuscation approaches can be assessed today, and what may be useful in the future.
4. Organization of a shared task at PAN 2016 on author obfuscation to which the three obfuscators evaluated have been submitted. Moreover, we experiment with peer evaluation in shared tasks by inviting participants as well as interested third parties to co-evaluate the obfuscators with regard to the three aforementioned performance dimensions.

In what follows, Section 2 surveys the related work on author obfuscation, and Section 3 systematically reviews and organizes the corresponding evaluation methodology; here, the obfuscation impact measures are introduced. Section 4 reviews the obfuscation approaches that have been submitted to our shared task, and Section 5 reports on their evaluation against the state of the art in authorship verification, including the results of the outlined peer evaluation initiative.

## 2 Related Work

The literature on author obfuscation goes into three directions: (1) obfuscation generation, (2) obfuscation evaluation, and (3) obfuscation detection and reversal. This section reviews the contributions that have been made to date.

Obfuscation generation approaches divide into manual, computer-assisted, and automatic ones. The manual approaches include a study by Brennan and Greenstadt [11] who asked 12 laymen writers to mask their writing style and to imitate another author's style. The obtained results indicate that non-professional writers are capable of influencing their style to a point at which automatic authorship attribution performs no better than random. These results have been later replicated by Brennan *et al.* [10] who additionally employed crowdsourcing via Amazon's Mechanical Turk to increase the number of human subjects by 45 writers. Both datasets have been published as the (Extended) Brennan-Greenstadt Corpus.[2] Almishari *et al.* [3] also employ Amazon's

[2] https://psal.cs.drexel.edu/index.php/Main_Page

Mechanical Turk to obfuscate the reviews of 40 Yelp users, where up to 5 reviews per user were obfuscated many times over on Mechanical Turk, and up to 99 reviews per user were used to check whether the original user could still be identified among the 40 candidate users. After obfuscation, the success rate at attributing obfuscated reviews to the correct user dropped from 95% to 55% under a standard authorship attribution model employing words, POS tags, POS bigrams, as well as character bigrams and trigrams as features.

Anonymouth is the name of a tool developed by McDonald *et al.* [72, 73] which belongs to the computer-assisted approaches for obfuscation generation. It supports its users to manually obfuscate a given document by analyzing whether it can still be attributed to its original author after each revision, and by identifying which of the underlying authorship attribution model's features perform best to suggest parts of the given document that should be changed in order to maximize the next revision's impact on classification performance. An important component of a computer-assisted author obfuscation tools is its underlying analysis to determine the success of incremental text revisions; Kacmarcik and Gamon [50] focus on such a component. They also select the best-performing features and artificially change them in order to reduce attribution performance. This approach specifically targets and defeats Koppel and Schler's [61] Unmasking. Conceivably, the analysis components of both Anonymouth and Kacmarcik and Gamon can also be applied as part of a fully automatic obfuscation generator, however, this has not been attempted so far. Recently, Le *et al.* [65] introduced a semi-automatic obfuscation approach that supersedes the aforementioned ones in terms of safety against de-obfuscation attacks.

Among the first to propose automatic obfuscation generation were Rao and Rohatgi [91], who suggested the use of round-trip machine translation: the to-be-obfuscated document is translated to an intermediate language and the result then back to its original language. More than one intermediate languages may be used in a row before returning to the initial one. Supposedly, the translation round-trip distorts the writing style of the original's author sufficiently to confuse an authorship analysis. With the rise of machine translation systems, this approach has been studied many times to date, becoming a de-facto baseline for author obfuscation. However, the question whether this approach works is still undecided; some find that it does not perform well in terms of safety, soundness, and sensibleness of the obfuscated text [10, 13], whereas others do find some merits [3, 56]. At any rate, the use of machine translation for obfuscation may have limits since the existing systems must be treated as a black box and the obfuscation results can hardly be controlled. Besides machine translation, Khosmood and Levinson [57, 55] develop an obfuscation framework which operationalizes the imitation of an author's writing style by transforming the style of a given document iteratively via style-changing text operations toward the writing style of a set of target documents. Their approach builds on a style comparison component not unlike the aforementioned analysis component of Anonymouth, which determines the success of manual obfuscation and suggests where to make further manual text operations. The style comparison component controls which of a set of style-changing text operations are automatically applied to get closer to the target style. Khosmood [56, 54] further proposes a number of text operations at the sentence level, including active-to-passive transformation, diction improvement, abstractions, synonym replacement, simplification, and round-trip translation. Independently, Xu *et al.* [112] study within-language machine-translation as a way of transforming and imitating another author's style. Unlike the aforementioned machine learning approaches, here, the machine translation approach is specifically trained on texts from the source style and the target style so as to allow for accurate style paraphrases, rendering the system less of a black box than using round-trip translations for obfuscation. However, the approach is rendered less practical for its exceeding resource requirements in terms of samples of writing from the target style.

Obfuscation evaluation is about assessing the performance of an obfuscation approach. All of the aforementioned obfuscation approaches have been evaluated to some extent by their authors. However, little has been said to date on how an obfuscator should be evaluated; rather, evaluation setups have been created individually for each paper, rendering the reported results incomparable across papers. This is one of our primary contributions, and Section 3 introduces a comprehensive evaluation setup for author obfuscation under the three performance dimensions safety, soundness, and sensibleness. Nevertheless, our setup takes inspiration from the literature by collecting and organizing the previously employed evaluation procedures. For example, the most common evaluation approach under the safety dimension is to employ an existing authorship analysis approach to verify whether an obfuscated text can still be attributed to its original's author [3, 4, 11, 10, 13, 48, 49, 72]. Furthermore, some obfuscation approaches have been evaluated under the dimension of soundness, to ensure that an original's meaning does not get distorted by its obfuscation [56, 54], and sensibleness, to ensure that obfuscated texts are still human-readable [3]. On top of that, we report on the results of the first peer evaluation organized as part of our shared task. Participants and volunteers independently evaluated the obfuscation approaches submitted under the three aforementioned performance dimensions.

Finally, obfuscation detection and reversal is the task of deciding whether a given text has been obfuscated, and in that case, to undo the obfuscation in order to retrieve as much of the original text as possible. The possibility of reversing the effects of an (automatic) obfuscation threatens its safety. Afroz *et al.* [2] and Juola [46] have simultaneously shown that the texts of humans trying to mask their writing style or trying to imitate that of another author can be accurately detected as such. This gives rise to the detection of literary fraud, where an author may attempt to publish a text under the name of another author, imitating the latter's style. However, to simply remain anonymous the knowledge of the fact that a given text has likely been obfuscated is of no avail, since the obfuscation's measurable traces in a text do not necessarily give a clue about its original author. However, when it is possible to reverse the changes made via obfuscation, even if only partially, this puts users of the corresponding obfuscation approach at risk of being identified. In this regard, Le *et al.* [65] show that the semi-automatic obfuscation approaches of McDonald *et al.* [72] and Kacmarcik and Gamon [50] can be reversed and must therefore be considered unsafe.

## 3 Evaluating Author Obfuscation

This section collects and organizes the evaluation methodology for author obfuscation for the first time. We introduce three performance dimensions in which an author obfuscation approach should excel to be considered fit for practical use. Afterwards, we present an operationalization of each dimension based on both manual review as well as performance measurement.

### 3.1 The Three Dimensions of Obfuscation Evaluation

The performance of an author obfuscation approach obviously rests with its capability to achieve the goal of fooling forensic experts, be they software or human. However, this disregards writers and their target audience whose primary goal is to communicate, albeit safe from deanonymization. For them, the quality of an obfuscated text as well as whether its semantics are preserved are also important. In our survey of related work, these performance dimensions are often neglected or mentioned only in passing. Altogether, we call an obfuscation software

- **safe**, if its obfuscated texts can not be attributed to their original authors anymore,
- **sound**, if its obfuscated texts are textually entailed by their originals, and
- **sensible**, if its obfuscated texts are well-formed and inconspicuous.

These dimensions are orthogonal; an obfuscation software may meet each of them to various degrees of perfection. Each dimension keeps the others in check since trivial or flawed approaches stand no chance of achieving perfection in all three dimensions at the same time. The operationalization of each dimension, however, poses significant challenges in terms of resource requirements as well as scalability. In what follows, we review in detail the challenges involved in making each dimension measurable as well as how they have been operationalized in related work.

### 3.2 Evaluating Obfuscation Safety

The safety of an obfuscation approach depends on it withstanding three kinds of attacks: (1) manual authorship analyses, (2) automatic authorship analyses, and (3) deobfuscation attacks.

**(1) Manual authorship analyses** Manual authorship analyses can only be done by trained forensic linguists, so that this kind of analysis is expensive and therefore does not scale. Furthermore, it is unlikely that a forensic linguist would take part as human subject in an experiment to evaluate an obfuscation approach (not even anonymously), since the risks of suffering reputation damage from failing to beat it are too high, whereas they have little to gain otherwise. At any rate, beating one forensic linguist is insufficient to establish the safety of an obfuscation approach; the approach would have to be tested against a number of experts to raise sufficient confidence. Given these limitations, author obfuscation approaches are probably not going to be analyzed for safety against manual authorship analyses any time soon.

**(2) Automatic authorship analyses** Automatic authorship analyses are by comparison much more straightforward, practical, and scalable to be employed for obfuscation evaluation. Dozens of approaches have been proposed for the two author identification subtasks authorship attribution and authorship verification, so that evaluating an author obfuscation approach boils down to running the existing, pre-trained authorship analysis approaches against problem instances with and without obfuscated texts to observe the difference in performance. To be considered safe against automatic authorship analyses, an obfuscation approach should be able to defeat the best-performing authorship analysis approaches. In this connection, the related work on author obfuscation has employed a number of approaches for authorship attribution from the literature. Most safety evaluations rely on two basic feature sets, namely the so-called Basic 9 features (used in [10, 65, 72]) and the Writeprint features [1, 114] (used in [2, 3, 10, 65, 72]), on top of which Weka classifiers are applied. Some authors also evaluate their obfuscation approaches against other authorship analysis approaches such as the ones of Tweedie *et al.* [101], Clark and Hannon [22], and Koppel's Unmasking [61] as well as the freely available Signature Stylometric System[3] (see [11, 50]). Still, the number of authorship analysis approaches proposed to date by far surpasses the number of approaches employed in author obfuscation evaluations. The reason for this shortcoming may be found in the fact that hardly any implementations of the proposed authorship analysis approaches have surfaced to date: authors typically publish papers about their approaches but not their code base. Nevertheless, it must be conceded that the aforementioned approaches do not represent the landscape of authorship analysis approaches that have been proposed to date. To mitigate this problem for future author obfuscation evaluations, we have built two resources that allow for more comprehensive safety evaluations for author obfuscation for both authorship attribution and authorship verification: first, in [88], we report on the replication of 15 of the most influential authorship attribution approaches, the implementations of which are available as source code on GitHub.[4] Second, in our shared tasks on authorship verification at PAN 2013 to PAN 2015 [47, 97, 98], we report on the performances of a total of 49 pieces of software that have been submitted for evaluation. They are kept in working condition on the TIRA experimentation platform [32, 89], ready to be re-run on new datasets such as obfuscated versions of the test datasets of PAN 2013 to PAN 2015. In this paper, we employ them for the first time to evaluate three author obfuscation approaches for safety at scale.

*Measuring obfuscation impact.* When using existing authorship analysis approaches to measure the performance of an obfuscation approach, the impact obfuscation has on their classification performance is of interest. More specifically, for all problem instances where authors are correctly identified, the question is how many of them are not correctly identified, anymore, after obfuscation.

Let $d_a$ denote a document written by author $a$, and let $D_A$ denote a set of documents written by authors from the set of all authors $A$. We call $D = \langle d_u, D_A \rangle$ an instance of an authorship problem, where $d_u$ is a document written by an unknown author $u$ and $D_A$ is a set of documents of known authorship which may or may not contain a subset

[3] http://www.philocomp.net/humanities/signature.htm

[4] List of repositories: https://github.com/search?q=authorship+attribution+user:pan-webis-de

of documents $D_u \subseteq D_A$ written by the author $u \in A$ of $d_u$. If $D_A$ comprises only documents written by a single author $a$ so that $D_A = D_a$, $D$ is called an authorship verification problem, and otherwise an authorship attribution problem. If $D_A$ comprises documents written by more than one author, and if it can be guaranteed that one of them is the author of $d_u$, $D$ is a closed-class classification problem, and otherwise and open-class classification problem. Closed-class attribution problems can also be considered ranking problems, where the authors represented by disjunct subsets of $D_A$ are to be ranked so that the highest-ranking author $u$ is that of $d_u$.

We denote the universe of all authorship problem instances by $\mathcal{D}$, and $\gamma : \mathcal{D} \to A \cup \{\emptyset\}$ denotes the true mapping from $\mathcal{D}$ to the set of known authors $A$ for problems $D \in \mathcal{D}$ whose true author $u \in A$ of $d_u$ is among the candidates found in $D_A$, and to $\emptyset$ otherwise. An authorship analysis approach $y : \mathcal{D} \to A \cup \{\emptyset\}$ is an approximation of $\gamma$ that has been trained on $\mathcal{D}_{\mathrm{train}} \subset \mathcal{D}$. The extent to which the learned approximation of $y$ to $\gamma$ has been successful is evaluated using $\mathcal{D}_{\mathrm{test}} \subset \mathcal{D} \setminus \mathcal{D}_{\mathrm{train}}$ by checking whether $y$ returns answers matching those of $\gamma$ for the problem instances in $\mathcal{D}_{\mathrm{test}}$. A basic measure that is frequently applied to measure the performance of a given authorship analysis approach $y$ is its accuracy $\mathrm{acc}(y, \mathcal{D}_{\mathrm{test}})$ on a given test set $\mathcal{D}_{\mathrm{test}}$:

$$\mathrm{acc}(y, \mathcal{D}_{\mathrm{test}}) = \frac{|\{D \in \mathcal{D}_{\mathrm{test}} : y(D) = \gamma(D)\}|}{|\mathcal{D}_{\mathrm{test}}|}.$$

In this setting, an author obfuscation approach $o : \mathcal{D} \to \mathcal{D}$ maps the universe of authorship problems onto itself; here, $o(D) = \langle d_o, D_A \rangle$, where $d_o$ is the obfuscated version of $d_u \in D$ and $D_A \in D$ is kept as is. The true author of an obfuscated problem $o(D)$ is the same as without, say $\gamma(o(D)) = \gamma(D)$. But if an obfuscator works, an authorship analysis approach $y$ would return $y(o(D)) \neq \gamma(o(D))$. Let $o(\mathcal{D}) = \{o(D) : D \in \mathcal{D}\}$. A straightforward way to evaluate an author obfuscation approach $o$ is to apply it to the problem instances in $\mathcal{D}_{\mathrm{test}}$, measure the accuracy $\mathrm{acc}(y, o(\mathcal{D}_{\mathrm{test}}))$, and to calculate the performance delta:

$$\Delta_{\mathrm{acc}}(o, y, \mathcal{D}_{\mathrm{test}}) = \mathrm{acc}(y, o(\mathcal{D}_{\mathrm{test}})) - \mathrm{acc}(y, \mathcal{D}_{\mathrm{test}}). \tag{1}$$

However, in case $\mathcal{D}_{\mathrm{test}}$ comprises verification problems or open-class attribution problems, this measure takes into account problems where obfuscation need not be applied on the document of unknown authorship $d_u$, since its true author is not among the candidates $D_A$. Therefore, we consider only the subset $\mathcal{D}^{+}_{\mathrm{test}} \subseteq \mathcal{D}_{\mathrm{test}}$, which comprises only problem instances $D^{+} = \langle d_u, D_A \rangle$ where the true author $u$ of $d_u$ has written at least one document found in $D_A$. Measuring the accuracy of an authorship analysis approach $y$ on $\mathcal{D}^{+}_{\mathrm{test}}$ is equivalent to measuring recall, hence:

$$\mathrm{rec}(y, \mathcal{D}_{\mathrm{test}}) = \mathrm{acc}(y, \mathcal{D}^{+}_{\mathrm{test}}), \text{ and}$$

$$\Delta_{\mathrm{rec}}(o, y, \mathcal{D}_{\mathrm{test}}) = \mathrm{rec}(y, o(\mathcal{D}_{\mathrm{test}})) - \mathrm{rec}(y, \mathcal{D}_{\mathrm{test}}). \tag{2}$$

The domain of this measure is $[-1, 1]$, where $-1$ indicates the best possible performance of an obfuscator (i.e., flipping all decisions of an authorship analysis approach $y$ that makes no errors on unobfuscated texts), $0$ indicates the obfuscator has no impact, and a score greater than $0$ indicates the worst case, namely that the obfuscator somehow improves the classification performance of the given authorship analysis approach $y$ instead of decreasing it. In practice, the range of possible scores of $\varDelta_{\mathrm{rec}}$ is governed by the a priori performance $\mathrm{rec}(y, \mathcal{D}_{\mathrm{test}})$ of the authorship analysis approach $y$: $[-\mathrm{rec}(y, \mathcal{D}_{\mathrm{test}}), 1 - \mathrm{rec}(y, \mathcal{D}_{\mathrm{test}})]$. This means that $\varDelta_{\mathrm{rec}}$ does not reveal whether an obfuscation approach has accomplished everything it can against $y$. When achieving a score in the interval $(-1, 0)$ it remains unclear whether the obfuscator has flipped all of $y$'s correct authorship attributions, or not. To get an idea of the *relative impact* an obfuscator has on $y$'s recall, we apply the following normalizations dependent of $\varDelta_{\mathrm{rec}}$'s sign:

$$\mathrm{imp}(o, y, \mathcal{D}_{\mathrm{test}}) = \begin{cases} -\frac{\varDelta_{\mathrm{rec}}(o,y,\mathcal{D}_{\mathrm{test}})}{\mathrm{rec}(y,\mathcal{D}_{\mathrm{test}})} & \text{if } \varDelta_{\mathrm{rec}}(o, y, \mathcal{D}_{\mathrm{test}}) < 0, \\ -\frac{\varDelta_{\mathrm{rec}}(o,y,\mathcal{D}_{\mathrm{test}})}{1-\mathrm{rec}(y,\mathcal{D}_{\mathrm{test}})} & \text{else.} \end{cases} \tag{3}$$

The domain of this measure is, independent of the a priori performance of $y$, in the interval $[-1, 1]$, where $1$ indicates the best performance an obfuscator $o$ can reach by successfully obfuscating the problem instances where $y$ made a correct attribution before, and where $-1$ indicates that an obfuscator supports $y$ instead obstructing it by allowing it to correctly attribute problems it has not correctly attributed before. Note that we change the sign of the measure to emphasize that it captures obfuscator performance, and to allow for a more natural ordering. A potential drawback of measuring relative impact may be that the impact measured on authorship analysis approaches with a poor a priori performance may be overemphasized: for example, it may be much easier to flip the few correct attributions of an a priori poor-performing authorship analysis approach to earn a high relative impact than to flip the many correct attributions of a well-performing one. To mitigate this issue, a least-performance threshold under $\mathcal{D}_{\mathrm{test}}$ may be imposed that an authorship analysis approach $y$ must exceed to be considered attack-worthy by an obfuscator $o$.

As discussed at the outset, the performance of an obfuscation approach $o$ should not only be evaluated against a single authorship analysis approach $y$, but against as large a collection $Y$ of approaches as possible. After all, author obfuscation approaches are supposed to protect authors across the board of forensic analyses, and not just against specific specimen. Therefore, for a given collection of authorship analysis approaches $Y$, and for a given obfuscation approach $o$, we compute its average impact under $\mathcal{D}_{\mathrm{test}}$ as follows:

$$\mathrm{avg\ imp}(o, Y, \mathcal{D}_{\mathrm{test}}) = \frac{1}{|Y|} \sum_{y \in Y} \mathrm{imp}(o, y, \mathcal{D}_{\mathrm{test}}). \tag{4}$$

The average impact of different obfuscation approaches on a large number of authorship analysis approaches $Y$ allows for ranking among the obfuscation approaches in order to determine which of them performs best in terms of safety under $\mathcal{D}_{\mathrm{test}}$.

**(3) De-obfuscation attacks** De-obfuscation attacks include attempts to undo the effects an obfuscation approach has on a text, as well as analyses thereof that allow for a (semi-)accurate attribution of authorship despite the text having been obfuscated. The analytical nature of de-obfuscation attacks require a clear formulation of the assumptions under which de-obfuscation becomes possible, just like any proof of the safety of an obfuscation approach against de-obfuscation does. Such assumptions are sometimes enumerated as “attacker capabilities” or referred to as “threat model.” We propose to make the following general assumptions when analyzing an obfuscation approach’s safety against de-obfuscation:

- Kerckhoffs’ principle: the obfuscation approach used is public
- Data used during obfuscation is public except for the original text
- Seeds used to initialize pseudorandom number generators are secret
- No available meta data links an obfuscated text to its author

This way, the safety of an author obfuscation approach against de-obfuscation depends only on its merits at generating an irreversible obfuscation, and not on the fact that the approach or data used during obfuscation are secret. At the same time, if an obfuscation approach is deterministic and its text operations are easily recognizable and reversible to the original state, the approach must be considered unsafe and unfit for practical use. To date, the only systematic analysis of obfuscation approaches has been conducted by Le *et al.* [65] who show that the obfuscation approach of Kacmarcik and Gamon [50] can be completely reversed via backtracking, and that the safety against de-obfuscation of the approach implemented in Anonymouth by McDonald *et al.* [72] can be severely reduced in a closed-set attribution, increasing the probability of picking the correct author from 0.2 to 0.4. Given these results, the authors of a new obfuscation approach should always analyze its safety against de-obfuscation, whereas independent analyses of this kind are just as important to raise confidence. Unfortunately, de-obfuscation attacks elude performance measurement, since they will typically be tailored to the obfuscation approach attacked.

### 3.3 Evaluating Obfuscation Soundness

The soundness of an obfuscation approach depends on its ability to transfer the semantics of an original text to its obfuscated version. While the author of a text may value safety pretty high, the goal of writing a text is still to get a message across, which should remain untarnished by automatic obfuscation. A version of a text that conveys the same meaning as its original is called a paraphrase, and the goal of author obfuscation is to generate one under the constraint that the style of the original’s author is not recognizable, anymore (i.e., so that it is safe against forensic authorship analyses). Consequently, an author obfuscation approach must be evaluated whether and to what extent it generates paraphrases.

At the time of writing, research and development on paraphrase generation is mostly carried out at the sentence level; hardly any approaches exist that paraphrase at the paragraph level or even at the discourse level. Nevertheless, a paraphrase of a whole text may not only be done sentence-by-sentence, but it may also involve rearrangement of sentences, paragraphs, and entire lines of argumentative discourse. When evaluating soundness, a key challenge therefore is to trace the changes made by an obfuscator and to compare the parts of an obfuscated text to their unobfuscated counterparts in the original text. Given the possible non-linear changes that can be made during paraphrasing, an a posteriori comparison and judgment of the obfuscated text compared to its original is rendered difficult, since the apparent relations between an original text and its obfuscation may be ambiguous. While automatic obfuscation approaches can output which part of the original test went into generating which part of its obfuscation, this may not be as straightforward for manual obfuscations, unless the manual text editing operations are traced minutely as they happen.

Concerning evaluation, research on paraphrase generation relies almost unanimously on manual review. Nevertheless, inspired by the success of the well-known BLEU metric for machine translation evaluation, several performance measures for paraphrasing evaluation have been proposed, namely ParaMetric by Callison-Burch *et al.* [14], PEM by Liu *et al.* [68], PINC by Chen and Dolan [20], PARADIGM by Weese *et al.* [110], and APEM (for Korean) by Moon *et al.* [78]. All of these metrics are designed for sentence-wise paraphrase evaluation; the only evaluation of passage-length paraphrases has been reported by Burrows *et al.* [12] but no metrics have been proposed, whereas Xu *et al.* [112] propose three metrics that measure the quality of a paraphrase under the constraint that it is supposed to match a given author's style. The latter's metrics, however, require large corpora in both the original text's style and the style of the author to be imitated, which limits their applicability in many scenarios. In the literature for author obfuscation, also manual reviews with regard to soundness prevail: while the approaches to manual and semi-automatic obfuscation proposed do not require extensive soundness reviews, since the human subjects taking part in user studies can be trusted to produce sound paraphrases, the automatic obfuscation approach of Khosmood and Levinson [56] has been manually reviewed. The authors divide the generated paraphrases into the three basic categories of "correct", "passable", and "incorrect" paraphrases.

In the long run, the aforementioned measures may prove to be useful also for evaluating the soundness of author obfuscation approaches. At this time, however, we prefer manual review despite its disadvantages in terms of overhead, since the literature has not settled on a metric of choice, yet. Instead, to facilitate and scale manual reviews of obfuscated texts in comparison to their original texts, we develop a visual analytics tool for text comparison. The tool features various text comparison visualizations that assist manual soundness review: visualizations are applied to monitor the changes made by an author obfuscation approach. Figures 2, 3, and 4 show examples of the visualizations in action, contrasting the three approaches submitted to our shared task. As a brief explanation, the visualizations show the original text and the obfuscated text at the same time. The text is arranged in phrase, where each line either shows a large phrase the two compared texts have in common, or two stacked phrases where the two texts differ. This allows for quick comprehension of the effects an obfuscator has, as well as for quick judgments, thus significantly decreasing the time for manual review.

*Relaxing Soundness Constraints.* The constraint that both original and obfuscation possess the exact same semantics may be relaxed to some extent: it may be sufficient if the statements made in the obfuscated version of a text will be considered true under the presumption that the corresponding statements in the original are true (i.e., if the obfuscated statements follow logically from their respective originals). This is called textual entailment, and an obfuscated text would be considered entailed under the original text in such a situation. Relaxing the soundness constraint to allow for textual entailment opens a much wider space of possible obfuscations compared to paraphrases. A comprehensive survey of algorithms to recognize textual entailment as well as a series of corresponding shared tasks on this subject has been given by Dagan *et al.* [23]; these algorithms may serve as a basis for the development of new obfuscation soundness metrics. Another potential relaxation that arises from allowing textually entailed obfuscation is the possibility of allowing for summary obfuscation: for example, by summarizing a text or certain passages thereof, the original author's style may be significantly changed while maintaining at least the gist of the intended message. Conceivably, in some situations where author obfuscation is applied, summarization may be an acceptable route as means for safe obfuscation. In this connection, summaries are often textually entailed by the summarized text.

Finally, when relaxing the soundness constraints, one may also question how exact the original message has to be transferred into the obfuscated text. For example, in early machine translation systems as well as the ones deployed at scale today, not all translations are perfect, but the translation results are still useful to get a broad, and sometimes even a detailed understanding of a text in a foreign language. The same may apply to author obfuscation: to get a message across, it may not be necessary that every last detail gets transmitted correctly (as in properly paraphrased or textually entailed), but it may be sufficient for the reader to get a "readable" text whose message can be discerned with reasonable effort. While the goal of automatic author obfuscation should be to be perfectly sound and to generate actual paraphrases, early systems that do not come close to this requirement may still be useful in practice despite their deficiencies.

### 3.4 Evaluating Obfuscation Sensibleness

The sensibleness of an obfuscation approach depends on its ability (1) to create readable, ideally grammatically well-formed text, and (2) to hide the fact that a given obfuscated text has been obfuscated.

**(1) Obfuscation grammaticality** Next to being safe against forensic analyses and sound, another desired property of an obfuscated text is that it is well-formed in terms of grammar and that it fits into its genre. Automatic grammar checking has a long history in natural language processing and computer linguistics. Almost all approaches developed so far are designed to find specific error types in an ungrammatical text. For evaluation, however, the specific errors made by an obfuscator may be of less interest as opposed to deciding which parts of an obfuscated text are grammatical and which are not, for whatever reason. This latter task of *judging grammaticality* of a given piece of text is by far less often studied [19, 21, 25, 41, 87, 100, 108, 109, 111]: the proposed approaches to classifying a given sentence as being grammatical or not rely mostly on features extracted from statistical natural language parsers, whose confidence in their parsing results as well as features extracted from their resulting parse trees of grammatical and ungrammatical sentences are used to train linear classifiers at recognizing whether a given sentence is grammatical or not. The performances reported vary between 50% to more than 90% detection accuracy of ungrammatical sentences, dependent on the test dataset employed, whereas the results are largely incomparable for lack of a common baseline or a standardized benchmark dataset. Dependent on how successful grammaticality classification will become in the future, these approaches can be employed to build an effective performance measure for author obfuscation approaches. In fact, they may also be applied as components within an author obfuscation approach to a priori judge whether a given change will result in a grammatical obfuscated text. Until then, the only means left to evaluate a given author obfuscation approach is manual review.

Again, just like for obfuscation soundness, relaxing the criteria for obfuscation grammaticality may be reasonable in certain situations: for example, an obfuscator may be allowed to return ungrammatical text as long as it can be understood sufficiently well, or even as a means to mislead readers into thinking that an obfuscated text has been written by a second-language speaker of a given language. In this connection, manual review cannot be entirely avoided when evaluating author obfuscation approaches, since the line between what is acceptable and what is not is much more blurry.

**(2) Hiding obfuscation style** Although the safety of an obfuscated text and therefore the obfuscation approach used to generate it cannot rely on the fact that readers of the obfuscated text do not know that it has been obfuscated, hiding obfuscation is still an interesting side-goal of obfuscation generation. The inconspicuousness of an obfuscated text may serve as a first line of defense that forecloses detailed investigations. This is particularly important for automatic author obfuscation approaches that easily defeat automatic authorship analysis approaches but are vulnerable to manual authorship analyses by forensic experts. Conceivably, there are not enough forensic experts to analyze all texts that may be desirable to be analyzed, so that automatic authorship approaches will be applied to attain scale, whereas manual analyses will only be conducted in cases of doubt or suspicion. Avoiding to raise suspicion is therefore a worthwhile goal for an author obfuscation approach.

In the related work on author obfuscation, Afroz *et al.* [2] and Juola [46] have shown simultaneously, independently of one another that humans trying to mask their own style, and humans trying to imitate that of another author leave measurable traces in their writing that allow for automatically discriminating texts where authors have tried to alter their style from texts where authors have written in their genuine style (e.g., without making a conscious effort at altering it). The fact that humans leave such traces is no indication of whether automatic obfuscation approaches will do so as well, but it is very likely that they do. The question remains whether and how an automatic author obfuscation approach can be taught either to randomize the traces, or to blend in in a way so that the style of the obfuscated text cannot be distinguished, anymore, from the style of a single human writing genuinely. Evaluating this aspect of author obfuscation approaches will rely on applying the aforementioned approaches at detecting style deception to check whether they can be successfully lead astray, whereas manually reviewing for style deception is infeasible since subtle traces of obfuscation that are revealed only via statistical analyses may be lost on a human reviewer.

### 3.5 Obfuscation Efficiency

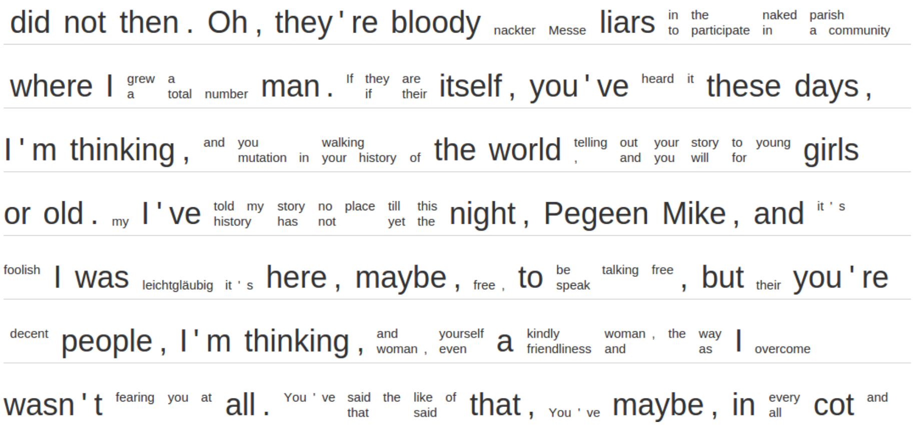
Since there are typically many alternative ways to paraphrase a given statement in order to obfuscate it, and even more so considering an entire text, the question arises which alternatives are better than others. One possible way to decide this question is to search for the alternative which is least different from the original but still achieves the goals of safety, soundness, and sensibleness. In this connection, Kacmarcik and Gamon [50] have proposed the "amount of work" as an efficiency measure for author obfuscation, namely the number of changes per 1000 words. They claim that their approach obfuscates a text with as little as 14 changes per 1000 words. Suppose two given obfuscation approaches $o_1$ and $o_2$ achieve sufficient safety, soundness, and sensibleness, where $o_1$ does so with the least possible amount of changes to the original text, and $o_2$ does so by making significantly more changes, which of the two is to be preferred? There is no straightforward answer to this question; while lazy approaches appeal by their efficiency, investing more work into generating an obfuscation may be worthwhile to attain an obfuscation that is not only safe against the state of the art but potentially also against future authorship analysis approaches that apply new forms of analyses. In this regard, alternative obfuscations that maximize the difference to the original text may be more interesting. However, maximizing the distance of an obfuscation to its original text under some style model may not be a strategy that is safe against de-obfuscation attacks in all situations. For example, Le *et al.* [65] show that in a closed-class attribution scenario, maximizing the distance, or moving the style towards the centroid of a given set of authors, provides sufficient information for a de-obfuscation attack. Altogether, while measuring obfuscation efficiency in terms of number of changes made to the original text per unit is interesting, and while it is also interesting to know how little work is necessary to achieve safety, soundness, and sensibleness today, this measure is insufficient to rank two obfuscation approaches.

## 4 Survey of Submitted Obfuscation Approaches

The three approaches submitted to our shared task follow three rather different strategies: round-trip translations, replacement of at most one frequent word per sentence, and style feature changes. While replacement of one word per sentence is a rather conservative strategy in that it changes the to-be-obfuscated text only slightly, the other two approaches change the text more substantially.

**Keswani *et al.*** The approach of Keswani *et al.* [52] is based on round-trip translation. Since access to the translation APIs of big commercial search engines like Google, Microsoft Bing, or Yandex is disallowed during the testing period of the author obfuscation task to prevent the test data from leaking, Keswani *et al.* employ the open source Moses SMT toolkit [60] trained on the Europarl corpus [59]. The original text is translated from English to German, the German text is translated to French, and the French text is translated back to English. The presumption of Keswani *et al.* is that the original text will be sufficiently changed during the translation to obfuscate its author.

As for the resulting text, our evaluation showed that hardly any human-readable text is produced (not even on a sentence level). One reason might be that the Europarl corpus is not particularly suited as a training corpus for the different genres of our test datasets. In the current form of Keswani *et al.*'s translation obfuscation, especially the soundness and the sensibleness suffer from severe errors that render the texts non-readable and not understandable by a human. Figure 2 shows the output for the problem instance 429.


did not then . Oh , they ' re bloody nackter Messe liars in to the participate naked in parish a community

where I grew a a total number man . If if they are their itself , you ' ve heard it these days ,

I ' m thinking , and you mutation in walking your history of the world telling , out and your you story will to for young girls

or old . my I ' ve told history my story has no not place till yet this the night , Pegeen Mike , and it ' s

foolish I was leichtgläubig it ' s here , maybe , free , to be speak talking free , but their you ' re

decent people , I ' m thinking , and woman , yourself even a kindly friendliness woman , and the way as I overcome

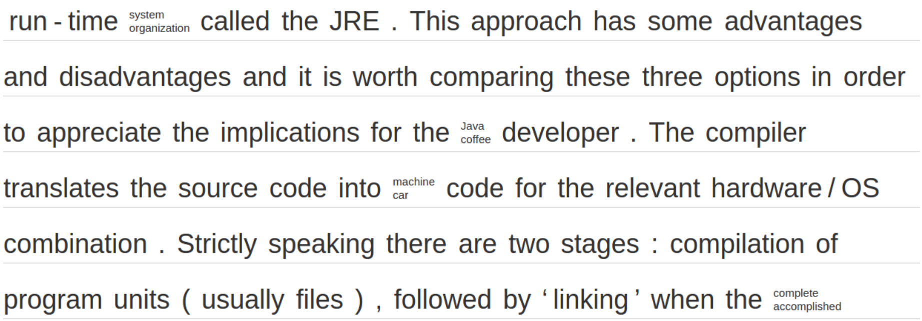
wasn ' t fearing you at all . You ' ve said that the like said of that , You ' ve maybe , in every all cot and

**Figure 2.** Difference visualization of the original text and the output of Keswani *et al.*'s obfuscation approach for problem instance 429 (for the differences there are two lines, original on top). There are many changes that make the text almost impossible to understand; even containing translation artifacts like the German words "nackter Messe" or "leichtgläubig" in the example.

Especially the "missing" grammaticality of the output substantially decreases sensibleness and soundness at the same time since the texts do not make sense at all. A couple of translation artifacts (remaining German or French words from the intermediate translations) would raise some suspicions even if the text was readable. However, do-it-yourself statistical machine translation will improve over time such that the text quality will get better with new translation approaches and larger corpora. In this respect, employing some of the best public translation approaches in the form of the proprietary big commercial search engine APIs might be interesting for the future although potentially difficult with the current evaluation setting at PAN.

**Mansoorizadeh *et al.*** The approach of Mansoorizadeh *et al.* [70] focuses on attacking the feature of word frequencies that are the basis of many verification approaches. The idea is to exchange some occurrences of the 200 most frequent words in the to-be-obfuscated text with synonyms obtained from WordNet [76] and scored based on their similarity to the original to-be-replaced word. For the most similar candidates, the commonness in the context of the to-be-replaced word is measured using a 4-gram language model trained on the Brown corpus [27]. Instead of the Brown corpus some bigger corpora allowing for longer $n$-grams might be worth a look at (e.g., the Google n-grams [9] underlying the Netspeak queries [90] that Stein *et al.* used for paraphrasing in acrostic generation [99]). A crucial point of Mansoorizadeh *et al.*'s approach is that it replaces at most one (!) word per sentence—rendering the paraphrasing rather conservative.

Since not that many changes are introduced overall, the texts in general remain as readable and grammatically correct as the original text and usually are fully understandable. Still, we spotted some issues, an example of which is shown in Figure 3. In problem instance 5 (a text about Java programming), the word "machine" that refers to a computer is consistently changed to "car," the word "Java" is sometimes changed to "coffee," etc. This makes the obfuscations less sensible and sound than in case of



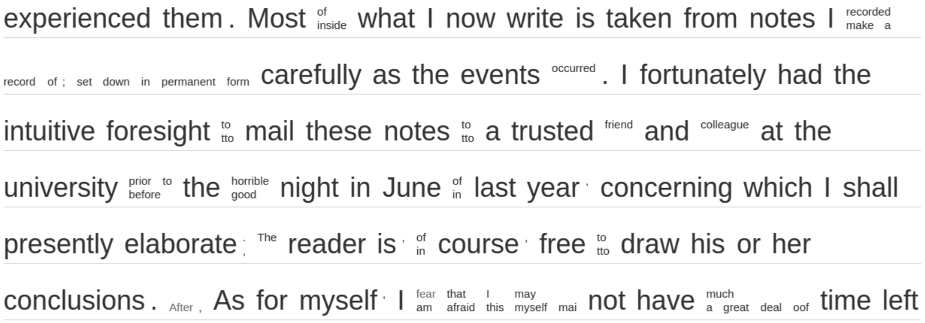
run - time system organization called the JRE . This approach has some advantages
and disadvantages and it is worth comparing these three options in order
to appreciate the implications for the Java coffee developer . The compiler
translates the source code into machine car code for the relevant hardware / OS
combination . Strictly speaking there are two stages : compilation of
program units ( usually files ) , followed by ' linking ' when the complete accomplished

**Figure 3.** Difference visualization of the original text and the output of Mansoorizadeh *et al.*'s obfuscation approach for problem instance 5 (for the differences there are two lines, original on top). Only few changes can be observed that still sometimes make the text less understandable (e.g., "machine" to "car" or "Java" to "coffee" in this example).

most other problem instances. Although the text quality is not reduced in most cases, the conservative paraphrasing also is the reason for a suboptimal safety score since the slight changes on average cannot obfuscate authorship very well. Allowing for some more changes per sentence and taking other than the most frequent words into account might be good future options that could be tested.

**Mihaylova *et al.*** The approach of Mihaylova *et al.* [75] targets a wide range of style-indicating features that are frequently used in author identification approaches: sentence length, punctuation, stop words, parts of speech, all caps, and individual word frequencies. As a background corpus (in the current approach formed by some books from Project Gutenberg and the PAN training datasets), the expected values of the features are computed. For a to-be-obfuscated text, the scores for the same feature types are calculated and then tried to be transformed more to the background average, applying a wide range of rule-based and random text operations. In particular, the operations employed are merging/splitting sentences, removing/inserting commas, removing stop words, improving/impairing spelling, exchanging words with synonyms, applying rules from the paraphrase database [30], lowercasing long, all-caps words, switching American and British spelling, inserting random functional words as noise, and replacing word shortenings, abbreviations, numbers, and possessive expressions.

In sum, the approach of Mihaylova *et al.* changes the text a lot since there are no restrictions on the number of changes per sentence. However, the context is usually not taken into account (i.e., whether a particular word or expression is frequent or common enough in the given context around the replacement). Thus, a lot of the changes look rather odd to a human—also the typos and randomly inserted words—and sometimes even change the meaning completely (e.g., "horrible night" changed to "good night" in problem instance 134, shown in Figure 4). Still, the soundness and sensibleness on average are slightly better than for the round-trip machine translation approach of Keswani *et al.* Also, the quality of the produced text might be improved a lot by not overdoing the spelling errors (i.e., not introducing a spelling error in every occur-



experienced them . Most of inside what I now write is taken from notes I recorded make a

record of ; set down in permanent form carefully as the events occurred . I fortunately had the

intuitive foresight to tto mail these notes to tto a trusted friend and colleague at the

university prior before to the horrible good night in June of in last year , concerning which I shall

presently elaborate ; The reader is , of in course , free to tto draw his or her

conclusions . After , As for myself , I fear am that afraid I this may myself mai not have much a great deal oof time left

**Figure 4.** Difference visualization of the original text and the output of Mihaylova *et al.*'s obfuscation approach for problem instance 134 (for the differences there are two lines, original on top). There are many changes that make the text more difficult to understand; examples are the exchange of "recorded" with two (!) longer explanations of the word in the first row, the insertion of random words like "After" in the last line, or the always incorrectly spelled word "tto."

rence of some word) and probably even more by taking context into account for word replacements that can increase the chance of a more common formulation resulting in more "meaningful" text.

# 5 Evaluation

We automatically evaluate the safety of the three obfuscation approaches against 44 authorship verifiers which have been submitted to the previous three shared tasks on authorship identification at PAN 2013 through PAN 2015, and we manually assess sensibleness and soundness of the obfuscated texts of each obfuscator.

## 5.1 Safety

Our safety evaluation is built with an eye on reproducibility, so that future evaluations of author obfuscators may be conducted with ease. In what follows, we detail our setup, the datasets used, and report on the results obtained.

**Evaluation Setup** The scale of our safety evaluation is made possible based on our long-term evaluation-as-a-service initiative [39], and the development of the corresponding cloud-based evaluation platform TIRA [32, 89].[5] TIRA facilitates software submissions for shared task competitions so that the organizers of a shared task can ask participants to submit their software instead of just its run output. At PAN, we have successfully invited software submissions for various shared tasks since 2012, all of which are still archived and available for reuse on TIRA. This is also the case for the past three editions of the shared task on authorship verification organized at PAN 2013 through PAN 2015. A total of 49 pieces of software have been submitted over all three years by as many research teams from all over the world, 44 of which were eligible for

[5] www.tira.io

our evaluation.[6] This collection of software represents the state of the art in authorship verification, implementing many different paradigms of tackling this task as well as hundreds of different features. The best-performing approaches of each year are those of Seidman [95] submitted to PAN 2013, Fréry *et al.* [29] submitted to PAN 2014 and Modaresi and Gross [77] submitted to PAN 2014 (which outperforms Fréry's approach on a different test dataset), and, the approach of Bagnall [5] submitted to PAN 2015. Seidman implements the Impostor's Method of Koppel and Winter [62], Fréry implements a "traditional" classification approach based on style-indicating features and linear classifiers, whereas Modaresi employs fuzzy clustering and Bagnall a multi-headed recurrent deep neural network instead of a linear classifier. The range of approaches implemented is too broad to be reviewed in detail here, but a complete survey can be found in the overview papers of the respective shared tasks in [47, 97, 98]. This collection of authorship verifiers is a unique resource for reproducible evaluations on authorship verification, and it forms a solid basis for the evaluation of author obfuscators at scale. Supported by TIRA, the total time to evaluate all 44 authorship verification on obfuscated versions of PAN's test datasets amounted to less than two man-weeks work time.

The participants of the shared task on author obfuscation, too, have been invited to submit their obfuscation approaches in the form of software to TIRA. This way, any newly submitted authorship verification software can be immediately evaluated for vulnerabilities against the submitted obfuscators, and vice versa.

**Evaluation Datasets** The test datasets on which our evaluation is based correspond to those used at the shared tasks on authorship verification at PAN 2013 to PAN 2015, covering a selection of genres:

- *PAN13.* Collection of English computer science textbook excerpts. Formulas and source code were removed. Problem instances comprise one document of unknown authorship and average 4 documents of known authorship. The training dataset comprises 10 problem instances at an average 1000 words per document, and the test dataset 30 problem instances at the same average document length.
- *PAN14 EE.* Collection of English essays written by English-as-a-second-language students at different language proficiency levels. Essays were divide into age-based clusters before forming verification problems. Problem instances comprise one document of unknown authorship and an average 2.6 documents of known authorship. The training dataset comprises 200 problem instances at an average 848 words per document, and the test dataset 200 problem instances at an average 833 words per document.
- *PAN14 EN.* Collection of English horror fiction novel excerpts from H.P. Lovecraft's "Cthulhu Mythos". The genre is selected very narrowly to strip away any cross-topic or cross-genre effects typically found in other collections. Many unfamiliar terms are found in these documents, creating strong indicators of shared topic and style. Most texts' themes have a strong negative coloring. Lovacraft's original

[6] Five approaches had to be excluded: the three approaches submitted by Halvani *et al.* [36, 37, 38] have been deleted from TIRA at the request of the authors; the approach by Veenman and Li [106] was dysfunctional, and the approach of Vartapetiance and Gillam [105] was not submitted for evaluation on the English portion of the PAN test datasets.

**Table 1.** Safety evaluation of three obfuscators against sets of 26-36 authorship verification approaches submitted to PAN 2013 through PAN 2015. Column group "PAN measures" shows the average performance delta on the evaluation measures ROC AUC, C@1, and the final score AUC · C@1 applied at PAN. Row groups show safety under four English PAN test datasets, rows within row groups are ordered by average impact ($\mathrm{avg\ imp}$).

| **Obfuscator** | | **Verifier** | **Dataset** | | **PAN measures** | | | **Section 3** | | **Impact** |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\lvert Y \rvert$ | $\mathcal{D}_{\mathrm{test}}$ | $\lvert \mathcal{D}^{+}_{\mathrm{test}} \rvert$ | $\Delta_{\mathrm{AUC}}$ | $\Delta_{\mathrm{C@1}}$ | $\Delta_{\mathrm{final}}$ | $\Delta_{\mathrm{acc}}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{avg\ imp}$ |
| Mihaylova *et al.* | [75] | 36 | PAN13 | 14 | **-0.1066** | **-0.0759** | **-0.1030** | **-0.1389** | **-0.2778** | **0.4690** |
| Keswani *et al.* | [52] | 36 | PAN13 | 14 | -0.0908 | -0.0695 | -0.0940 | -0.1148 | -0.2361 | 0.4245 |
| Mansoorizadeh *et al.* | [70] | 36 | PAN13 | 14 | -0.0422 | -0.0254 | -0.0392 | -0.0463 | -0.0933 | 0.1442 |
| Mihaylova *et al.* | [75] | 26 | PAN14 EE | 100 | **-0.1305** | **-0.1088** | **-0.1144** | **-0.1229** | **-0.2304** | **0.4891** |
| Keswani *et al.* | [52] | 26 | PAN14 EE | 100 | -0.1085 | -0.0870 | -0.0960 | -0.0975 | -0.1873 | 0.4058 |
| Mansoorizadeh *et al.* | [70] | 26 | PAN14 EE | 100 | -0.0514 | -0.0463 | -0.0473 | -0.0577 | -0.1038 | 0.2512 |
| Mihaylova *et al.* | [75] | 36 | PAN14 EN | 100 | **-0.1613** | **-0.1050** | **-0.1260** | **-0.1456** | **-0.2456** | **0.4750** |
| Keswani *et al.* | [52] | 36 | PAN14 EN | 100 | -0.1020 | -0.0704 | -0.0845 | -0.1074 | -0.1783 | 0.3769 |
| Mansoorizadeh *et al.* | [70] | 36 | PAN14 EN | 100 | -0.0579 | -0.0408 | -0.0493 | -0.0665 | -0.0958 | 0.2345 |
| Mihaylova *et al.* | [75] | 35 | PAN15 | 250 | **-0.1074** | **-0.0927** | **-0.1090** | **-0.1050** | **-0.2009** | **0.3649** |
| Keswani *et al.* | [52] | 35 | PAN15 | 250 | -0.0599 | -0.0468 | -0.0612 | -0.0645 | -0.1298 | 0.2543 |
| Mansoorizadeh *et al.* | [70] | 35 | PAN15 | 250 | -0.0375 | -0.0339 | -0.0420 | -0.0502 | -0.0994 | 0.1952 |

writings as well as modern fan-fiction form part of the dataset. Problem instances comprise two documents. The training dataset comprises 100 problem instances at an average 3138 words per document, and the test dataset 200 problem instances at an average 6104 words per document.

- *PAN15.* Collection of English dialog lines from plays, excluding speaker names, stage directions, lists of characters, etc. Problem instances comprise two documents; cases where the author of the two documents match are from different plays of the same author. The training dataset comprises 100 problem instances at an average 366 words per document, and the test dataset 500 problem instances at an average 536 words per document.

All datasets have a balanced ratio of problem instances where the author of the documents of known authorship is the same as that of the document of unknown authorship to problem instances where this is not the case. The training datasets were release to participants so that they could develop their obfuscation approaches against it. The test datasets were kept private. Participants who made a successful software submission could run their software on TIRA against the test datasets, while TIRA prevents any direct access of participants to the test datasets hosted there, and takes precautions against data leaks. This way, any optimization of approaches against test datasets is rendered impossible.

**Evaluation Results** Table 1 shows the results of our safety evaluation of the three submitted author obfuscation approaches against at total of 44 authorship verification approaches on the aforementioned four PAN evaluation datasets. The best-performing approach across all performance measures and across all datasets is the author obfuscator of Mihaylova *et al.* [75]. In terms of average impact ($\mathrm{avg\ imp}$), it manages to flip between 46% and 49% of the correct authorship attributions of the verifiers on the datasets PAN13, PAN14 EE, and PAN14 EN, but only about 36% on the PAN15 dataset. While the approach of Keswani *et al.* [52] achieves an average impact close

**Table 2.** "Hall of shame" for the three author obfuscation approaches: the table lists all authorship verifiers that were supported instead of obstructed when applying the obfuscation to a given test dataset. The support is measured as positive performance delta and negative relative impact, respectively. Many of these discrepancies can be explained and dismissed for various reasons outlined in Section 5.1; for three cases, however, no explanation could be found (marked by a **?**).

| **Verifier** | **Dataset** | **PAN measures** | | | **Section 3** | | | | **Impact** | **Dismissed** |
|---|---|---|---|---|---|---|---|---|---|---|
| Team [Reference] | $\mathcal{D}_{\text{test}}$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | acc | $\Delta_{\text{acc}}$ | rec | $\Delta_{\text{rec}}$ | imp | Reasons |
| *Mihaylova* et al. *[75] supports the following verifiers:* | | | | | | | | | | |
| Mechti [74] | PAN13 | 0.09 | 0.08 | 0.10 | 0.60 | -0.03 | 0.50 | 0.14 | -0.29 | 3 |
| Layton [64] | PAN14 EE | 0.07 | -0.21 | -0.09 | 0.58 | 0.06 | 0.72 | 0.11 | -0.39 | 2 |
| Kern [51] | PAN14 EN | 0.27 | -0.01 | 0.13 | 0.57 | 0.02 | 0.92 | 0.03 | -0.38 | 2 |
| Kocher [58] | PAN14 EN | 0.02 | -0.11 | -0.06 | 0.64 | 0.05 | 0.70 | 0.10 | -0.33 | **?** |
| Maitra [69] | PAN14 EN | -0.53 | -0.12 | -0.36 | 0.71 | -0.24 | 0.65 | 0.19 | -0.54 | 4 |
| Mechti [74] | PAN14 EN | -0.04 | 0.00 | -0.02 | 0.61 | -0.03 | 0.63 | 0.02 | -0.05 | 1 |
| Mechti [74] | PAN15 | 0.02 | -0.01 | 0.01 | 0.51 | -0.02 | 0.26 | 0.01 | -0.01 | 1, 2 |
| van Dam [102] | PAN15 | 0.00 | -0.01 | -0.00 | 0.59 | -0.01 | 0.60 | 0.00 | -0.01 | 1, 2 |
| *Keswani* et al. *[52] supports the following verifiers:* | | | | | | | | | | |
| Bartoli [7] | PAN13 | 0.14 | 0.07 | 0.11 | 0.67 | 0.03 | 0.36 | 0.07 | -0.11 | 3 |
| Maitra [69] | PAN13 | 0.09 | 0.07 | 0.09 | 0.60 | 0.03 | 0.29 | 0.07 | -0.10 | 3 |
| Mechti [74] | PAN13 | 0.11 | 0.02 | 0.07 | 0.60 | 0.03 | 0.50 | 0.14 | -0.29 | 3 |
| Layton [64] | PAN14 EE | 0.09 | -0.23 | -0.10 | 0.58 | 0.08 | 0.72 | 0.15 | -0.54 | 2 |
| Bartoli [7] | PAN14 EN | 0.01 | -0.02 | -0.01 | 0.62 | 0.01 | 0.77 | 0.01 | -0.04 | 1 |
| Maitra [69] | PAN14 EN | -0.21 | -0.14 | -0.21 | 0.71 | -0.20 | 0.65 | 0.26 | -0.74 | 4 |
| Moreau [80] | PAN14 EN | 0.03 | 0.03 | 0.04 | 0.63 | 0.01 | 0.72 | 0.01 | -0.04 | 1 |
| Kocher [58] | PAN15 | 0.01 | -0.01 | 0.00 | 0.70 | 0.00 | 0.60 | 0.01 | -0.02 | 1 |
| Mechti [74] | PAN15 | -0.00 | -0.01 | -0.01 | 0.51 | -0.01 | 0.26 | 0.00 | -0.01 | 1, 2 |
| Pacheco [83] | PAN15 | 0.04 | 0.02 | 0.04 | 0.71 | 0.01 | 0.59 | 0.02 | -0.06 | 1 |
| van Dam [102] | PAN15 | 0.00 | -0.00 | -0.00 | 0.59 | -0.00 | 0.60 | 0.04 | -0.10 | 2 |
| *Mansoorizadeh* et al. *[70] supports the following verifiers:* | | | | | | | | | | |
| Bagnall [5] | PAN13 | 0.01 | -0.02 | -0.00 | 0.80 | 0.07 | 0.93 | 0.07 | -1.00 | 3 |
| Harvey [40] | PAN13 | 0.02 | 0.00 | 0.01 | 0.60 | 0.03 | 0.50 | 0.07 | -0.14 | 3 |
| Jayapal [45] | PAN13 | 0.00 | 0.03 | 0.02 | 0.60 | 0.03 | 0.36 | 0.07 | -0.11 | 3 |
| Moreau [81] | PAN13 | 0.00 | 0.03 | 0.02 | 0.73 | 0.03 | 0.71 | 0.07 | -0.25 | 3 |
| Castillo [17] | PAN14 EN | 0.01 | 0.01 | 0.01 | 0.62 | 0.01 | 0.63 | 0.02 | -0.05 | 1 |
| Gutierrez [35] | PAN14 EN | 0.06 | -0.01 | 0.02 | 0.56 | 0.03 | 0.41 | 0.01 | -0.02 | 1, 2 |
| Kocher [58] | PAN14 EN | -0.01 | -0.03 | -0.02 | 0.64 | 0.01 | 0.70 | 0.01 | -0.03 | 1 |
| Maitra [69] | PAN14 EN | -0.18 | -0.14 | -0.20 | 0.71 | -0.19 | 0.65 | 0.29 | -0.83 | 4 |
| Mechti [74] | PAN14 EN | 0.01 | 0.00 | 0.01 | 0.61 | 0.00 | 0.63 | 0.06 | -0.16 | **?** |
| Modaresi [77] | PAN14 EN | 0.05 | 0.04 | 0.06 | 0.72 | 0.04 | 0.68 | 0.07 | -0.22 | **?** |
| Bartoli [7] | PAN15 | 0.03 | 0.03 | 0.04 | 0.58 | 0.00 | 0.87 | 0.01 | -0.06 | 1, 2 |
| Castillo [17] | PAN15 | 0.00 | 0.00 | 0.00 | 0.64 | 0.00 | 0.83 | 0.00 | -0.02 | 1 |
| Harvey [40] | PAN15 | -0.00 | 0.01 | 0.00 | 0.62 | 0.00 | 0.70 | 0.01 | -0.03 | 1 |
| Pimas [85] | PAN15 | 0.00 | 0.00 | 0.00 | 0.51 | 0.00 | 0.56 | 0.00 | -0.01 | 1, 2 |

to Mihaylova *et al.*'s on the former two datasets, its performance is much worse on the latter two. The approach of Mansoorizadeh *et al.* [70] achieves about half the average impact of Mihaylova *et al.*'s approach at best. For completeness sake, Table 1 also reports the performance deltas obtained for accuracy and recall as described in Section 3, as well as the performance deltas the obfuscators achieve with respect to the performance measures applied at PAN's shared tasks on authorship verification. The performance values are numerically much smaller, but their trend is the same as that of the average impact measure.

However, while the averaged performance deltas and the average impact all point into the direction that obfuscation approaches decrease the performance of authorship verification approaches, this is not unanimously the case. Tables 3, 4, and 5 show the performances when applying the individual authorship verification approaches on the four respective test datasets, each one obfuscated once by the approaches of Mihaylova *et al.* (Table 3), Keswani *et al.* (Table 4), and Mansoorizadeh *et al.* (Table 5). Notably, all obfuscators appear to *improve* a number of verifiers instead of obstructing them. Table 2 collects all cases where an authorship verification approach benefits from obfuscation. However, not all of these cases spell out failure for the obfuscation approaches:

- *Reason 1.* Negative impacts greater than -0.1 can be explained by random variation of the authorship verification approach.
- *Reason 2.* Authorship verification approaches which have an a priori performance of less than $0.6\ \mathrm{acc}$ may perhaps not be trusted to make reasonable decisions in the face of text alterations.
- *Reason 3.* High improvements on the PAN13 dataset are easily obtained by flipping only a few decisions, since the dataset is very small compared to the others.
- *Reason 4.* In three cases, the high negative obfuscation impacts on Maitra's [69] approach go along with a high negative accuracy performance delta ($\Delta_{\mathrm{acc}}$), which indicates that obfuscation has caused this verifier to shift a high number of its decisions from "different author" to "same author," but also on problem instances where no obfuscation has been applied (i.e., problems whose documents have indeed been written by different authors). This suggests that the decisions of Maitra's approach are not independent of one another but relative to decisions on other problem instances, which must be considered erroneous behavior.

These reasons notwithstanding, some cases remain where a priori well-performing obfuscators significantly improve after obfuscation, and the question remains why this is the case. It remains unclear whether these are random fluctuations which are to be expected when applying many authorship verification approaches implemented by as many different people onto obfuscated datasets, or whether something like this can be expected in practice as well. After applying all these rules of the thumb, only three cases remain unexplained, namely the negative obfuscation impact of Mihaylova *et al.*'s obfuscator on Kocher's verifier on the dataset PAN14 EN, and that of Mansoorizadeh *et al.*'s obfuscator on Mechti's verifier and Modaresi's verifier on the dataset PAN14 EN. Since in all three cases, the PAN14 EN dataset is involved, it's characteristics may cause this behavior, but this is just a speculation.

Altogether, we draw three conclusions from these results: (1) even basic author obfuscation can already achieve some degree of safety against state-of-the-art automatic forensic authorship analyses, (2) the state of the art in authorship verification is extremely vulnerable to obfuscation, and (3) obfuscation approaches must be evaluated in as many different situations as possible to identify odd behavior of both the obfuscator as well as authorship verifiers. The latter conclusion is particularly important since it forces us to now take the effectiveness of authorship verification technology reported in the literature with a grain of salt, especially in adversarial scenarios: its application in court is rendered doubtful, since texts of disputed authorship may have been tampered with to influence the court's decision.

### 5.2 Sensibleness and Soundness

A human assessor skimmed through a random subset of ten obfuscated texts for each approach using the aforementioned visual analytics tool outlining the differences of the original and obfuscated texts (some example screenshots are given in Section 4). The

**Table 3.** Safety evaluation of the obfuscator of Mihaylova *et al.* [75]. Each table shows the performances and performance deltas of various authorship verification approaches submitted to PAN 2013 through PAN 2015 when run on test datasets that have been obfuscated by this obfuscator. Verifiers that failed to process a dataset (e.g., for being incompatible or not scalable) have been omitted from the tables. Verifiers whose optimal classification threshold $\tau$ that maximizes classification accuracy $\mathrm{acc}$ on the unobfuscated test dataset turned out to be negative or positive infinity (i.e., marking all problem instances “same author” or “different author”, respectively) were omitted from forming the average performances reported in Table 1.

**PAN 2013 test dataset**

| Verifier | | | PAN measures | | | Section 3 | | | | Impact |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | 0.478 | 0.04 | 0.07 | 0.09 | 0.80 | -0.07 | 0.93 | 0.00 | 0.00 |
| Bartoli | [7] | 0.647 | -0.11 | 0.03 | -0.04 | 0.67 | -0.13 | 0.36 | -0.29 | 0.80 |
| Bobicev | [8] | 0.5144 | -0.29 | -0.20 | -0.26 | 0.67 | -0.17 | 0.50 | -0.36 | 0.71 |
| Castillo | [17] | 0.4 | -0.10 | 0.00 | -0.04 | 0.53 | -0.13 | 0.36 | -0.29 | 0.80 |
| Castro | [18] | 0.9 | -0.04 | -0.03 | -0.06 | 0.93 | -0.03 | 1.00 | -0.07 | 0.07 |
| Feng | [24] | 0.46 | -0.22 | -0.27 | -0.30 | 0.77 | -0.30 | 0.79 | -0.64 | 0.82 |
| Fratila | [28] | 0.38 | -0.41 | -0.27 | -0.33 | 0.67 | -0.33 | 0.71 | -0.71 | 1.00 |
| Fréry | [29] | 0.333 | -0.21 | -0.13 | -0.18 | 0.63 | -0.20 | 0.57 | -0.43 | 0.75 |
| Ghaeini | [31] | 0.46 | -0.42 | -0.24 | -0.41 | 0.80 | -0.30 | 0.71 | -0.64 | 0.90 |
| Gómez-Adorno | [33] | inf | 0.00 | 0.00 | 0.00 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Grozea | [34] | 0.05 | 0.02 | 0.03 | 0.02 | 0.53 | 0.00 | 1.00 | 0.00 | 0.00 |
| Gutierrez | [35] | 0.8182 | 0.02 | 0.06 | 0.06 | 0.77 | -0.03 | 0.79 | -0.07 | 0.09 |
| Harvey | [40] | 0.001 | -0.19 | 0.00 | -0.10 | 0.60 | -0.17 | 0.50 | -0.36 | 0.71 |
| Hürlimann | [42] | 0.6394 | -0.25 | -0.17 | -0.24 | 0.70 | -0.27 | 0.36 | -0.29 | 0.80 |
| Jankowska | [43] | 0.59 | -0.15 | -0.07 | -0.16 | 0.80 | -0.27 | 0.64 | -0.57 | 0.89 |
| Jankowska | [44] | 0.615 | -0.16 | -0.10 | -0.19 | 0.80 | -0.20 | 0.64 | -0.43 | 0.67 |
| Jayapal | [45] | 1 | 0.00 | 0.00 | 0.00 | 0.60 | 0.00 | 0.36 | 0.00 | 0.00 |
| Kern | [51] | 0.5 | 0.15 | -0.20 | -0.02 | 0.57 | 0.00 | 0.07 | 0.00 | 0.00 |
| Khonji | [53] | 0.444 | -0.34 | -0.32 | -0.43 | 0.80 | -0.30 | 0.93 | -0.64 | 0.69 |
| Kocher | [58] | 0.484 | -0.08 | -0.16 | -0.16 | 0.67 | -0.03 | 1.00 | -0.07 | 0.07 |
| Layton | [64] | inf | 0.15 | -0.23 | 0.02 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Layton | [63] | 0.7057 | -0.08 | -0.13 | -0.13 | 0.67 | -0.13 | 0.29 | -0.29 | 1.00 |
| Ledesma | [66] | inf | 0.00 | -0.13 | -0.07 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Maitra | [69] | 0.8 | 0.04 | 0.07 | 0.06 | 0.60 | -0.03 | 0.29 | -0.07 | 0.25 |
| Mayor | [71] | 0.1 | -0.14 | -0.09 | -0.15 | 0.73 | -0.10 | 0.79 | -0.36 | 0.45 |
| Mechti | [74] | 0.469 | 0.09 | 0.08 | 0.10 | 0.60 | -0.03 | 0.50 | 0.14 | -0.29 |
| Modaresi | [77] | 0.392 | -0.06 | -0.07 | -0.06 | 0.57 | -0.13 | 0.79 | -0.29 | 0.36 |
| Moreau | [81] | 1 | 0.00 | -0.30 | -0.15 | 0.73 | -0.30 | 0.71 | -0.64 | 0.90 |
| Moreau | [80] | 0.6215 | -0.28 | 0.00 | -0.13 | 0.70 | -0.27 | 0.93 | -0.43 | 0.46 |
| Nikolov | [82] | 0.448 | -0.15 | 0.00 | -0.08 | 0.60 | -0.13 | 0.29 | -0.29 | 1.00 |
| Pacheco | [83] | 0.7223 | -0.06 | 0.00 | -0.05 | 0.60 | -0.23 | 0.71 | -0.50 | 0.70 |
| Petmanson | [84] | 0.59 | -0.39 | -0.20 | -0.32 | 0.73 | -0.20 | 0.57 | -0.43 | 0.75 |
| Sari | [93] | 0.546 | -0.02 | 0.00 | -0.01 | 0.53 | -0.13 | 0.93 | -0.29 | 0.31 |
| Satyam | [94] | 0.423 | 0.02 | 0.07 | 0.03 | 0.53 | 0.00 | 1.00 | 0.00 | 0.00 |
| Seidman | [95] | 1 | -0.01 | 0.00 | -0.01 | 0.77 | -0.03 | 0.71 | -0.07 | 0.10 |
| Solórzano | [96] | 0.812 | -0.15 | -0.10 | -0.11 | 0.57 | -0.17 | 0.64 | -0.36 | 0.56 |
| van Dam | [102] | 1 | 0.00 | -0.07 | -0.03 | 0.60 | -0.07 | 0.57 | -0.07 | 0.13 |
| Vartapetiance | [103] | inf | 0.00 | -0.40 | -0.20 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vartapetiance | [104] | inf | 0.00 | 0.00 | 0.00 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vilarino | [107] | 1 | 0.00 | -0.03 | -0.02 | 0.67 | -0.03 | 0.36 | -0.07 | 0.20 |
| Zamani | [113] | 0.997 | 0.07 | 0.00 | 0.05 | 0.80 | -0.07 | 0.64 | -0.14 | 0.22 |

**PAN 2014 EN test dataset**

| Verifier | | | PAN measures | | | Section 3 | | | | Impact |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | 0.418 | -0.32 | -0.28 | -0.32 | 0.70 | -0.29 | 0.81 | -0.28 | 0.35 |
| Bartoli | [7] | 0.613 | -0.34 | -0.04 | -0.18 | 0.62 | -0.26 | 0.77 | -0.52 | 0.68 |
| Bobicev | [8] | 0.3406 | -0.24 | -0.05 | -0.15 | 0.57 | -0.19 | 0.42 | -0.37 | 0.88 |
| Castillo | [17] | 1 | -0.42 | -0.31 | -0.32 | 0.62 | -0.31 | 0.63 | -0.62 | 0.98 |
| Castro | [18] | 0.8 | -0.31 | -0.24 | -0.24 | 0.62 | -0.22 | 0.76 | -0.44 | 0.58 |
| Feng | [24] | 0.59 | -0.58 | -0.41 | -0.36 | 0.64 | -0.46 | 0.88 | -0.87 | 0.99 |
| Fréry | [29] | 1 | -0.22 | -0.16 | -0.19 | 0.61 | -0.22 | 0.69 | -0.43 | 0.62 |
| Gómez-Adorno | [33] | 1 | -0.01 | -0.01 | -0.01 | 0.57 | -0.01 | 0.90 | -0.02 | 0.02 |
| Grozea | [34] | 0.43 | -0.19 | -0.19 | -0.18 | 0.60 | -0.17 | 0.72 | -0.34 | 0.47 |
| Gutierrez | [35] | 1 | -0.10 | -0.05 | -0.07 | 0.56 | -0.10 | 0.41 | -0.21 | 0.51 |
| Harvey | [40] | 0.002 | -0.10 | -0.03 | -0.06 | 0.55 | -0.09 | 0.18 | -0.18 | 1.00 |
| Hürlimann | [42] | 0.4895 | -0.50 | -0.44 | -0.32 | 0.61 | -0.46 | 0.60 | -0.54 | 0.90 |
| Jankowska | [43] | 0.16 | -0.21 | -0.07 | -0.11 | 0.52 | -0.19 | 0.80 | -0.38 | 0.48 |
| Jankowska | [44] | 0.284 | -0.21 | -0.02 | -0.10 | 0.56 | -0.16 | 0.84 | -0.31 | 0.37 |
| Jayapal | [45] | 1 | 0.00 | -0.06 | -0.03 | 0.65 | -0.06 | 0.46 | -0.12 | 0.26 |
| Kern | [51] | 0.31 | 0.27 | -0.01 | 0.13 | 0.57 | 0.02 | 0.92 | 0.03 | -0.38 |
| Khonji | [53] | 0.735 | -0.34 | -0.11 | -0.25 | 0.72 | -0.28 | 0.61 | -0.52 | 0.85 |
| Kocher | [58] | 0.45 | 0.02 | -0.11 | -0.06 | 0.64 | 0.05 | 0.70 | 0.10 | -0.33 |
| Layton | [64] | -inf | 0.00 | 0.00 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Layton | [63] | 1 | -0.01 | -0.01 | -0.01 | 0.51 | -0.01 | 0.98 | -0.01 | 0.01 |
| Ledesma | [66] | 1 | 0.00 | -0.05 | -0.02 | 0.52 | -0.05 | 0.10 | -0.09 | 0.90 |
| Maitra | [69] | 0.64 | -0.53 | -0.12 | -0.36 | 0.71 | -0.24 | 0.65 | 0.19 | -0.54 |
| Mayor | [71] | 0.2 | -0.05 | -0.01 | -0.04 | 0.61 | -0.06 | 0.71 | -0.10 | 0.14 |
| Mechti | [74] | 0.833 | -0.04 | 0.00 | -0.02 | 0.61 | -0.03 | 0.63 | 0.02 | -0.05 |
| Modaresi | [77] | 0.395 | -0.32 | -0.29 | -0.34 | 0.72 | -0.30 | 0.68 | -0.58 | 0.85 |
| Moreau | [81] | 1 | 0.00 | -0.11 | -0.06 | 0.59 | -0.11 | 0.25 | -0.22 | 0.88 |
| Moreau | [79] | 0.8037 | -0.14 | -0.09 | -0.11 | 0.61 | -0.11 | 0.42 | -0.21 | 0.50 |
| Moreau | [80] | 0.5627 | -0.07 | 0.01 | -0.03 | 0.63 | -0.08 | 0.72 | -0.12 | 0.17 |
| Nikolov | [82] | 0.451 | 0.09 | 0.00 | 0.04 | 0.52 | -0.03 | 0.06 | -0.06 | 1.00 |
| Pacheco | [83] | 0.7114 | -0.08 | 0.00 | -0.04 | 0.59 | -0.12 | 0.26 | -0.24 | 0.92 |
| Petmanson | [84] | 0.1 | -0.20 | -0.07 | -0.12 | 0.54 | -0.19 | 0.49 | -0.38 | 0.78 |
| Satyam | [94] | 0.523 | -0.34 | -0.19 | -0.26 | 0.63 | -0.20 | 0.56 | -0.40 | 0.71 |
| Seidman | [95] | 1 | -0.01 | -0.01 | -0.01 | 0.52 | -0.01 | 0.99 | 0.00 | 0.00 |
| Solórzano | [96] | 0.594 | -0.06 | -0.05 | -0.05 | 0.54 | -0.09 | 0.75 | -0.18 | 0.24 |
| van Dam | [102] | 1 | 0.00 | -0.03 | -0.01 | 0.52 | -0.03 | 0.50 | -0.03 | 0.06 |
| Vartapetiance | [103] | 1 | 0.00 | -0.07 | -0.04 | 0.52 | -0.07 | 0.14 | -0.14 | 1.00 |
| Vartapetiance | [104] | -inf | -0.03 | -0.03 | -0.03 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Vilarino | [107] | -inf | 0.00 | 0.01 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Zamani | [113] | 0.456 | -0.26 | -0.14 | -0.24 | 0.69 | -0.19 | 0.90 | -0.27 | 0.30 |

**PAN 2014 EE test dataset**

| Verifier | | | PAN measures | | | Section 3 | | | | Impact |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | -inf | -0.15 | -0.15 | -0.15 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Bartoli | [7] | 0.611 | -0.31 | -0.15 | -0.20 | 0.59 | -0.18 | 0.43 | -0.35 | 0.81 |
| Bobicev | [8] | 0.6704 | -0.15 | -0.08 | -0.10 | 0.52 | -0.09 | 0.33 | -0.13 | 0.39 |
| Castillo | [17] | 0.7 | -0.08 | -0.09 | -0.08 | 0.58 | -0.09 | 0.59 | -0.17 | 0.29 |
| Castro | [18] | 0.8 | -0.11 | -0.05 | -0.09 | 0.60 | -0.11 | 0.42 | -0.21 | 0.50 |
| Fréry | [29] | 0.5 | -0.15 | -0.12 | -0.18 | 0.72 | -0.13 | 0.80 | -0.25 | 0.31 |
| Gómez-Adorno | [33] | -inf | -0.02 | -0.02 | -0.02 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Grozea | [34] | 0.81 | -0.19 | -0.17 | -0.14 | 0.53 | -0.10 | 0.28 | -0.20 | 0.71 |
| Gutierrez | [35] | -inf | -0.30 | -0.25 | -0.21 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Harvey | [40] | 0.014 | -0.11 | -0.05 | -0.08 | 0.59 | -0.12 | 0.37 | -0.24 | 0.65 |
| Hürlimann | [42] | 0.3552 | -0.32 | -0.28 | -0.20 | 0.58 | -0.37 | 0.89 | -0.60 | 0.67 |
| Jankowska | [43] | 0.62 | -0.27 | -0.18 | -0.19 | 0.55 | -0.24 | 0.53 | -0.47 | 0.89 |
| Jankowska | [44] | 0.5 | -0.21 | -0.20 | -0.18 | 0.55 | -0.21 | 0.50 | -0.42 | 0.84 |
| Jayapal | [45] | -inf | 0.00 | -0.06 | -0.03 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Kern | [51] | -inf | 0.11 | -0.12 | -0.01 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Khonji | [53] | 0.482 | -0.31 | -0.14 | -0.22 | 0.62 | -0.17 | 0.39 | -0.32 | 0.82 |
| Kocher | [58] | 0.482 | -0.03 | -0.07 | -0.06 | 0.62 | -0.04 | 0.63 | -0.07 | 0.11 |
| Layton | [64] | 1 | 0.07 | -0.21 | -0.09 | 0.58 | 0.06 | 0.72 | 0.11 | -0.39 |
| Layton | [63] | 0.7057 | -0.12 | -0.12 | -0.13 | 0.61 | -0.12 | 0.30 | -0.24 | 0.80 |
| Ledesma | [66] | 1 | 0.00 | -0.17 | -0.08 | 0.55 | -0.17 | 0.40 | -0.33 | 0.83 |
| Maitra | [69] | 0.5 | -0.25 | -0.17 | -0.20 | 0.57 | -0.17 | 0.98 | -0.33 | 0.34 |
| Mayor | [71] | 0.2 | -0.06 | -0.04 | -0.05 | 0.57 | -0.08 | 0.49 | -0.12 | 0.24 |
| Mechti | [74] | 0.667 | 0.00 | 0.00 | 0.00 | 0.53 | -0.04 | 0.21 | -0.03 | 0.14 |
| Modaresi | [77] | 0.757 | -0.19 | -0.13 | -0.16 | 0.63 | -0.20 | 0.50 | -0.29 | 0.58 |
| Moreau | [81] | -inf | 0.00 | -0.24 | -0.12 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Moreau | [79] | 0.5177 | -0.22 | -0.13 | -0.19 | 0.61 | -0.15 | 0.79 | -0.29 | 0.37 |
| Moreau | [80] | 0.6246 | -0.05 | -0.02 | -0.03 | 0.59 | -0.07 | 0.22 | -0.12 | 0.55 |
| Nikolov | [82] | 0.448 | 0.05 | -0.01 | 0.02 | 0.59 | -0.13 | 0.29 | -0.26 | 0.90 |
| Pacheco | [83] | -inf | -0.49 | 0.00 | -0.25 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Petmanson | [84] | -inf | -0.10 | -0.09 | -0.08 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Satyam | [94] | 0.479 | -0.17 | -0.11 | -0.17 | 0.71 | -0.14 | 0.68 | -0.28 | 0.41 |
| Seidman | [95] | -inf | -0.12 | -0.11 | -0.09 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Solórzano | [96] | 0.907 | -0.00 | -0.01 | -0.00 | 0.52 | -0.03 | 0.09 | -0.05 | 0.56 |
| van Dam | [102] | -inf | 0.00 | -0.10 | -0.05 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vartapetiance | [103] | -inf | 0.00 | -0.11 | -0.06 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vartapetiance | [104] | 1 | -0.14 | -0.14 | -0.12 | 0.52 | -0.14 | 0.84 | -0.27 | 0.32 |
| Vilarino | [107] | -inf | 0.00 | -0.01 | -0.00 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Zamani | [113] | 0.488 | -0.06 | -0.03 | -0.05 | 0.58 | -0.05 | 0.76 | -0.06 | 0.08 |

**PAN 2015 test dataset**

| Verifier | | | PAN measures | | | Section 3 | | | | Impact |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | 0.562 | -0.24 | -0.19 | -0.29 | 0.77 | -0.19 | 0.70 | -0.26 | 0.36 |
| Bartoli | [7] | 0.455 | -0.21 | -0.16 | -0.18 | 0.58 | -0.12 | 0.87 | -0.25 | 0.28 |
| Bobicev | [8] | 0.6893 | -0.21 | -0.12 | -0.18 | 0.67 | -0.15 | 0.66 | -0.31 | 0.47 |
| Castillo | [17] | 0.7 | -0.11 | -0.12 | -0.13 | 0.64 | -0.12 | 0.83 | -0.24 | 0.29 |
| Castro | [18] | 0.7 | -0.09 | -0.05 | -0.09 | 0.71 | -0.06 | 0.79 | -0.12 | 0.16 |
| Fréry | [29] | 0.143 | -0.14 | -0.07 | -0.09 | 0.55 | -0.15 | 0.52 | -0.30 | 0.58 |
| Gómez-Adorno | [33] | 1 | -0.06 | -0.06 | -0.06 | 0.53 | -0.06 | 0.96 | -0.08 | 0.08 |
| Grozea | [34] | 0.09 | -0.09 | -0.06 | -0.08 | 0.56 | -0.09 | 0.29 | -0.17 | 0.60 |
| Gutierrez | [35] | 0.7273 | -0.16 | -0.12 | -0.18 | 0.70 | -0.11 | 0.72 | -0.22 | 0.30 |
| Harvey | [40] | 0.502 | -0.12 | -0.10 | -0.13 | 0.62 | -0.11 | 0.70 | -0.22 | 0.31 |
| Hürlimann | [42] | 0.5238 | -0.41 | -0.32 | -0.33 | 0.64 | -0.30 | 0.56 | -0.33 | 0.59 |
| Jankowska | [43] | 0.6 | -0.13 | -0.08 | -0.12 | 0.62 | -0.13 | 0.67 | -0.26 | 0.38 |
| Jankowska | [44] | 0.407 | -0.12 | -0.15 | -0.16 | 0.67 | -0.06 | 0.83 | -0.12 | 0.14 |
| Jayapal | [45] | 1 | 0.00 | -0.01 | -0.01 | 0.51 | -0.01 | 0.63 | -0.02 | 0.03 |
| Kern | [51] | -inf | 0.36 | -0.12 | 0.14 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Khonji | [53] | 0.333 | -0.34 | -0.18 | -0.31 | 0.76 | -0.27 | 0.84 | -0.53 | 0.63 |
| Kocher | [58] | 0.5 | -0.03 | -0.11 | -0.10 | 0.70 | -0.08 | 0.60 | -0.16 | 0.26 |
| Layton | [64] | -inf | 0.00 | 0.00 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Layton | [63] | 0.2943 | -0.07 | 0.00 | -0.04 | 0.67 | -0.07 | 0.37 | -0.14 | 0.39 |
| Ledesma | [66] | 1 | 0.00 | -0.22 | -0.11 | 0.62 | -0.22 | 0.50 | -0.44 | 0.87 |
| Maitra | [69] | 0.42 | -0.18 | -0.17 | -0.18 | 0.59 | -0.11 | 0.82 | -0.22 | 0.26 |
| Mayor | [71] | 0.8 | -0.13 | -0.08 | -0.12 | 0.64 | -0.11 | 0.45 | -0.18 | 0.41 |
| Mechti | [74] | 0.917 | 0.02 | -0.01 | 0.01 | 0.51 | -0.02 | 0.26 | 0.01 | -0.01 |
| Modaresi | [77] | 0.201 | -0.13 | -0.12 | -0.09 | 0.50 | -0.00 | 1.00 | -0.00 | 0.00 |
| Moreau | [81] | -inf | 0.00 | -0.01 | -0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Moreau | [80] | 0.5536 | -0.16 | -0.09 | -0.15 | 0.70 | -0.15 | 0.60 | -0.29 | 0.49 |
| Nikolov | [82] | 0.534 | -0.02 | -0.18 | -0.09 | 0.56 | -0.12 | 0.25 | -0.25 | 1.00 |
| Pacheco | [83] | 0.5650 | -0.03 | 0.02 | -0.00 | 0.71 | -0.08 | 0.59 | -0.15 | 0.26 |
| Petmanson | [84] | 0.99 | -0.05 | -0.03 | -0.04 | 0.55 | -0.03 | 0.30 | -0.06 | 0.21 |
| Pimas | [85] | 0.108 | -0.02 | -0.02 | -0.02 | 0.51 | -0.02 | 0.56 | -0.04 | 0.08 |
| Posadas-Durán | [86] | 0.8316 | -0.22 | -0.13 | -0.19 | 0.66 | -0.16 | 0.66 | -0.66 | 1.00 |
| Sari | [93] | 0.544 | 0.19 | 0.00 | 0.10 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Satyam | [94] | 0.145 | -0.14 | 0.00 | -0.07 | 0.72 | -0.20 | 0.48 | -0.40 | 0.83 |
| Seidman | [95] | 1 | -0.06 | -0.06 | -0.08 | 0.68 | -0.06 | 0.42 | -0.13 | 0.31 |
| Solórzano | [96] | 0.691 | -0.03 | -0.02 | -0.02 | 0.54 | -0.02 | 0.36 | -0.04 | 0.12 |
| van Dam | [102] | 1 | 0.00 | -0.01 | -0.00 | 0.59 | -0.01 | 0.60 | 0.00 | -0.01 |
| Vartapetiance | [103] | -inf | 0.00 | -0.03 | -0.02 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Vartapetiance | [104] | 1 | -0.19 | -0.19 | -0.20 | 0.60 | -0.19 | 0.40 | -0.38 | 0.95 |
| Vilarino | [107] | -inf | 0.00 | 0.00 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Zamani | [113] | 0.761 | -0.07 | -0.05 | -0.08 | 0.71 | -0.07 | 0.54 | -0.07 | 0.13 |

**Table 4.** Safety evaluation of the obfuscator of Keswani *et al.* [52]. Each table shows the performances and performance deltas of various authorship verification approaches submitted to PAN 2013 through PAN 2015 when run on test datasets that have been obfuscated by this obfuscator. Verifiers that failed to process a dataset (e.g., for being incompatible or not scalable) have been omitted from the tables. Verifiers whose optimal classification threshold $\tau$ that maximizes classification accuracy $\mathrm{acc}$ on the unobfuscated test dataset turned out to be negative or positive infinity (i.e., marking all problem instances "same author" or "different author", respectively) were omitted from forming the average performances reported in Table 1.

**PAN 2013 test dataset**

| Verifier | | | PAN measures | | | Section 3 | | | | Impact |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | 0.478 | -0.04 | -0.06 | -0.08 | 0.80 | -0.03 | 0.93 | -0.07 | 0.08 |
| Bartoli | [7] | 0.647 | 0.14 | 0.07 | 0.11 | 0.67 | 0.03 | 0.36 | 0.07 | -0.11 |
| Bobicev | [8] | 0.5144 | -0.25 | -0.16 | -0.23 | 0.67 | -0.17 | 0.50 | -0.36 | 0.71 |
| Castillo | [17] | 0.4 | -0.06 | 0.03 | -0.01 | 0.53 | -0.10 | 0.36 | -0.21 | 0.60 |
| Castro | [18] | 0.9 | -0.06 | -0.03 | -0.08 | 0.93 | -0.07 | 1.00 | -0.14 | 0.14 |
| Feng | [24] | 0.46 | -0.03 | -0.13 | -0.12 | 0.77 | -0.13 | 0.79 | -0.29 | 0.36 |
| Fratila | [28] | 0.38 | -0.28 | -0.23 | -0.27 | 0.67 | -0.20 | 0.71 | -0.43 | 0.60 |
| Fréry | [29] | 0.333 | -0.28 | -0.17 | -0.22 | 0.63 | -0.27 | 0.57 | -0.57 | 1.00 |
| Ghaeini | [31] | 0.46 | -0.39 | -0.26 | -0.40 | 0.80 | -0.30 | 0.71 | -0.64 | 0.90 |
| Gómez-Adorno | [33] | inf | 0.00 | 0.00 | 0.00 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Grozea | [34] | 0.05 | 0.06 | 0.00 | 0.02 | 0.53 | -0.03 | 1.00 | -0.07 | 0.07 |
| Gutierrez | [35] | 0.8182 | -0.16 | -0.11 | -0.17 | 0.77 | -0.13 | 0.79 | -0.29 | 0.36 |
| Harvey | [40] | 0.001 | -0.19 | 0.00 | -0.10 | 0.60 | -0.17 | 0.50 | -0.36 | 0.71 |
| Hürlimann | [42] | 0.6394 | -0.10 | -0.02 | -0.08 | 0.70 | -0.10 | 0.36 | -0.14 | 0.40 |
| Jankowska | [43] | 0.59 | -0.15 | -0.07 | -0.15 | 0.80 | -0.20 | 0.64 | -0.43 | 0.67 |
| Jankowska | [44] | 0.615 | -0.19 | -0.13 | -0.24 | 0.80 | -0.20 | 0.64 | -0.43 | 0.67 |
| Jayapal | [45] | 1 | 0.00 | -0.07 | -0.03 | 0.60 | -0.07 | 0.36 | -0.14 | 0.40 |
| Kern | [51] | 0.5 | -0.06 | -0.20 | -0.10 | 0.57 | -0.03 | 0.07 | -0.07 | 1.00 |
| Khonji | [53] | 0.444 | -0.23 | -0.22 | -0.32 | 0.80 | -0.17 | 0.93 | -0.36 | 0.38 |
| Kocher | [58] | 0.484 | -0.03 | -0.04 | -0.04 | 0.67 | -0.03 | 1.00 | -0.07 | 0.07 |
| Layton | [64] | inf | 0.26 | -0.23 | 0.07 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Layton | [63] | 0.7057 | -0.15 | -0.13 | -0.16 | 0.67 | -0.13 | 0.29 | -0.29 | 1.00 |
| Ledesma | [66] | inf | 0.00 | 0.00 | 0.00 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Maitra | [69] | 0.8 | 0.09 | 0.07 | 0.09 | 0.60 | 0.03 | 0.29 | 0.07 | -0.10 |
| Mayor | [71] | 0.1 | -0.21 | -0.12 | -0.20 | 0.73 | -0.17 | 0.79 | -0.36 | 0.45 |
| Mechti | [74] | 0.469 | 0.11 | 0.02 | 0.07 | 0.60 | 0.03 | 0.50 | 0.14 | -0.29 |
| Modaresi | [77] | 0.392 | -0.25 | -0.17 | -0.17 | 0.57 | -0.23 | 0.79 | -0.50 | 0.64 |
| Moreau | [81] | 1 | 0.00 | -0.23 | -0.12 | 0.73 | -0.23 | 0.71 | -0.50 | 0.70 |
| Moreau | [80] | 0.6215 | -0.28 | 0.00 | -0.13 | 0.70 | -0.13 | 0.93 | -0.14 | 0.15 |
| Nikolov | [82] | 0.448 | -0.15 | 0.00 | -0.08 | 0.60 | -0.13 | 0.29 | -0.29 | 1.00 |
| Pacheco | [83] | 0.7223 | 0.01 | 0.00 | 0.00 | 0.60 | -0.10 | 0.71 | -0.21 | 0.30 |
| Petmanson | [84] | 0.59 | -0.21 | -0.13 | -0.20 | 0.73 | -0.13 | 0.57 | -0.29 | 0.50 |
| Sari | [93] | 0.546 | -0.04 | 0.00 | -0.02 | 0.53 | -0.13 | 0.93 | -0.29 | 0.31 |
| Satyam | [94] | 0.423 | 0.08 | 0.13 | 0.09 | 0.53 | 0.00 | 1.00 | 0.00 | 0.00 |
| Seidman | [95] | 1 | -0.04 | -0.03 | -0.06 | 0.77 | -0.07 | 0.71 | -0.14 | 0.20 |
| Solórzano | [96] | 0.812 | -0.03 | 0.07 | 0.01 | 0.57 | -0.17 | 0.64 | -0.36 | 0.56 |
| van Dam | [102] | 1 | 0.00 | -0.07 | -0.03 | 0.60 | -0.07 | 0.57 | -0.07 | 0.13 |
| Vartapetiance | [103] | inf | 0.00 | -0.40 | -0.20 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vartapetiance | [104] | inf | 0.00 | 0.00 | 0.00 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vilarino | [107] | 1 | 0.00 | -0.10 | -0.05 | 0.67 | -0.10 | 0.36 | -0.21 | 0.60 |
| Zamani | [113] | 0.997 | 0.09 | 0.00 | 0.07 | 0.80 | -0.03 | 0.64 | -0.07 | 0.11 |

**PAN 2014 EE test dataset**

| Verifier | | | PAN measures | | | Section 3 | | | | Impact |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | -inf | -0.05 | -0.06 | -0.06 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Bartoli | [7] | 0.611 | -0.18 | -0.08 | -0.12 | 0.59 | -0.11 | 0.43 | -0.21 | 0.49 |
| Bobicev | [8] | 0.6704 | -0.15 | -0.09 | -0.10 | 0.52 | -0.10 | 0.33 | -0.18 | 0.55 |
| Castillo | [17] | 0.7 | -0.07 | -0.07 | -0.07 | 0.58 | -0.07 | 0.59 | -0.13 | 0.22 |
| Castro | [18] | 0.8 | -0.08 | -0.03 | -0.06 | 0.60 | -0.09 | 0.42 | -0.18 | 0.43 |
| Fréry | [29] | 0.5 | -0.05 | -0.05 | -0.07 | 0.72 | -0.04 | 0.80 | -0.07 | 0.09 |
| Gómez-Adorno | [33] | -inf | -0.10 | -0.10 | -0.09 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Grozea | [34] | 0.81 | -0.05 | -0.05 | -0.04 | 0.53 | -0.05 | 0.28 | -0.10 | 0.36 |
| Gutierrez | [35] | -inf | -0.31 | -0.26 | -0.22 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Harvey | [40] | 0.014 | -0.10 | -0.05 | -0.07 | 0.59 | -0.10 | 0.37 | -0.19 | 0.51 |
| Hürlimann | [42] | 0.3552 | -0.30 | -0.28 | -0.20 | 0.58 | -0.27 | 0.89 | -0.47 | 0.53 |
| Jankowska | [43] | 0.62 | -0.22 | -0.13 | -0.16 | 0.55 | -0.20 | 0.53 | -0.39 | 0.74 |
| Jankowska | [44] | 0.5 | -0.16 | -0.18 | -0.15 | 0.55 | -0.18 | 0.50 | -0.35 | 0.70 |
| Jayapal | [45] | -inf | 0.00 | -0.06 | -0.03 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Kern | [51] | -inf | 0.11 | -0.14 | -0.03 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Khonji | [53] | 0.482 | -0.29 | -0.15 | -0.21 | 0.62 | -0.19 | 0.39 | -0.35 | 0.90 |
| Kocher | [58] | 0.482 | -0.08 | -0.08 | -0.09 | 0.62 | -0.09 | 0.63 | -0.18 | 0.29 |
| Layton | [64] | 1 | 0.09 | -0.23 | -0.10 | 0.58 | 0.08 | 0.72 | 0.15 | -0.54 |
| Layton | [63] | 0.7057 | -0.13 | -0.14 | -0.14 | 0.61 | -0.14 | 0.30 | -0.27 | 0.90 |
| Ledesma | [66] | 1 | 0.00 | -0.07 | -0.03 | 0.55 | -0.07 | 0.40 | -0.13 | 0.33 |
| Maitra | [69] | 0.5 | -0.24 | -0.18 | -0.20 | 0.57 | -0.19 | 0.98 | -0.37 | 0.38 |
| Mayor | [71] | 0.2 | -0.07 | -0.03 | -0.05 | 0.57 | -0.05 | 0.49 | -0.08 | 0.16 |
| Mechti | [74] | 0.667 | 0.03 | 0.02 | 0.02 | 0.53 | 0.02 | 0.21 | 0.00 | 0.00 |
| Modaresi | [77] | 0.757 | -0.09 | -0.02 | -0.06 | 0.63 | -0.11 | 0.50 | -0.15 | 0.30 |
| Moreau | [81] | -inf | 0.00 | -0.20 | -0.10 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Moreau | [79] | 0.5177 | -0.22 | -0.15 | -0.19 | 0.61 | -0.15 | 0.79 | -0.29 | 0.37 |
| Moreau | [80] | 0.6246 | -0.09 | -0.04 | -0.06 | 0.59 | -0.09 | 0.22 | -0.13 | 0.59 |
| Nikolov | [82] | 0.448 | 0.05 | -0.01 | 0.02 | 0.59 | -0.09 | 0.29 | -0.18 | 0.62 |
| Pacheco | [83] | -inf | -0.44 | 0.00 | -0.22 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Petmanson | [84] | -inf | -0.02 | -0.03 | -0.02 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Satyam | [94] | 0.479 | -0.25 | -0.16 | -0.24 | 0.71 | -0.20 | 0.68 | -0.39 | 0.57 |
| Seidman | [95] | -inf | -0.15 | -0.12 | -0.11 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Solórzano | [96] | 0.907 | -0.07 | -0.03 | -0.04 | 0.52 | -0.04 | 0.09 | -0.08 | 0.89 |
| van Dam | [102] | -inf | 0.00 | -0.11 | -0.06 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vartapetiance | [103] | -inf | 0.00 | -0.11 | -0.06 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vartapetiance | [104] | 1 | -0.02 | -0.02 | -0.02 | 0.52 | -0.02 | 0.84 | -0.04 | 0.05 |
| Vilarino | [107] | -inf | 0.00 | -0.01 | -0.00 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Zamani | [113] | 0.488 | -0.07 | -0.01 | -0.04 | 0.58 | -0.06 | 0.76 | -0.11 | 0.14 |

**PAN 2014 EN test dataset**

| Verifier | | | PAN measures | | | Section 3 | | | | Impact |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | 0.418 | -0.10 | -0.06 | -0.11 | 0.70 | -0.13 | 0.81 | -0.06 | 0.07 |
| Bartoli | [7] | 0.613 | 0.01 | -0.02 | -0.01 | 0.62 | 0.01 | 0.77 | 0.01 | -0.04 |
| Bobicev | [8] | 0.3406 | -0.23 | -0.05 | -0.14 | 0.57 | -0.17 | 0.42 | -0.36 | 0.86 |
| Castillo | [17] | 1 | -0.31 | -0.27 | -0.28 | 0.62 | -0.27 | 0.63 | -0.53 | 0.84 |
| Castro | [18] | 0.8 | -0.01 | 0.00 | -0.00 | 0.62 | -0.02 | 0.76 | -0.04 | 0.05 |
| Feng | [24] | 0.59 | -0.54 | -0.37 | -0.35 | 0.64 | -0.43 | 0.88 | -0.83 | 0.94 |
| Fréry | [29] | 1 | -0.10 | -0.04 | -0.07 | 0.61 | -0.08 | 0.69 | -0.15 | 0.22 |
| Gómez-Adorno | [33] | 1 | -0.10 | -0.10 | -0.10 | 0.57 | -0.10 | 0.90 | -0.11 | 0.12 |
| Grozea | [34] | 0.43 | -0.02 | -0.04 | -0.03 | 0.60 | -0.03 | 0.72 | -0.05 | 0.07 |
| Gutierrez | [35] | 1 | -0.12 | -0.05 | -0.08 | 0.56 | -0.15 | 0.41 | -0.28 | 0.68 |
| Harvey | [40] | 0.002 | -0.07 | -0.03 | -0.05 | 0.55 | -0.07 | 0.18 | -0.14 | 0.78 |
| Hürlimann | [42] | 0.4895 | -0.33 | -0.28 | -0.26 | 0.61 | -0.29 | 0.60 | -0.36 | 0.60 |
| Jankowska | [43] | 0.16 | -0.19 | -0.08 | -0.11 | 0.52 | -0.16 | 0.80 | -0.31 | 0.39 |
| Jankowska | [44] | 0.284 | -0.30 | -0.02 | -0.14 | 0.56 | -0.27 | 0.84 | -0.54 | 0.64 |
| Jayapal | [45] | 1 | 0.00 | -0.07 | -0.03 | 0.65 | -0.07 | 0.46 | -0.13 | 0.28 |
| Kern | [51] | 0.31 | 0.12 | 0.00 | 0.06 | 0.57 | 0.00 | 0.92 | 0.00 | 0.00 |
| Khonji | [53] | 0.735 | -0.38 | -0.16 | -0.29 | 0.72 | -0.29 | 0.61 | -0.55 | 0.90 |
| Kocher | [58] | 0.45 | -0.01 | -0.07 | -0.05 | 0.64 | -0.02 | 0.70 | -0.04 | 0.06 |
| Layton | [64] | -inf | 0.00 | 0.00 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Layton | [63] | 1 | -0.04 | -0.04 | -0.03 | 0.51 | -0.04 | 0.98 | -0.07 | 0.07 |
| Ledesma | [66] | 1 | 0.00 | -0.03 | -0.01 | 0.52 | -0.03 | 0.10 | -0.05 | 0.50 |
| Maitra | [69] | 0.64 | -0.21 | -0.14 | -0.21 | 0.71 | -0.20 | 0.65 | 0.26 | -0.74 |
| Mayor | [71] | 0.2 | -0.06 | -0.06 | -0.07 | 0.61 | -0.05 | 0.71 | -0.03 | 0.04 |
| Mechti | [74] | 0.833 | -0.01 | 0.00 | -0.00 | 0.61 | -0.03 | 0.63 | -0.02 | 0.03 |
| Modaresi | [77] | 0.395 | -0.11 | -0.09 | -0.14 | 0.72 | -0.09 | 0.68 | -0.20 | 0.29 |
| Moreau | [81] | 1 | 0.00 | -0.10 | -0.05 | 0.59 | -0.10 | 0.25 | -0.19 | 0.76 |
| Moreau | [79] | 0.8037 | -0.20 | -0.09 | -0.14 | 0.61 | -0.17 | 0.42 | -0.33 | 0.79 |
| Moreau | [80] | 0.5627 | 0.03 | 0.03 | 0.04 | 0.63 | 0.01 | 0.72 | 0.01 | -0.04 |
| Nikolov | [82] | 0.451 | 0.08 | 0.00 | 0.04 | 0.52 | -0.03 | 0.06 | -0.06 | 1.00 |
| Pacheco | [83] | 0.7114 | -0.09 | 0.00 | -0.04 | 0.59 | -0.12 | 0.26 | -0.23 | 0.88 |
| Petmanson | [84] | 0.1 | -0.07 | -0.06 | -0.06 | 0.54 | -0.12 | 0.49 | -0.24 | 0.49 |
| Satyam | [94] | 0.523 | -0.23 | -0.11 | -0.18 | 0.63 | -0.14 | 0.56 | -0.28 | 0.50 |
| Seidman | [95] | 1 | -0.01 | -0.01 | -0.01 | 0.52 | -0.01 | 0.99 | -0.03 | 0.03 |
| Solórzano | [96] | 0.594 | -0.09 | -0.07 | -0.07 | 0.54 | -0.13 | 0.75 | -0.25 | 0.33 |
| van Dam | [102] | 1 | 0.00 | -0.05 | -0.03 | 0.52 | -0.05 | 0.50 | -0.05 | 0.10 |
| Vartapetiance | [103] | 1 | 0.00 | -0.07 | -0.04 | 0.52 | -0.07 | 0.14 | -0.14 | 1.00 |
| Vartapetiance | [104] | -inf | -0.01 | -0.01 | -0.01 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Vilarino | [107] | -inf | 0.00 | 0.00 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Zamani | [113] | 0.456 | -0.02 | -0.01 | -0.01 | 0.69 | -0.03 | 0.90 | -0.05 | 0.06 |

**PAN 2015 test dataset**

| Verifier | | | PAN measures | | | Section 3 | | | | Impact |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | 0.562 | -0.06 | -0.04 | -0.07 | 0.77 | -0.06 | 0.70 | -0.11 | 0.15 |
| Bartoli | [7] | 0.455 | -0.05 | -0.04 | -0.05 | 0.58 | -0.03 | 0.87 | -0.06 | 0.07 |
| Bobicev | [8] | 0.6893 | -0.09 | -0.02 | -0.07 | 0.67 | -0.08 | 0.66 | -0.15 | 0.23 |
| Castillo | [17] | 0.7 | -0.06 | -0.05 | -0.07 | 0.64 | -0.05 | 0.83 | -0.10 | 0.13 |
| Castro | [18] | 0.7 | -0.03 | -0.01 | -0.03 | 0.71 | -0.01 | 0.79 | -0.03 | 0.04 |
| Fréry | [29] | 0.143 | -0.00 | 0.01 | 0.00 | 0.55 | -0.01 | 0.52 | -0.03 | 0.05 |
| Gómez-Adorno | [33] | 1 | -0.05 | -0.05 | -0.05 | 0.53 | -0.05 | 0.96 | -0.04 | 0.05 |
| Grozea | [34] | 0.09 | -0.04 | -0.01 | -0.03 | 0.56 | -0.05 | 0.29 | -0.10 | 0.33 |
| Gutierrez | [35] | 0.7273 | -0.08 | -0.05 | -0.09 | 0.70 | -0.06 | 0.72 | -0.13 | 0.18 |
| Harvey | [40] | 0.502 | -0.04 | -0.01 | -0.03 | 0.62 | -0.01 | 0.70 | -0.02 | 0.03 |
| Hürlimann | [42] | 0.5238 | -0.22 | -0.17 | -0.21 | 0.64 | -0.18 | 0.56 | -0.21 | 0.37 |
| Jankowska | [43] | 0.6 | -0.11 | -0.06 | -0.10 | 0.62 | -0.11 | 0.67 | -0.22 | 0.33 |
| Jankowska | [44] | 0.407 | -0.14 | -0.17 | -0.18 | 0.67 | -0.09 | 0.83 | -0.18 | 0.21 |
| Jayapal | [45] | 1 | 0.00 | -0.02 | -0.01 | 0.51 | -0.02 | 0.63 | -0.05 | 0.08 |
| Kern | [51] | -inf | 0.11 | -0.04 | 0.05 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Khonji | [53] | 0.333 | -0.14 | -0.15 | -0.19 | 0.76 | -0.11 | 0.84 | -0.22 | 0.26 |
| Kocher | [58] | 0.5 | 0.01 | -0.01 | 0.00 | 0.70 | 0.00 | 0.60 | 0.01 | -0.02 |
| Layton | [64] | -inf | 0.00 | 0.00 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Layton | [63] | 0.2943 | -0.15 | 0.00 | -0.08 | 0.67 | -0.15 | 0.37 | -0.30 | 0.82 |
| Ledesma | [66] | 1 | 0.00 | -0.11 | -0.05 | 0.62 | -0.11 | 0.50 | -0.21 | 0.42 |
| Maitra | [69] | 0.42 | -0.09 | -0.07 | -0.09 | 0.59 | -0.05 | 0.82 | -0.10 | 0.12 |
| Mayor | [71] | 0.8 | -0.05 | -0.02 | -0.04 | 0.64 | -0.04 | 0.45 | -0.03 | 0.07 |
| Mechti | [74] | 0.917 | -0.00 | -0.01 | -0.01 | 0.51 | -0.01 | 0.26 | 0.00 | -0.01 |
| Modaresi | [77] | 0.201 | 0.03 | 0.03 | 0.02 | 0.50 | -0.00 | 1.00 | 0.00 | 0.00 |
| Moreau | [81] | -inf | 0.00 | -0.01 | -0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Moreau | [80] | 0.5536 | -0.15 | -0.09 | -0.14 | 0.70 | -0.17 | 0.60 | -0.33 | 0.55 |
| Nikolov | [82] | 0.534 | -0.02 | -0.18 | -0.09 | 0.56 | -0.12 | 0.25 | -0.25 | 1.00 |
| Pacheco | [83] | 0.5650 | 0.04 | 0.02 | 0.04 | 0.71 | 0.01 | 0.59 | 0.02 | -0.06 |
| Petmanson | [84] | 0.99 | -0.04 | -0.03 | -0.03 | 0.55 | -0.03 | 0.30 | -0.05 | 0.17 |
| Pimas | [85] | 0.108 | -0.01 | -0.01 | -0.01 | 0.51 | -0.02 | 0.56 | -0.03 | 0.06 |
| Posadas-Durán | [86] | 0.8316 | -0.16 | -0.07 | -0.13 | 0.66 | -0.16 | 0.66 | -0.66 | 1.00 |
| Sari | [93] | 0.544 | 0.06 | 0.00 | 0.03 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Satyam | [94] | 0.145 | -0.14 | 0.00 | -0.07 | 0.72 | -0.18 | 0.48 | -0.36 | 0.75 |
| Seidman | [95] | 1 | -0.14 | -0.14 | -0.17 | 0.68 | -0.14 | 0.42 | -0.30 | 0.72 |
| Solórzano | [96] | 0.691 | -0.05 | -0.02 | -0.04 | 0.54 | -0.05 | 0.36 | -0.10 | 0.27 |
| van Dam | [102] | 1 | 0.00 | -0.00 | -0.00 | 0.59 | -0.00 | 0.60 | 0.04 | -0.10 |
| Vartapetiance | [103] | -inf | 0.00 | -0.03 | -0.02 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Vartapetiance | [104] | 1 | -0.12 | -0.12 | -0.13 | 0.60 | -0.12 | 0.40 | -0.24 | 0.58 |
| Vilarino | [107] | -inf | 0.00 | 0.00 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Zamani | [113] | 0.761 | 0.02 | 0.03 | 0.03 | 0.71 | -0.01 | 0.54 | -0.03 | 0.05 |

**Table 5.** Safety evaluation of the obfuscator of Mansoorizadeh *et al.* [70]. Each table shows the performances and performance deltas of various authorship verification approaches submitted to PAN 2013 through PAN 2015 when run on test datasets that have been obfuscated by this obfuscator. Verifiers that failed to process a dataset (e.g., for being incompatible or not scalable) have been omitted from the tables. Verifiers whose optimal classification threshold $\tau$ that maximizes classification accuracy $\mathrm{acc}$ on the unobfuscated test dataset turned out to be negative or positive infinity (i.e., marking all problem instances "same author" or "different author", respectively) were omitted from forming the average performances reported in Table 1.

**PAN 2013 test dataset**

| **Verifier** | | | **PAN measures** | | | **Section 3** | | | | **Impact** |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | 0.478 | 0.01 | -0.02 | -0.00 | 0.80 | 0.07 | 0.93 | 0.07 | -1.00 |
| Bartoli | [7] | 0.647 | 0.09 | 0.07 | 0.09 | 0.67 | -0.03 | 0.36 | -0.07 | 0.20 |
| Bobicev | [8] | 0.5144 | -0.23 | -0.20 | -0.23 | 0.67 | -0.20 | 0.50 | -0.43 | 0.86 |
| Castillo | [17] | 0.4 | 0.02 | 0.03 | 0.03 | 0.53 | 0.00 | 0.36 | 0.00 | 0.00 |
| Castro | [18] | 0.9 | -0.03 | 0.00 | -0.02 | 0.93 | -0.03 | 1.00 | -0.07 | 0.07 |
| Feng | [24] | 0.46 | 0.01 | 0.03 | 0.03 | 0.77 | 0.00 | 0.79 | 0.00 | 0.00 |
| Fratila | [28] | 0.38 | 0.00 | 0.00 | 0.00 | 0.67 | 0.00 | 0.71 | 0.00 | 0.00 |
| Fréry | [29] | 0.333 | -0.18 | -0.13 | -0.16 | 0.63 | -0.17 | 0.57 | -0.36 | 0.63 |
| Ghaeini | [31] | 0.46 | -0.33 | -0.19 | -0.34 | 0.80 | -0.20 | 0.71 | -0.43 | 0.60 |
| Gómez-Adorno | [33] | inf | 0.00 | 0.00 | 0.00 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Grozea | [34] | 0.05 | -0.02 | 0.00 | -0.01 | 0.53 | -0.03 | 1.00 | -0.07 | 0.07 |
| Gutierrez | [35] | 0.8182 | -0.06 | -0.01 | -0.06 | 0.77 | -0.07 | 0.79 | -0.14 | 0.18 |
| Harvey | [40] | 0.001 | 0.02 | 0.00 | 0.01 | 0.60 | 0.03 | 0.50 | 0.07 | -0.14 |
| Hürlimann | [42] | 0.6394 | -0.08 | -0.05 | -0.09 | 0.70 | -0.17 | 0.36 | -0.14 | 0.40 |
| Jankowska | [43] | 0.59 | -0.04 | 0.00 | -0.03 | 0.80 | -0.07 | 0.64 | -0.14 | 0.22 |
| Jankowska | [44] | 0.615 | -0.06 | -0.03 | -0.08 | 0.80 | -0.07 | 0.64 | -0.14 | 0.22 |
| Jayapal | [45] | 1 | 0.00 | 0.03 | 0.02 | 0.60 | 0.03 | 0.36 | 0.07 | -0.11 |
| Kern | [51] | 0.5 | 0.08 | -0.13 | -0.02 | 0.57 | 0.00 | 0.07 | 0.00 | 0.00 |
| Khonji | [53] | 0.444 | -0.02 | -0.02 | -0.03 | 0.80 | -0.03 | 0.93 | -0.07 | 0.08 |
| Kocher | [58] | 0.484 | -0.00 | 0.00 | -0.00 | 0.67 | -0.03 | 1.00 | -0.07 | 0.07 |
| Layton | [64] | inf | 0.00 | -0.10 | -0.02 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Layton | [63] | 0.7057 | -0.04 | -0.03 | -0.04 | 0.67 | -0.03 | 0.29 | -0.07 | 0.25 |
| Ledesma | [66] | inf | 0.00 | -0.03 | -0.02 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Maitra | [69] | 0.8 | 0.00 | 0.02 | 0.01 | 0.60 | -0.03 | 0.29 | -0.07 | 0.25 |
| Mayor | [71] | 0.1 | -0.09 | -0.07 | -0.11 | 0.73 | -0.07 | 0.79 | -0.14 | 0.18 |
| Mechti | [74] | 0.469 | 0.00 | -0.02 | -0.01 | 0.60 | -0.07 | 0.50 | -0.07 | 0.14 |
| Modaresi | [77] | 0.392 | -0.03 | 0.07 | 0.01 | 0.57 | 0.00 | 0.79 | 0.00 | 0.00 |
| Moreau | [81] | 1 | 0.00 | 0.03 | 0.02 | 0.73 | 0.03 | 0.71 | 0.07 | -0.25 |
| Moreau | [80] | 0.6215 | -0.07 | 0.00 | -0.03 | 0.70 | 0.00 | 0.93 | -0.07 | 0.08 |
| Nikolov | [82] | 0.448 | -0.10 | 0.00 | -0.06 | 0.60 | -0.10 | 0.29 | -0.21 | 0.75 |
| Pacheco | [83] | 0.7223 | 0.11 | -0.05 | 0.05 | 0.60 | 0.00 | 0.71 | 0.00 | 0.00 |
| Petmanson | [84] | 0.59 | -0.48 | -0.23 | -0.37 | 0.73 | -0.23 | 0.57 | -0.50 | 0.88 |
| Sari | [93] | 0.546 | 0.00 | 0.00 | 0.00 | 0.53 | 0.00 | 0.93 | 0.00 | 0.00 |
| Satyam | [94] | 0.423 | 0.01 | 0.07 | 0.03 | 0.53 | 0.00 | 1.00 | 0.00 | 0.00 |
| Seidman | [95] | 1 | 0.00 | 0.00 | 0.00 | 0.77 | 0.00 | 0.71 | 0.00 | 0.00 |
| Solórzano | [96] | 0.812 | -0.06 | 0.00 | -0.03 | 0.57 | -0.10 | 0.64 | -0.21 | 0.33 |
| van Dam | [102] | 1 | 0.00 | -0.07 | -0.03 | 0.60 | -0.07 | 0.57 | -0.07 | 0.13 |
| Vartapetiance | [103] | inf | 0.00 | -0.40 | -0.20 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vartapetiance | [104] | inf | 0.00 | 0.00 | 0.00 | 0.53 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vilarino | [107] | 1 | 0.00 | 0.00 | 0.00 | 0.67 | 0.00 | 0.36 | 0.00 | 0.00 |
| Zamani | [113] | 0.997 | 0.05 | 0.00 | 0.03 | 0.80 | -0.03 | 0.64 | -0.07 | 0.11 |

**PAN 2014 EE test dataset**

| **Verifier** | | | **PAN measures** | | | **Section 3** | | | | **Impact** |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | -inf | -0.08 | -0.07 | -0.08 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Bartoli | [7] | 0.611 | 0.01 | 0.00 | 0.00 | 0.59 | -0.05 | 0.43 | -0.10 | 0.23 |
| Bobicev | [8] | 0.6704 | -0.17 | -0.11 | -0.11 | 0.52 | -0.10 | 0.33 | -0.18 | 0.55 |
| Castillo | [17] | 0.7 | -0.01 | -0.01 | -0.01 | 0.58 | -0.01 | 0.59 | -0.02 | 0.03 |
| Castro | [18] | 0.8 | -0.03 | -0.02 | -0.03 | 0.60 | -0.04 | 0.42 | -0.07 | 0.17 |
| Fréry | [29] | 0.5 | -0.02 | -0.01 | -0.02 | 0.72 | -0.02 | 0.80 | -0.03 | 0.04 |
| Gómez-Adorno | [33] | -inf | -0.02 | -0.02 | -0.02 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Grozea | [34] | 0.81 | -0.04 | -0.02 | -0.03 | 0.53 | -0.04 | 0.28 | -0.07 | 0.25 |
| Gutierrez | [35] | -inf | -0.30 | -0.26 | -0.21 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Harvey | [40] | 0.014 | -0.01 | -0.00 | -0.01 | 0.59 | -0.01 | 0.37 | -0.01 | 0.03 |
| Hürlimann | [42] | 0.3552 | -0.29 | -0.26 | -0.19 | 0.58 | -0.31 | 0.89 | -0.54 | 0.61 |
| Jankowska | [43] | 0.62 | -0.12 | -0.05 | -0.08 | 0.55 | -0.12 | 0.53 | -0.24 | 0.45 |
| Jankowska | [44] | 0.5 | -0.08 | -0.09 | -0.08 | 0.55 | -0.08 | 0.50 | -0.16 | 0.32 |
| Jayapal | [45] | -inf | 0.00 | -0.05 | -0.03 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Kern | [51] | -inf | 0.05 | -0.02 | 0.02 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Khonji | [53] | 0.482 | -0.07 | -0.05 | -0.07 | 0.62 | -0.06 | 0.39 | -0.10 | 0.26 |
| Kocher | [58] | 0.482 | -0.01 | -0.01 | -0.02 | 0.62 | -0.03 | 0.63 | -0.05 | 0.08 |
| Layton | [64] | 1 | -0.00 | -0.13 | -0.07 | 0.58 | -0.01 | 0.72 | -0.02 | 0.03 |
| Layton | [63] | 0.7057 | -0.08 | -0.08 | -0.09 | 0.61 | -0.08 | 0.30 | -0.16 | 0.53 |
| Ledesma | [66] | 1 | 0.00 | -0.13 | -0.07 | 0.55 | -0.13 | 0.40 | -0.26 | 0.65 |
| Maitra | [69] | 0.5 | -0.04 | -0.03 | -0.04 | 0.57 | -0.03 | 0.98 | -0.06 | 0.06 |
| Mayor | [71] | 0.2 | -0.04 | -0.01 | -0.02 | 0.57 | -0.02 | 0.49 | -0.04 | 0.08 |
| Mechti | [74] | 0.667 | -0.05 | -0.01 | -0.03 | 0.53 | -0.09 | 0.21 | -0.11 | 0.52 |
| Modaresi | [77] | 0.757 | -0.11 | -0.08 | -0.10 | 0.63 | -0.13 | 0.50 | -0.21 | 0.42 |
| Moreau | [81] | -inf | 0.00 | -0.04 | -0.02 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Moreau | [79] | 0.5177 | -0.00 | 0.00 | 0.00 | 0.61 | -0.01 | 0.79 | -0.01 | 0.01 |
| Moreau | [80] | 0.6246 | -0.04 | -0.01 | -0.03 | 0.59 | -0.07 | 0.22 | -0.07 | 0.32 |
| Nikolov | [82] | 0.448 | -0.01 | -0.01 | -0.01 | 0.59 | -0.01 | 0.29 | -0.02 | 0.07 |
| Pacheco | [83] | -inf | -0.30 | 0.00 | -0.15 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Petmanson | [84] | -inf | -0.19 | -0.15 | -0.14 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Satyam | [94] | 0.479 | -0.06 | -0.06 | -0.08 | 0.71 | -0.05 | 0.68 | -0.09 | 0.13 |
| Seidman | [95] | -inf | -0.12 | -0.10 | -0.09 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Solórzano | [96] | 0.907 | -0.01 | 0.01 | -0.00 | 0.52 | -0.03 | 0.09 | -0.06 | 0.67 |
| van Dam | [102] | -inf | 0.00 | -0.07 | -0.04 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vartapetiance | [103] | -inf | 0.00 | -0.11 | -0.06 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Vartapetiance | [104] | 1 | 0.00 | 0.00 | 0.00 | 0.52 | 0.00 | 0.84 | 0.00 | 0.00 |
| Vilarino | [107] | -inf | 0.00 | -0.01 | -0.00 | 0.50 | 0.00 | 0.00 | 0.00 | 0.00 |
| Zamani | [113] | 0.488 | -0.05 | -0.05 | -0.05 | 0.58 | -0.03 | 0.76 | -0.02 | 0.03 |

**PAN 2014 EN test dataset**

| **Verifier** | | | **PAN measures** | | | **Section 3** | | | | **Impact** |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | 0.418 | -0.20 | -0.16 | -0.21 | 0.70 | -0.19 | 0.81 | -0.12 | 0.15 |
| Bartoli | [7] | 0.613 | -0.07 | 0.01 | -0.03 | 0.62 | -0.06 | 0.77 | -0.11 | 0.14 |
| Bobicev | [8] | 0.3405 | -0.30 | -0.05 | -0.17 | 0.57 | -0.20 | 0.42 | -0.41 | 0.98 |
| Castillo | [17] | 1 | 0.01 | 0.01 | 0.01 | 0.62 | 0.01 | 0.63 | 0.02 | -0.05 |
| Castro | [18] | 0.8 | -0.18 | -0.15 | -0.16 | 0.62 | -0.13 | 0.76 | -0.25 | 0.33 |
| Feng | [24] | 0.59 | -0.05 | 0.00 | -0.03 | 0.64 | -0.05 | 0.88 | -0.06 | 0.07 |
| Fréry | [29] | 1 | -0.03 | -0.01 | -0.02 | 0.61 | -0.03 | 0.69 | -0.06 | 0.09 |
| Gómez-Adorno | [33] | 1 | -0.02 | -0.02 | -0.02 | 0.57 | -0.02 | 0.90 | -0.03 | 0.03 |
| Grozea | [34] | 0.43 | -0.03 | -0.04 | -0.03 | 0.60 | -0.03 | 0.72 | -0.06 | 0.08 |
| Gutierrez | [35] | 1 | 0.06 | -0.01 | 0.02 | 0.56 | 0.03 | 0.41 | 0.01 | -0.02 |
| Harvey | [40] | 0.002 | -0.02 | -0.02 | -0.02 | 0.55 | -0.03 | 0.18 | -0.06 | 0.33 |
| Hürlimann | [42] | 0.4895 | -0.36 | -0.28 | -0.27 | 0.61 | -0.31 | 0.60 | -0.33 | 0.55 |
| Jankowska | [43] | 0.16 | -0.14 | -0.04 | -0.07 | 0.52 | -0.11 | 0.80 | -0.21 | 0.26 |
| Jankowska | [44] | 0.284 | -0.19 | -0.02 | -0.09 | 0.56 | -0.15 | 0.84 | -0.29 | 0.35 |
| Jayapal | [45] | 1 | 0.00 | -0.13 | -0.06 | 0.65 | -0.13 | 0.46 | -0.25 | 0.54 |
| Kern | [51] | 0.31 | 0.05 | -0.00 | 0.03 | 0.57 | -0.02 | 0.92 | -0.04 | 0.04 |
| Khonji | [53] | 0.735 | -0.05 | -0.01 | -0.04 | 0.72 | -0.08 | 0.61 | -0.11 | 0.18 |
| Kocher | [58] | 0.45 | -0.01 | -0.03 | -0.02 | 0.64 | 0.01 | 0.70 | 0.01 | -0.03 |
| Layton | [64] | -inf | 0.01 | -0.00 | 0.00 | 0.50 | 0.01 | 1.00 | 0.00 | 0.00 |
| Layton | [63] | 1 | 0.00 | 0.00 | 0.00 | 0.51 | 0.00 | 0.98 | -0.01 | 0.01 |
| Ledesma | [66] | 1 | 0.00 | -0.05 | -0.03 | 0.52 | -0.05 | 0.10 | -0.10 | 1.00 |
| Maitra | [69] | 0.64 | -0.18 | -0.14 | -0.20 | 0.71 | -0.19 | 0.65 | 0.29 | -0.83 |
| Mayor | [71] | 0.2 | -0.01 | -0.01 | -0.01 | 0.61 | -0.01 | 0.71 | -0.01 | 0.01 |
| Mechti | [74] | 0.833 | 0.01 | 0.00 | 0.01 | 0.61 | 0.00 | 0.63 | 0.06 | -0.16 |
| Modaresi | [77] | 0.395 | 0.05 | 0.04 | 0.06 | 0.72 | 0.04 | 0.68 | 0.07 | -0.22 |
| Moreau | [81] | 1 | 0.00 | -0.01 | -0.00 | 0.59 | -0.01 | 0.25 | -0.01 | 0.04 |
| Moreau | [79] | 0.8037 | -0.14 | -0.07 | -0.11 | 0.61 | -0.13 | 0.42 | -0.26 | 0.62 |
| Moreau | [80] | 0.5627 | -0.03 | 0.00 | -0.02 | 0.63 | -0.06 | 0.72 | -0.02 | 0.03 |
| Nikolov | [82] | 0.451 | 0.09 | 0.00 | 0.04 | 0.52 | -0.03 | 0.06 | -0.06 | 1.00 |
| Pacheco | [83] | 0.7114 | 0.01 | 0.00 | 0.01 | 0.59 | -0.06 | 0.26 | -0.11 | 0.42 |
| Petmanson | [84] | 0.1 | -0.25 | -0.08 | -0.14 | 0.54 | -0.23 | 0.49 | -0.46 | 0.94 |
| Satyam | [94] | 0.523 | -0.10 | -0.05 | -0.09 | 0.63 | -0.10 | 0.56 | -0.20 | 0.36 |
| Seidman | [95] | 1 | 0.01 | 0.01 | 0.01 | 0.52 | 0.01 | 0.99 | 0.00 | 0.00 |
| Solórzano | [96] | 0.594 | -0.01 | -0.02 | -0.01 | 0.54 | -0.04 | 0.75 | -0.08 | 0.11 |
| van Dam | [102] | 1 | 0.00 | -0.02 | -0.01 | 0.52 | -0.02 | 0.50 | -0.03 | 0.06 |
| Vartapetiance | [103] | 1 | 0.00 | -0.07 | -0.04 | 0.52 | -0.07 | 0.14 | -0.14 | 1.00 |
| Vartapetiance | [104] | -inf | 0.00 | 0.00 | 0.00 | 0.50 | 0.01 | 1.00 | 0.00 | 0.00 |
| Vilarino | [107] | -inf | 0.00 | -0.00 | -0.00 | 0.50 | 0.01 | 1.00 | 0.00 | 0.00 |
| Zamani | [113] | 0.456 | -0.03 | -0.05 | -0.05 | 0.69 | -0.03 | 0.90 | -0.03 | 0.03 |

**PAN 2015 test dataset**

| **Verifier** | | | **PAN measures** | | | **Section 3** | | | | **Impact** |
|---|---|---|---|---|---|---|---|---|---|---|
| Team | [Reference] | $\tau$ | $\Delta_{\text{AUC}}$ | $\Delta_{\text{C@1}}$ | $\Delta_{\text{final}}$ | $\mathrm{acc}$ | $\Delta_{\mathrm{acc}}$ | $\mathrm{rec}$ | $\Delta_{\mathrm{rec}}$ | $\mathrm{imp}$ |
| Bagnall | [5] | 0.562 | -0.13 | -0.09 | -0.16 | 0.77 | -0.12 | 0.70 | -0.20 | 0.28 |
| Bartoli | [7] | 0.455 | 0.03 | 0.03 | 0.04 | 0.58 | 0.00 | 0.87 | 0.01 | -0.06 |
| Bobicev | [8] | 0.6893 | -0.21 | -0.11 | -0.18 | 0.67 | -0.17 | 0.66 | -0.34 | 0.52 |
| Castillo | [17] | 0.7 | 0.00 | 0.00 | 0.00 | 0.64 | 0.00 | 0.83 | 0.00 | -0.02 |
| Castro | [18] | 0.7 | -0.01 | -0.01 | -0.01 | 0.71 | -0.00 | 0.79 | -0.01 | 0.01 |
| Fréry | [29] | 0.143 | -0.00 | 0.01 | 0.00 | 0.55 | -0.01 | 0.52 | -0.02 | 0.03 |
| Gómez-Adorno | [33] | 1 | -0.01 | -0.01 | -0.01 | 0.53 | -0.01 | 0.96 | -0.01 | 0.01 |
| Grozea | [34] | 0.09 | -0.09 | -0.06 | -0.08 | 0.56 | -0.08 | 0.29 | -0.16 | 0.57 |
| Gutierrez | [35] | 0.7273 | -0.11 | -0.09 | -0.13 | 0.70 | -0.08 | 0.72 | -0.18 | 0.24 |
| Harvey | [40] | 0.502 | -0.00 | 0.01 | 0.00 | 0.62 | 0.00 | 0.70 | 0.01 | -0.03 |
| Hürlimann | [42] | 0.5238 | -0.22 | -0.17 | -0.21 | 0.64 | -0.18 | 0.56 | -0.19 | 0.34 |
| Jankowska | [43] | 0.6 | -0.07 | -0.05 | -0.07 | 0.62 | -0.07 | 0.67 | -0.14 | 0.21 |
| Jankowska | [44] | 0.407 | -0.10 | -0.12 | -0.13 | 0.67 | -0.06 | 0.83 | -0.12 | 0.14 |
| Jayapal | [45] | 1 | 0.00 | -0.02 | -0.01 | 0.51 | -0.02 | 0.63 | -0.03 | 0.05 |
| Kern | [51] | -inf | 0.05 | -0.00 | 0.03 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Khonji | [53] | 0.333 | -0.06 | -0.05 | -0.08 | 0.76 | -0.06 | 0.84 | -0.10 | 0.12 |
| Kocher | [58] | 0.5 | 0.01 | 0.00 | 0.00 | 0.70 | -0.01 | 0.60 | -0.01 | 0.02 |
| Layton | [64] | -inf | 0.00 | 0.00 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Layton | [63] | 0.2943 | -0.04 | 0.00 | -0.02 | 0.67 | -0.04 | 0.37 | -0.08 | 0.20 |
| Ledesma | [66] | 1 | 0.00 | -0.15 | -0.08 | 0.62 | -0.15 | 0.50 | -0.31 | 0.61 |
| Maitra | [69] | 0.42 | -0.01 | -0.01 | -0.01 | 0.59 | -0.00 | 0.82 | -0.00 | 0.00 |
| Mayor | [71] | 0.8 | -0.01 | 0.00 | -0.01 | 0.64 | -0.01 | 0.45 | -0.02 | 0.04 |
| Mechti | [74] | 0.917 | -0.04 | -0.01 | -0.02 | 0.51 | -0.07 | 0.26 | -0.04 | 0.17 |
| Modaresi | [77] | 0.201 | 0.05 | 0.03 | 0.03 | 0.50 | -0.00 | 1.00 | -0.00 | 0.00 |
| Moreau | [81] | -inf | 0.00 | 0.01 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Moreau | [80] | 0.5536 | -0.07 | -0.04 | -0.07 | 0.70 | -0.07 | 0.60 | -0.14 | 0.23 |
| Nikolov | [82] | 0.534 | -0.02 | -0.18 | -0.09 | 0.56 | -0.12 | 0.25 | -0.25 | 1.00 |
| Pacheco | [83] | 0.5650 | -0.02 | 0.01 | -0.00 | 0.71 | -0.06 | 0.59 | -0.12 | 0.20 |
| Petmanson | [84] | 0.99 | -0.00 | -0.05 | -0.03 | 0.55 | -0.04 | 0.30 | -0.08 | 0.27 |
| Pimas | [85] | 0.108 | 0.00 | 0.00 | 0.00 | 0.51 | 0.00 | 0.56 | 0.00 | -0.01 |
| Posadas-Durán | [86] | 0.8316 | -0.10 | -0.01 | -0.06 | 0.66 | -0.16 | 0.66 | -0.66 | 1.00 |
| Sari | [93] | 0.544 | 0.06 | 0.00 | 0.03 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Satyam | [94] | 0.145 | -0.07 | 0.00 | -0.04 | 0.72 | -0.06 | 0.48 | -0.12 | 0.25 |
| Seidman | [95] | 1 | -0.04 | -0.04 | -0.05 | 0.68 | -0.04 | 0.42 | -0.10 | 0.23 |
| Solórzano | [96] | 0.691 | 0.01 | 0.02 | 0.01 | 0.54 | -0.01 | 0.36 | -0.02 | 0.07 |
| van Dam | [102] | 1 | 0.00 | -0.02 | -0.01 | 0.59 | -0.02 | 0.60 | -0.02 | 0.03 |
| Vartapetiance | [103] | -inf | 0.00 | -0.03 | -0.02 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Vartapetiance | [104] | 1 | -0.01 | -0.01 | -0.02 | 0.60 | -0.01 | 0.40 | -0.03 | 0.07 |
| Vilarino | [107] | -inf | 0.00 | 0.00 | 0.00 | 0.50 | 0.00 | 1.00 | 0.00 | 0.00 |
| Zamani | [113] | 0.761 | -0.01 | 0.00 | -0.01 | 0.71 | -0.02 | 0.54 | -0.01 | 0.02 |

assessor's observation on these sample cases was that most of the paraphrased texts of each particular approach have a very similar characteristic with respect to sensibleness and soundness. The final decision then was to base the manual assessment on just one random text from each year of PAN's test datasets, excluding the original texts from language learners (their original text quality might already be suboptimal): in-depth manual assessment was performed on problem instances 5, 134, and 429.

The human assessor started by reading the obfuscated texts without knowing which was the output of what approach. During this reading phase, the assessor marked up errors (typos, grammar) and assigned school grades (on a scale from 1 (excellent) to 5 (fail)) for the sensibleness of each of the sample problem instances. As a result, the obfuscated texts of Mansoorizadeh *et al.*'s approach got a grade 2 for all three cases mainly due to the many punctuation problems where a white space was inserted before every punctuation mark. The texts of Mihaylova *et al.*'s obfuscator all got a grade 4 due to the many grammatical errors, capitalizing issues (lower-case sentence starts, capitalized words in the middle of sentences), punctuation problems (many missing or useless punctuations), and consistent typos ("tto" and "oof"). The assessor noted that the texts were difficult to read due to the many problems but that grade 4 was given to show the difference to the even worse texts of Keswani *et al.*'s obfuscator. For Keswani *et al.*'s approach, our assessor noted that the texts were impossible to read or understand with lots of grammatical errors, capitalization problems at sentence beginnings, etc. The assessor even wanted to stop reading before finishing the whole text due to the "painful" experience.

After grading the sensibleness of the obfuscated texts, the assessor read the original texts and used the visual analytics tool highlighting the textual differences in various ways to assess the soundness of the obfuscated texts on a three-point scale as correct, passable, or incorrect. The obfuscated texts of Mihaylova *et al.*'s and Keswani *et al.*'s approaches were both judged "incorrect" for all three cases since they are almost impossible to read. Mihaylova *et al.*'s obfuscator might produce slightly more sound texts than Keswani *et al.*'s, yet, the assessor did not want to assign a "passable" to any of these but suggested to further differentiate the point scale as a future evaluation improvement. Not that surprising, Mansoorizadeh *et al.*'s very conservative approach achieved "correct" as a score for the problem instances 134 and 429. However, some of the changes for problem instance 5 did change the meaning a lot ("car" instead of "machine" when a computer was meant, etc.) such that for problem instance 5, Mansoorizadeh *et al.*'s approach only got a "passable" score.

### 5.3 Peer Evaluation Results

Three groups participated in the peer evaluation; one external group (Liebeck *et al.* [67]) and two of the author obfuscation task participants (Mansoorizadeh *et al.* and Mihaylova *et al.*).

As for the safety evaluation, all three groups employ the GLAD authorship verification approach [42] and checked how many of its decisions are changed by obfuscation. Interestingly, two of the peer review teams agree that Mihaylova *et al.*'s approach is safest against the GLAD system while Keswani *et al.*'s obfuscator is ranked second;

in Mansoorizadeh *et al.* peer review, however, Keswani *et al.* is ranked before Mansoorizadeh *et al.*, and Mihaylova *et al.* is ranked third, which is claimed to have hardly any effect. In our evaluation against many more than just one verification approach, we agree with the former two peer reviews in that Mihaylova *et al.* is the safest approach while Keswani *et al.*'s approach ranks second.

As for soundness, Liebeck *et al.* suggest an automatic measure based on approaches submitted to the shared tasks on semantic textual similarity (STS) organized at SemEval, where an optimum soundness would achieve a score of 5 and the worst soundness a score of 0. In Liebeck *et al.*'s evaluation, all three obfuscation approaches achieve scores above 4, where Mansoorizadeh *et al.* performs almost perfect and Keswani *et al.* the worst. Still, given our own soundness evaluation, the scores for the approaches of Keswani *et al.* and Mihaylova *et al.* with 4+ out of 5 seem a too optimistic and should be much closer to 0. Similarly, also the peer review of Mansoorizadeh *et al.* scores soundness from 5 down to 0 but by a manual judgment. Here, their own obfuscation approach outperforms the others by far (4.86), while Mihaylova *et al.* achieve a 3.93 and Keswani *et al.* a 2.02; the same ordering as in Liebeck *et al.*'s evaluation but probably also rather high scores for the two worse approaches. The peer review of Mihaylova *et al.* is also based on manual scoring for soundness and seems to come to the conclusion that their own approach is the best, followed by Mansoorizadeh *et al.*'s obfuscations. However, this seems to be a rather biased outcome since our own manual soundness evaluation showed the soundness of Mihaylova *et al.*'s obfuscation approach to be inferior to Mansoorizadeh *et al.*'s approach.

As for sensibleness, all three peer evaluators opt for a manual analysis and rank the obfuscations of Mansoorizadeh *et al.* clearly more sensible than Mihaylova *et al.*'s that again are more sensible than Keswani *et al.*'s obfuscations. Interestingly, Mansoorizadeh *et al.* employ a scale from 0–5 with 5 as most sensible, Liebeck *et al.* employ a three point scale, while Mihalyova *et al.* only have a two point scale (sensible or not) and grade half of a sample of their own approach's obfuscations as sensible. Just as is the case for Mihalyova *et al.*'s peer review of soundness, this somewhat contradicts our own manual sensibleness evaluation, but in sum the ordering of the approaches of all three peer reviews is consistent with our own sensibleness grading.

Not surprisingly, the approach of Mansoorizadeh *et al.* that hardly changes anything in a text, except for introducing many spaces before punctuation marks, achieves good and very good scores for sensibleness and soundness but is the least safe of the tested obfuscators. Although hardly being sound or sensible, the texts produced by the safest obfuscator of Mihaylova *et al.* have a slightly better quality compared to the round-trip translations produced by Keswani *et al.*'s obfuscator. Most of the three external peer reviews agree with these evaluation results at least on the relative ordering of the individual obfuscators.

## 6 Conclusion and Outlook

We have conducted the first large-scale evaluation of author obfuscation approaches in terms of their safety against the state of the art in authorship verification. A total of 44 verification approaches have been tested as to their vulnerability to obfuscation, and we found that many of them are indeed vulnerable to a greater or lesser extent. Moreover, for the first time, we have shown that author obfuscation technology can take on many authorship verification approach simultaneously, which is a must if this technology is supposed to be useful in practice. The best-performing obfuscator flips on average about 47% of an authorship verifier's decisions towards choosing "different author" when the opposite decision would have been correct. The obfuscation approaches evaluated have been collected via a shared task on author obfuscation that we organized at PAN 2016; three obfuscators have been submitted which are now hosted on the TIRA evaluation-as-a-service platform, ready for re-evaluation against new authorship verification approaches. Furthermore, we have systematically reviewed the literature on author obfuscation and collected and organized for the first time its evaluation methodology, introducing the three main performance dimensions of an author obfuscator: safety, soundness, and sensibleness.

There are still many open challenges when it comes to evaluating author obfuscation approaches properly and at scale, some requiring original research into new technologies that are capable of recognizing paraphrases, textual entailment, grammaticality, and style deception. Conceivably, approaches to these problems can be devised which are tailored to the evaluation of author obfuscation approaches and therefore exploit certain aspect of this application domain to achieve better performance than in the general case. We leave a more detailed investigation in this direction for future work.

### Acknowledgements

We thank the participating teams of this shared task. Our special thanks go to Adobe Systems Inc. for sponsoring the event.

## Bibliography

1. Abbasi, A., Chen, H.: Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Trans. Inf. Syst. 26(2), 7:1–7:29 (Apr 2008), http://doi.acm.org/10.1145/1344411.1344413
2. Afroz, S., Brennan, M., Greenstadt, R.: Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In: 2012 IEEE Symposium on Security and Privacy. pp. 461–475 (May 2012)
3. Almishari, M., Oguz, E., Tsudik, G.: Fighting Authorship Linkability with Crowdsourcing. In: Sala, A., Goel, A., Gummadi, K. (eds.) Proceedings of the second ACM conference on Online social networks, COSN 2014, Dublin, Ireland, October 1-2, 2014. pp. 69–82. ACM (2014), http://doi.acm.org/10.1145/2660460.2660486
4. Backes, M., Berrang, P., Manoharan, P.: Poster: Assessing the effectiveness of countermeasures against authorship recognition. In: 2015 IEEE Symposium on Security and Privacy, SP 2015, San Jose, CA, USA, May 17-21, 2015. IEEE Computer Society (2015), http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=7160813
5. Bagnall, D.: Author Identification using multi-headed Recurrent Neural Networks—Notebook for PAN at CLEF 2015. In: [16]
6. Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.): CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal. CEUR Workshop Proceedings, CEUR-WS.org (2016), http://www.clef-initiative.eu/publication/working-notes

7. Bartoli, A., Dagri, A., De Lorenzo, A., Medvet, E., Tarlao, F.: An Author Verification Approach Based on Differential Features—Notebook for PAN at CLEF 2015. In: [16]
8. Bobicev, V.: Authorship Detection with PPM—Notebook for PAN at CLEF 2013. In: [26]
9. Brants, T., Franz, A.: Web 1T 5-gram Version 1. Linguistic Data Consortium LDC2006T13, Philadelphia (2006)
10. Brennan, M., Afroz, S., Greenstadt, R.: Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity. ACM Trans. Inf. Syst. Secur. 15(3), 12 (2012), http://doi.acm.org/10.1145/2382448.2382450
11. Brennan, M., Greenstadt, R.: Practical Attacks Against Authorship Recognition Techniques. In: Haigh, K., Rychtyckyj, N. (eds.) Proceedings of the Twenty-First Conference on Innovative Applications of Artificial Intelligence, July 14-16, 2009, Pasadena, California, USA. AAAI (2009), http://aaai.org/ocs/index.php/IAAI/IAAI09/paper/view/257
12. Burrows, S., Potthast, M., Stein, B.: Paraphrase Acquisition via Crowdsourcing and Machine Learning. Transactions on Intelligent Systems and Technology (ACM TIST) 4(3), 43:1–43:21 (Jun 2013), http://dl.acm.org/citation.cfm?id=2483676
13. Caliskan, A., Greenstadt, R.: Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text. In: Sixth IEEE International Conference on Semantic Computing, ICSC 2012, Palermo, Italy, September 19-21, 2012. pp. 121–125. IEEE Computer Society (2012), http://dx.doi.org/10.1109/ICSC.2012.46
14. Callison-Burch, C., Cohn, T., Lapata, M.: Parametric: An automatic evaluation metric for paraphrasing. In: Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). pp. 97–104. Coling 2008 Organizing Committee, Manchester, UK (August 2008), http://www.aclweb.org/anthology/C08-1013
15. Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.): CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK. CEUR Workshop Proceedings, CEUR-WS.org (2014), http://www.clef-initiative.eu/publication/working-notes
16. Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.): CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France. CEUR Workshop Proceedings, CEUR-WS.org (2015), http://www.clef-initiative.eu/publication/working-notes
17. Castillo, E., Cervantes, O., Vilariño, D., Pinto, D., , León, S.: Unsupervised Method for the Authorship Identification Task—Notebook for PAN at CLEF 2014. In: [15]
18. Castro, D., Adame, Y., Pelaez, M., Muñoz, R.: Authorship Verification, Combining Linguistic Features and Different Similarity Functions—Notebook for PAN at CLEF 2015. In: [16]
19. Chang, C.Y., Clark, S.: Linguistic steganography using automatically generated paraphrases. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. pp. 591–599. Association for Computational Linguistics, Los Angeles, California (June 2010), http://www.aclweb.org/anthology/N10-1084
20. Chen, D., Dolan, W.: Collecting Highly Parallel Data for Paraphrase Evaluation. In: Lin, D., Matsumoto, Y., Mihalcea, R. (eds.) Proceedings of the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 190–200. Association for Computational Linguistics, Portland, Oregon (Jun 2011)
21. Cherry, C., Quirk, C.: Discriminative, Syntactic Language Modeling through Latent SVMs. In: Proceedings of AMTA (2008), http://research.microsoft.com/pubs/72874/lsvm_amta.pdf

22. Clark, J.H., Hannon, C.J.: A Classifier System for Author Recognition Using Synonym-Based Features, pp. 839–849. Springer (2007), http://dx.doi.org/10.1007/978-3-540-76631-5_80
23. Dagan, I., Roth, D., Sammons, M., Zanzotto, F.M.: Recognizing Textual Entailment: Models and Applications. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2013), http://dx.doi.org/10.2200/S00509ED1V01Y201305HLT023
24. Feng, V., Hirst, G.: Authorship Verification with Entity Coherence and Other Rich Linguistic Features—Notebook for PAN at CLEF 2013. In: [26]
25. Ferraro, F., Post, M., Van Durme, B.: Judging grammaticality with count-induced tree substitution grammars. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP. pp. 116–121. Association for Computational Linguistics, Montréal, Canada (June 2012), http://www.aclweb.org/anthology/W12-2013
26. Forner, P., Navigli, R., Tufis, D. (eds.): CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain (2013), http://www.clef-initiative.eu/publication/working-notes
27. Francis, W.N., Kucera, H.: Brown corpus manual. Brown University (1979)
28. Fratila, S.: Submission to the Author Identification Task from the Polytechnic University of Bucharest, Romania. http://www.uni-weimar.de/medien/webis/events/pan-13 (2013), http://www.clef-initiative.eu/publication/working-notes
29. Fréry, J., Largeron, C., Juganaru-Mathieu, M.: UJM at CLEF in Author Identification—Notebook for PAN at CLEF 2014. In: [15]
30. Ganitkevitch, J., Van Durme, B., Callison-Burch, C.: Ppdb: The paraphrase database. In: Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 9-14, 2013, Westin Peachtree Plaza Hotel, Atlanta, Georgia, USA. pp. 758–764 (2013)
31. Ghaeini, M.: Intrinsic Author Identification Using ModifiedWeighted KNN—Notebook for PAN at CLEF 2013. In: [26]
32. Gollub, T., Stein, B., Burrows, S.: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Hersh, B., Callan, J., Maarek, Y., Sanderson, M. (eds.) 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12). pp. 1125–1126. ACM (Aug 2012)
33. Gómez-Adorno, H., Sidorov, G., Pinto, D., Markov, I.: A Graph Based Authorship Identification Approach—Notebook for PAN at CLEF 2015. In: [16]
34. Grozea, C., Popescu, M.: Submission to the Author Identification Task from Fraunhofer FOKUS, Germany, and the University of Bucharest, Romania. http://www.uni-weimar.de/medien/webis/events/pan-13 (2013), http://www.clef-initiative.eu/publication/working-notes, From Fraunhofer FOKUS and the University of Bucharest
35. Gutierrez, J., Casillas, J., Ledesma, P., Fuentes, G., Meza, I.: Homotopy Based Classification for Author Verification Task—Notebook for PAN at CLEF 2015. In: [16]
36. Halvani, O., Steinebach, M.: VEBAV - A Simple, Scalable and Fast Authorship Verification Scheme—Notebook for PAN at CLEF 2014. In: [15]
37. Halvani, O., Steinebach, M., Zimmermann, R.: Authorship Verification via k-Nearest Neighbor Estimation—Notebook for PAN at CLEF 2013. In: [26]
38. Halvani, O., Winter, C.: A Generic Authorship Verification Scheme Based on Equal Error Rates—Notebook for PAN at CLEF 2015. In: [16]
39. Hanbury, A., Müller, H., Balog, K., Brodt, T., Cormack, G., Eggel, I., Gollub, T., Hopfgartner, F., Kalpathy-Cramer, J., Kando, N., Krithara, A., Lin, J., Mercer, S., Potthast, M.: Evaluation-as-a-Service: Overview and Outlook. ArXiv e-prints (Dec 2015), http://arxiv.org/abs/1512.07454

40. Harvey, S.: Author Verification using PPM with Parts of Speech Tagging—Notebook for PAN at CLEF 2014. In: [15]
41. Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., Tetreault, J.: Predicting grammaticality on an ordinal scale. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 174–180. Association for Computational Linguistics, Baltimore, Maryland (June 2014), http://www.aclweb.org/anthology/P14-2029
42. Hürlimann, M., Weck, B., van den Berg, E., Šuster, S., Nissim, M.: GLAD: Groningen Lightweight Authorship Detection—Notebook for PAN at CLEF 2015. In: [16]
43. Jankowska, M., Kešelj, V., , Milios, E.: Proximity based One-class Classification with Common N-Gram Dissimilarity for Authorship Verification Task—Notebook for PAN at CLEF 2013. In: [26]
44. Jankowska, M., Kešelj, V., Milios, E.: Ensembles of Proximity-Based One-Class Classifiers for Author Verification—Notebook for PAN at CLEF 2014. In: [15]
45. Jayapal, A., Goswami, B.: Vector space model and Overlap metric for Author Identification—Notebook for PAN at CLEF 2013. In: [26]
46. Juola, P.: Detecting stylistic deception. In: Proceedings of the Workshop on Computational Approaches to Deception Detection. pp. 91–96. Association for Computational Linguistics, Avignon, France (April 2012), http://www.aclweb.org/anthology/W12-0414
47. Juola, P., Stamatatos, E.: Overview of the Author Identification Task at PAN 2013. In: [26]
48. Juola, P., Vescovi, D.: Empirical Evaluation of Authorship Obfuscation using JGAAP. In: Greenstadt, R. (ed.) Proceedings of the 3rd ACM Workshop on Security and Artificial Intelligence, AISec 2010, Chicago, Illinois, USA, October 8, 2010. pp. 14–18. ACM (2010), http://doi.acm.org/10.1145/1866423.1866427
49. Juola, P., Vescovi, D.: Analyzing Stylometric Approaches to Author Obfuscation. In: Peterson, G., Shenoi, S. (eds.) Advances in Digital Forensics VII - 7th IFIP WG 11.9 International Conference on Digital Forensics, Orlando, FL, USA, January 31 - February 2, 2011, Revised Selected Papers. IFIP Advances in Information and Communication Technology, vol. 361, pp. 115–125. Springer (2011), http://dx.doi.org/10.1007/978-3-642-24212-0_9
50. Kacmarcik, G., Gamon, M.: Obfuscating Document Stylometry to Preserve Author Anonymity. In: Calzolari, N., Cardie, C., Isabelle, P. (eds.) ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006. The Association for Computer Linguistics (2006), http://aclweb.org/anthology/P06-2058
51. Kern, R.: Grammar Checker Features for Author Identification and Author Profiling—Notebook for PAN at CLEF 2013. In: [26]
52. Keswani, Y., Trivedi, H., Mehta, P., Majumder, P.: Author Masking through Translation—Notebook for PAN at CLEF 2016. In: [6], http://ceur-ws.org/Vol-1609/
53. Khonji, M., Iraqi, Y.: A Slightly-modified GI-based Author-verifier with Lots of Features (ASGALF)—Notebook for PAN at CLEF 2014. In: [15]
54. Khosmood, F.: Comparison of Sentence-level Paraphrasing Approaches for Statistical Style Transformation. In: Proceedings of the 2012 International Conference on Artificial Intelligence. CSREA Press, Las Vegas (2012)
55. Khosmood, F., Levinson, R.: Toward Automated Stylistic Transformation of Natural Language Text. In: Proceedings of the Digital Humanities 2009, June 22-25. pp. 177–181 (2009)
56. Khosmood, F., Levinson, R.: Automatic Synonym and Phrase Replacement Show Promise for Style Transformation. In: Draghici, S., Khoshgoftaar, T., Palade, V., Pedrycz, W.,

Wani, M., Zhu, X. (eds.) The Ninth International Conference on Machine Learning and Applications, ICMLA 2010, Washington, DC, USA, 12-14 December 2010. pp. 958–961. IEEE Computer Society (2010), http://dx.doi.org/10.1109/ICMLA.2010.153

57. Khosmood, F., Levinson, R.A.: Automatic Natural Language Style Classification and Transformation. In: Proceedings of the 2008 BCS-IRSG Conference on Corpus Profiling. pp. 3–3. IRSG'08, British Computer Society, Swinton, UK, UK (2008), http://dl.acm.org/citation.cfm?id=2227976.2227980
58. Kocher, M., Savoy, J.: UniNE at CLEF 2015: Author Identification—Notebook for PAN at CLEF 2015. In: [16]
59. Koehn, P.: Europarl: A Parallel Corpus for Statistical Machine Translation. In: Conference Proceedings: the tenth Machine Translation Summit. pp. 79–86. AAMT, AAMT, Phuket, Thailand (2005), http://mt-archive.info/MTS-2005-Koehn.pdf
60. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. pp. 177–180. ACL'07, Association for Computational Linguistics, Stroudsburg, PA, USA (2007), http://dl.acm.org/citation.cfm?id=1557769.1557821
61. Koppel, M., Schler, J.: Authorship Verification as a One-Class Classification Problem. In: Brodley, C. (ed.) Proceedings of the Twenty-First International Conference on Machine Learning. pp. 1–7. ACM (Jul 2004)
62. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology 65(1), 178–187 (2014), http://dx.doi.org/10.1002/asi.22954
63. Layton, R.: A simple Local n-gram Ensemble for Authorship Verification—Notebook for PAN at CLEF 2014. In: [15]
64. Layton, R., Watters, P., Dazeley, R.: Local n-grams for Author Identification—Notebook for PAN at CLEF 2013. In: [26]
65. Le, H., Safavi-Naini, R., Galib, A.: Secure Obfuscation of Authoring Style. In: Akram, R., Jajodia, S. (eds.) Information Security Theory and Practice - 9th IFIP WG 11.2 International Conference, WISTP 2015 Heraklion, Crete, Greece, August 24-25, 2015 Proceedings. Lecture Notes in Computer Science, vol. 9311, pp. 88–103. Springer (2015), http://dx.doi.org/10.1007/978-3-319-24018-3_6
66. Ledesma, P., Fuentes, G., Jasso, G., Toledo, A., , Meza, I.: Distance learning for Author Verification—Notebook for PAN at CLEF 2013. In: [26]
67. Liebeck, M., Modaresi, P., Conrad, S.: Evaluating Safety, Soundness and Sensibleness of Obfuscation Systems—Notebook for PAN at CLEF 2016. In: [6], http://ceur-ws.org/Vol-1609/
68. Liu, C., Dahlmeier, D., Ng, H.T.: PEM: A paraphrase evaluation metric exploiting parallel texts. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. pp. 923–932. Association for Computational Linguistics, Cambridge, MA (October 2010), http://www.aclweb.org/anthology/D10-1090
69. Maitra, P., Ghosh, S., Das, D.: Authorship Verification - An Approach based on Random Forest—Notebook for PAN at CLEF 2015. In: [16]
70. Mansoorizadeh, M., Rahgooy, T., Aminiyan, M., Eskandari, M.: Author Obfuscation using WordNet and Language Models—Notebook for PAN at CLEF 2016. In: [6], http://ceur-ws.org/Vol-1609/
71. Mayor, C., Gutierrez, J., Toledo, A., Martinez, R., Ledesma, P., Fuentes, G., , Meza, I.: A Single Author Style Representation for the Author Verification Task—Notebook for PAN at CLEF 2014. In: [15]

72. McDonald, A., Afroz, S., Caliskan, A., Stolerman, A., Greenstadt, R.: Use Fewer Instances of the Letter "i": Toward Writing Style Anonymization. In: Fischer-Hübner, S., Wright, M. (eds.) Privacy Enhancing Technologies - 12th International Symposium, PETS 2012, Vigo, Spain, July 11-13, 2012. Proceedings. Lecture Notes in Computer Science, vol. 7384, pp. 299–318. Springer (2012), http://dx.doi.org/10.1007/978-3-642-31680-7_16
73. McDonald, A., Ulman, J., Barrowclift, M., Greenstadt, R.: Anonymouth Revamped: Getting Closer to Stylometric Anonymity. In: Kapadia, A., Caine, K., Camp, L., Lee, A., Patil, S., Reiter, M., Staddon, J. (eds.) Proceedings of the Workshop on Privacy Enhancing Tools PETools, Bloomington, Indiana, USA, July 9, 2013. (2013)
74. Mechti, S., Jaoua, M., Faiz, R., Belguith, L., Bsir, B.: On the Empirical Evaluation of Author Identification Hybrid Method—Notebook for PAN at CLEF 2015. In: [16]
75. Mihaylova, T., Karadjov, G., Nakov, P., Kiprov, Y., Georgiev, G., Koychev, I.: SU@PAN'2016: Author Obfuscation—Notebook for PAN at CLEF 2016. In: [6], http://ceur-ws.org/Vol-1609/
76. Miller, G.A.: Wordnet: A lexical database for english. Commun. ACM 38(11), 39–41 (Nov 1995), http://doi.acm.org/10.1145/219717.219748
77. Modaresi, P., Gross, P.: A Language Independent Author Verifier Using Fuzzy C-Means Clustering—Notebook for PAN at CLEF 2014. In: [15]
78. Moon, S.W., Gweon, G., Choi, H., Heo, J.: Apem: Automatic paraphrase evaluation using morphological analysis for the korean language. In: 2016 18th International Conference on Advanced Communication Technology (ICACT). pp. 680–684. IEEE (2016)
79. Moreau, E., Jayapal, A., , Vogel, C.: Author Verification: Exploring a Large set of Parameters using a Genetic Algorithm—Notebook for PAN at CLEF 2014. In: [15]
80. Moreau, E., Jayapal, A., Lynch, G., Vogel, C.: Author Verification: Basic Stacked Generalization Applied To Predictions from a Set of Heterogeneous Learners—Notebook for PAN at CLEF 2015. In: [16]
81. Moreau, E., Vogel, C.: Style-based Distance Features for Author Verification—Notebook for PAN at CLEF 2013. In: [26]
82. Nikolov, S., Tabakova, D., Savov, S., Kiprov, Y., Nakov, P.: SUPAN'2015: Experiments in Author Verification—Notebook for PAN at CLEF 2015. In: [16]
83. Pacheco, M., Fernandes, K., Porco, A.: Random Forest with Increased Generalization: A Universal Background Approach for Authorship Verification—Notebook for PAN at CLEF 2015. In: [16]
84. Petmanson, T.: Authorship Identification using Correlations of Frequent Features—Notebook for PAN at CLEF 2013. In: [26]
85. Pimas, O., Kröll, M., Kern, R.: Know-Center at PAN 2015 Author Identification—Notebook for PAN at CLEF 2015. In: [16]
86. Posadas-Durán, J.P., Sidorov, G., Batyrshin, I., Mirasol-Meléndez, E.: Author Verification Using Syntactic N-grams—Notebook for PAN at CLEF 2015. In: [16]
87. Post, M.: Judging grammaticality with tree substitution grammar derivations. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 217–222. Association for Computational Linguistics, Portland, Oregon, USA (June 2011), http://www.aclweb.org/anthology/P11-2038
88. Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J., Köhler, J., Lötzsch, W., Müller, F., Müller, M., Paßmann, R., Reinke, B., Rettenmeier, L., Rometsch, T., Sommer, T., Träger, M., Wilhelm, S., Stein, B., Stamatatos, E., Hagen, M.: Who Wrote the Web? Revisiting Influential Author Identification Research Applicable to Information Retrieval. In: Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Di Nunzio, G., Hauff, C., Silvello, G. (eds.) Advances in Information Retrieval. 38th European Conference on IR Resarch (ECIR 16). Lecture Notes in Computer Science, vol. 9626, pp. 393–407. Springer, Berlin Heidelberg New York (Mar 2016)

89. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (Sep 2014)
90. Potthast, M., Trenkmann, M., Stein, B.: Netspeak: Assisting Writers in Choosing Words. In: Gurrin, C., He, Y., Kazai, G., Kruschwitz, U., Little, S., Roelleke, T., Rüger, S., van Rijsbergen, K. (eds.) Advances in Information Retrieval. 32nd European Conference on Information Retrieval (ECIR 10). Lecture Notes in Computer Science, vol. 5993, p. 672. Springer, Berlin Heidelberg New York (Mar 2010)
91. Rao, J., Rohatgi, P.: Can Pseudonymity Really Guarantee Privacy? In: Bellovin, S., Rose, G. (eds.) 9th USENIX Security Symposium, Denver, Colorado, USA, August 14-17, 2000. USENIX Association (2000), https://www.usenix.org/conference/9th-usenix-security-symposium/can-pseudonymity-really-guarantee-privacy
92. Rosso, P., Rangel, F., Potthast, M., Stamatatos, E., Tschuggnall, M., Stein, B.: Overview of PAN'16—New Challenges for Authorship Analysis: Cross-genre Profiling, Clustering, Diarization, and Obfuscation. In: Fuhr, N., Quaresma, P., Larsen, B., Gonçalves, T., Balog, K., Macdonald, C., Cappellato, L., Ferro, N. (eds.) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 16). Springer, Berlin Heidelberg New York (Sep 2016)
93. Sari, Y., Stevenson, M.: A Machine Learning-based Intrinsic Method for Cross-topic and Cross-genre Authorship Verification—Notebook for PAN at CLEF 2015. In: [16]
94. Satyam, Anand, Dawn, A., , Saha, S.: Statistical Analysis Approach to Author Identification Using Latent Semantic Analysis—Notebook for PAN at CLEF 2014. In: [15]
95. Seidman, S.: Authorship Verification Using the Impostors Method—Notebook for PAN at CLEF 2013. In: [26]
96. Solórzano, J., Mijangos, V., Pimentel, A., López-Escobedo, F., Montes, A., Sierra, G.: Authorship Verification by Combining SVMs with Kernels Optimized for Different Feature Categories—Notebook for PAN at CLEF 2015. In: [16]
97. Stamatatos, E., amd Ben Verhoeven, W.D., Juola, P., López-López, A., Potthast, M., Stein, B.: Overview of the Author Identification Task at PAN 2015. In: [16]
98. Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast, M., Stein, B., Juola, P., Sanchez-Perez, M., Barrón-Cedeño, A.: Overview of the Author Identification Task at PAN 2014. In: [15]
99. Stein, B., Hagen, M., Bräutigam, C.: Generating Acrostics via Paraphrasing and Heuristic Search. In: Tsujii, J., Hajic, J. (eds.) 25th International Conference on Computational Linguistics (COLING 14). pp. 2018–2029. Association for Computational Linguistics (Aug 2014)
100. Sun, G., Liu, X., Cong, G., Zhou, M., Xiong, Z., Lee, J., Lin, C.: Detecting erroneous sentences using automatically mined sequential patterns. In: Carroll, J.A., van den Bosch, A., Zaenen, A. (eds.) ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic. The Association for Computational Linguistics (2007), http://aclweb.org/anthology-new/P/P07/P07-1011.pdf
101. Tweedie, F.J., Singh, S., Holmes, D.I.: Neural Network Applications in Stylometry: The Federalist Papers. Computers and the Humanities 30(1), 1–10 (1996), http://dx.doi.org/10.1007/BF00054024
102. van Dam, M.: A Basic Character N-gram Approach to Authorship Verification—Notebook for PAN at CLEF 2013. In: [26]

103. Vartapetiance, A., Gillam, L.: A Textual Modus Operandi: Surrey's Simple System for Author Identification—Notebook for PAN at CLEF 2013. In: [26]
104. Vartapetiance, A., Gillam, L.: A Trinity of Trials: Surrey's 2014 Attempts at Author Verification—Notebook for PAN at CLEF 2014. In: [15]
105. Vartapetiance, A., Gillam, L.: Adapting for Subject-Specific Term Length using Topic Cost in Author Verification—Notebook for PAN at CLEF 2015. In: [16]
106. Veenman, C., Li, Z.: Authorship Verification with Compression Features. In: [26]
107. Vilariño, D., Pinto, D., Gómez, H., León, S., Castillo, E.: Lexical-Syntactic and Graph-Based Features for Authorship Verification—Notebook for PAN at CLEF 2013. In: [26]
108. Wagner, J., Foster, J.: The effect of correcting grammatical errors on parse probabilities. In: Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09). pp. 176–179. Association for Computational Linguistics, Paris, France (October 2009), http://www.aclweb.org/anthology/W09-3827
109. Wagner, J., Foster, J., van Genabith, J.: Judging grammaticality: Experiments in sentence classification. CALICO Journal 26(3), 474–490 (2009)
110. Weese, J., Ganitkevitch, J., Callison-Burch, C.: Paradigm: Paraphrase diagnostics through grammar matching. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 192–201. Association for Computational Linguistics, Gothenburg, Sweden (April 2014), http://www.aclweb.org/anthology/E14-1021
111. Wong, S.M.J., Dras, M.: Parser features for sentence grammaticality classification. In: Proceedings of the Australasian Language Technology Association Workshop 2010. pp. 67–75. Melbourne, Australia (December 2010)
112. Xu, W., Ritter, A., Dolan, B., Grishman, R., Cherry, C.: Paraphrasing for style. In: Proceedings of COLING 2012. pp. 2899–2914. The COLING 2012 Organizing Committee, Mumbai, India (December 2012), http://www.aclweb.org/anthology/C12-1177
113. Zamani, H., Abnar, S., Dehghani, M., Forati, M., Babaei, P.: Submission to the Author Identification Task at PAN 2014. http://www.uni-weimar.de/medien/webis/events/pan-14 (2014), http://www.clef-initiative.eu/publication/working-notes, From the University of Tehran, Iran
114. Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 57(3), 378–393 (2006)