Overview of the Author Obfuscation Task at PAN 2018: A New Approach to Measuring Safety

Martin Potthast, Felix Schremmer, Matthias Hagen, and Benno Stein

Leipzig University
martin.potthast@uni-leipzig.de

University of Bonn
felix.schremmer@uni-bonn.de

Martin-Luther-Universität Halle-Wittenberg
matthias.hagen@informatik.uni-halle.de

Bauhaus-Universität Weimar
benno.stein@uni-weimar.de

Abstract In this paper, we evaluate seven author obfuscation approaches which are supposed to automatically mask an author's writing style in a given text to render automatic author identification impossible. The approaches are evaluated with regard to their safety, soundness, and sensibleness in terms of beating 44 author identification approaches, retaining the original meaning of the obfuscated text, and producing inconspicuous, human-readable obfuscations, respectively. Regarding the measurement of safety in particular, we introduce a set of new performance measures which are designed to render the performance of obfuscation approaches comparable as the numbers of author identification approaches and evaluation datasets increases, incorporating their respective performance and quality. Based on the new measures, we establish a world ranking of obfuscators.

1 Introduction

Author obfuscation is the adversary task to author identification. The goal of obfuscation is to render the identification of authors based on their writing style impossible or at least intractable. Consequently, an effective identification approach has to be robust against obfuscation, or else it cannot be trusted. The fundamental question underlying both tasks is whether writing style can be purposefully manipulated. We hypothesize that this is indeed the case, and that style manipulations sufficient to counter identification can be accomplished in a way indistinguishable from genuine writing. Our goal is to foster the development of new technology in this respect, and its evaluation.

We consider three performance dimensions according to which an author obfuscation approach must excel to be considered fit for practical use. Obviously, the obfuscation performance should depend on the capability of fooling forensic experts—be it a piece of software or a human. However, fulfilling this requirement in isolation will disregard writers and their target audience, whose primary goal is to communicate, albeit safe from deanonymization: the quality of an obfuscated text along with the fact that its semantics is preserved are equally important to fool authorship identification. We hence call an obfuscation software

1. **safe**, if its obfuscated texts cannot be attributed to their original authors anymore,
2. **sound**, if its obfuscated texts are paraphrases of their originals, and
3. **sensible**, if its obfuscated texts are well-formed and inconspicuous.

These dimensions are orthogonal; an obfuscation software may meet each of them to a certain degree of perfection. Related work on operationalizing different measures for these dimensions has been included in our recent overview [20]. In particular, for lack of suitable alternatives, we developed our own evaluation measures for the safety dimension, which were employed to evaluate five author obfuscation approaches in the past. In this paper, we build on this experience and redesign our suite of safety measures from the ground up in an attempt to rectify issues with the existing ones. For example, the new measures incorporate the notion of "case difficulty" of author identification cases, the a priori quality of identification approaches, and they prevent some forms of cheating.

We directly employ the new performance measures to evaluate the safety of seven author obfuscation approaches against 44 author identification approaches. This includes two obfuscation approaches submitted to our this year's shared task on author obfuscation at PAN 2018, as well as five that have been submitted to the two corresponding shared tasks in the past two years [10, 20]. The 44 authorship identification approaches have been obtained from the shared tasks on authorship verification—a specific variant of author identification where a pair of texts is checked for common authorship—organized at PAN 2013–2015 [14, 23, 22]. As for the evaluation of sensibleness and soundness, we stick to manual inspection and grading of examples as before.

In what follows, Section 2 introduces the new safety performance measures for author obfuscation, Section 3 reviews the two obfuscation approaches submitted this year, and Section 4 evaluates their performance in comparison to the five previously submitted ones. More detailed analyses of the new performance measure is found in the appendix.

2 Towards a World Ranking for Author Obfuscators

We propose a formal model to implement a kind of "obfuscator world ranking" in order to ease the comparison of new approaches to the state of the art in this growing field. The central building block regarding the safety dimension is a set of effective authorship verification algorithms, also called authorship verifiers for short. Authorship verification is about deciding whether or not a document has been written by a certain author, given one or multiple texts that are known to be written by this author (a one-class classification problem). Then, given a set of authorship verifiers and a corpus of verification problems, a to-be-evaluated obfuscator is run on the *positive* problems (those problems where the correct answer is "same author"), and it is checked whether for the obfuscated texts the verifier decisions' are "different authors".

Thanks to the organizers and participants of the PAN 13, PAN 14, and PAN 15 shared tasks in authorship verification, 44 working authorship verifiers are at our disposal for empirical analysis. For each combination of a positive verification problem

Table 1. Performance matrix for a single obfuscator o given a number of positive verification problems $p_1, p_2, \ldots$ and a number of authorship verifiers $av_1, av_2, \ldots$ A table entry at position (i, j) encodes the following information: "T $\rightarrow$ _" (F $\rightarrow$ _) indicates for authorship verifier av_j a true positive (false negative) decision on the i-th positive verification problem p_i. Likewise, _$\rightarrow$ F indicates for the obfuscator in question that it could "force" the authorship verifier av_j to return a wrong decision (= different authors) on p_i.

Obfuscator o	av_1	av_2	av_3	$\ldots$
p_1	T $\rightarrow$ T	T $\rightarrow$ F	F $\rightarrow$ F	
p_2	F $\rightarrow$ T	F $\rightarrow$ F	T $\rightarrow$ F	$\vdots$
p_3	T $\rightarrow$ T	F $\rightarrow$ F	T $\rightarrow$ T	
$\ldots$		$\ldots$		

and a verifier, we can check how a verifier decides before and after applying the to-be-evaluated obfuscator, obtaining a performance matrix as given in Table 1. An entry of the form T $\rightarrow$ F indicates that the obfuscator o successfully fooled the authorship verifier associated to this column on the verification problem in the respective line.

Each successful entry in the performance matrix should increase the overall safety score of the respective obfuscator, while entries of the form F $\rightarrow$ T should decrease the score. The exact influence on the final score depends on the set of verifiers used and the "difficulty" of the problem instance (e.g., how many verifiers can identify the authorship before obfuscation).

Actually, Table 1 shows a simplified view of the real situation since an authorship verifier typically returns confidence scores instead of a plain binary decision. A confidence score in $[0, 0.5)$ indicates a (gradually) negative answer (= different authors) whereas a confidence score in $(0.5, 1]$ indicates a (gradually) positive answer. In practice, however, a sensible interpretation of the confidence scores requires a verifier-specific approach—or the computation of standard normally-distributed confidence scores. Here we will choose an individual confidence threshold for each verifier, optimizing its accuracy on the original instances from the verification task. The same threshold will then be used for the decisions on the obfuscated instances.

As for safety evaluation, in previous years we assumed that the used verifiers are *deterministic* in the sense that they always report the same answer for the same problem, regardless of the history of other problems they have seen. Note that the particular design of the testing scenario used at PAN (all test cases are provided at once) allows a verifier to consider all test documents when classifying an individual problem. E.g., a verifier could exploit *global* assumptions such as information about the ratio of positive versus negative problems in a setup. In such cases the entries of the performance matrices (see Table 1) may change every time the verifier is run on another sequence of problem instances.

It should be noted that the verifiers that participated in the PAN verification tasks are *obfuscation-unaware*, i.e., they assume that no obfuscation has been applied and thus take no measures to revert or reduce potential obfuscation effects. In the terminology of Potthast et al. [20], the verifiers perform *automated authorship identification*, as op-

posed to *de-obfuscation attacks*. Being obfuscation-unaware is not necessary for safety evaluation; however, in the long run, obfuscation technology should aim to defeat both automated authorship identification and de-obfuscation attempts.

In the following we present three axioms that should be fulfilled by a measure that quantifies the obfuscation safety of an obfuscation algorithm.

1. Fooling an effective verifier should be scored higher than fooling a less effective one, where effectiveness may be measured as the verifier's true positive rate.
2. Fooling a verifier on an unambiguous problem should be scored higher than fooling it on an ambiguous one. Here, a problem is unambiguous if it is decided correctly by many verifiers.
3. Fooling two dissimilar verifiers from a set of equally effective verifiers should be scored higher than fooling two similar verifiers from this set. Here, two verifiers are called similar if they often come to the same decision.

Axiom 1 relates to the verifier effectiveness, Axiom 2 relates to the problem unambiguity, and Axiom 3 relates to the verifier-problem coverage.

The first and third criterion together should help to prevent the creators of obfuscation algorithms from "boosting" their score by submitting many variations of a particular verifier which is especially vulnerable to their obfuscation approach (or, similarly, by submitting variants of approaches that would lower the scores of other obfuscators). The second criterion puts emphasis on those problems for which many of the state-of-the-art verification approaches perform well—in a nutshell: unambiguous implies "easy for attribution, difficult for obfuscation".

2.1 Definition of the New Safety Measure

For a formal treatment of authorship verification and obfuscation, we adopt the terminology outlined in [20]. Generally, an *authorship problem* is a tuple $\langle d_u, D_A \rangle$ consisting of one document of unknown authorship d_u and a set of documents D_A such that each document has a known author in the set A of authors. If A consists of a single author a, we call it *authorship verification problem*. We denote by $\gamma(\langle d_u, D_A \rangle) \in A \cup \{\varnothing\}$ the true author of d_u, if he/she is in A, and otherwise $\varnothing$.

Let $\mathcal{D}$ be the set of all authorship problems. An authorship analysis approach v is a computable function $v(\langle d_u, D_A \rangle) \in A \cup \{\varnothing\}$ which is an approximation for γ. v has been trained on a subset of $\mathcal{D}$, the training set, on which γ is known. v is evaluated using a test set $\mathcal{D}_{\text{test}} \subset \mathcal{D}$, which is disjoint from the training set. In the specific case of authorship verification, we let $\mathcal{D}^{+}_{\text{test}} \subset \mathcal{D}_{\text{test}}$ be the subset of verification problems $\langle d_u, D_a \rangle$ where $\gamma(\langle d_u, D_a \rangle) = a$.

An obfuscator o is a mapping $o(\langle d_u, D_a \rangle) = \langle \widetilde{d_u}, D_a \rangle$ which obfuscates the text d_u of the author a, possibly using the other available files for that author D_a. The aim of the obfuscation is to change the result of verifiers to $\varnothing$.

With the basic terminology at hand, we turn to the construction of the new measure: Let V be the set of verifiers under consideration. For $v \in V$ and $d \in \mathcal{D}_{\text{test}}$, let $c(v, d) = 1$ indicate that v outputs the *correct* answer to the problem d (whether it is

"same author" or "different authors") and let $c(v,d) = 0$ indicate a wrong answer. The *accuracy* and *effectiveness* of a verifier $v \in V$ on the problem set $\mathcal{D}_{\text{test}}$ are defined as follows:

$$\text{accuracy}(v, \mathcal{D}_{\text{test}}) = \frac{\sum_{d \in \mathcal{D}_{\text{test}}} c(v,d)}{|\mathcal{D}_{\text{test}}|},$$

$$\text{effectiveness}(v, \mathcal{D}_{\text{test}}) = \max(0, 2 \cdot \text{accuracy}(v, \mathcal{D}_{\text{test}}) - 1).$$

These definitions work best with a *balanced* problem set $\mathcal{D}_{\text{test}}$, where the set of positive problems $\mathcal{D}^+_{\text{test}}$ and that of negative problems $\mathcal{D}_{\text{test}} \setminus \mathcal{D}^+_{\text{test}}$ have approximately the same size. In a balanced situation, a verifier with effectiveness 0 does no better than guessing, whereas a verifier with effectiveness 1 is always correct.

To measure the *similarity* between two verifiers $v, w \in V$, we do not directly compare their answers but focus on the errors they make and assume that two verifiers are similar if they tend to make the same errors. We compute the Pearson correlation coefficient between the corresponding *error-vectors* $c(v,\cdot)$ and $c(w,\cdot)$:

$$\rho(v,w) = \frac{\sum_{d \in \mathcal{D}_{\text{test}}} (c(v,d) - \overline{v})(c(w,d) - \overline{w})}{\sqrt{\sum_{d \in \mathcal{D}_{\text{test}}} (c(v,d) - \overline{v})^2} \sqrt{\sum_{d \in \mathcal{D}_{\text{test}}} (c(w,d) - \overline{w})^2}},$$

where $\overline{v} = \text{accuracy}(v)$ and $\overline{w} = \text{accuracy}(w)$.

There are edge cases with vanishing denominators: In case that $c(v,d) = 0$ for all $d \in \mathcal{D}_{\text{test}}$ or $c(v,d) = 1$ for all $d \in \mathcal{D}_{\text{test}}$ (similarly for w), the above expression is undefined and we set $\rho(v,w) := 0$ if $v \neq w$ and $\rho(v,v) = 1$. Note that such verifiers have not been observed in practice yet—it would imply the existence of a perfect verifier for that particular set of problems. But in cases where $\lceil_{\text{test}}$ is small, it could easily happen—however, such small scenarios are not our aim.

For each verifier $v \in V$, we define its *coverage* as a real number in $(0,1]$ with the following intuition: If the verifier is unique of its kind (correlation with other verifiers near zero), the coverage should be 1. If there are $k > 0$ other verifiers which give answers almost equal to v, the coverage of v and its k related verifiers should be $\approx \frac{1}{k+1}$, such that these "redundant" verifiers together via their coverage scores will contribute as much to an obfuscator's performance as one verifier which is unique of its kind will. This motivates the following definition, where $\tau \in [0,1]$ is a fixed constant:

$$\text{coverage}(v) = \left(\sum_{\substack{w \in V: \\ \rho(v,w) \geq \tau}} \rho(v,w) \right)^{-1}.$$

Since $\rho(v,v) = 1$, the coverage is always in $(0,1]$. Choosing a larger value for τ diminishes the influence of many small correlations in contrast to a few bigger ones. We pick $\tau = 0.5$ to capture all real similarities while reducing the noise of correlations which are rather coincidences on a finite set of test instances. For each verifier $v \in V$, we then define its *importance* as

$$\text{importance}(v) = \text{effectiveness}(v) \cdot \text{coverage}(v).$$

We now quantify the unambiguity of a problem $d \in \mathcal{D}^+_{\text{test}}$ as a weighted average of the verifiers giving the correct answer to d:

$$\text{unambiguity}(d) = \frac{\sum_{v \in V} c(v,d) \cdot \text{importance}(v)}{\sum_{v \in V} \text{importance}(v)}.$$

With the above definitions, we can now define our measure of obfuscator safety. Thinking in terms of the performance matrix from Table 1, an entry of the form T $\rightarrow$ F (i.e., successful obfuscation), corresponding to a verifier $v \in V$ and a problem $d \in \mathcal{D}_{\text{test}}$, gives points equal to

$$\text{importance}(v) \cdot \text{unambiguity}(d).$$

For entries of the form F $\rightarrow$ T, the same amount describes the number of points subtracted for a counter-productive obfuscation attempt. The *world ranking score* of the obfuscator o is the sum of the points awarded or subtracted for each combination of verifier and verification problem.

To put it differently, recall that $o(d)$ denotes the obfuscated problem for $d \in \mathcal{D}^+_{\text{test}}$, where the texts of known authorship are unchanged but the text of unknown authorship is obfuscated. Denote by $c(v, o(d))$ the answer of the verifier $v \in V$ to the obfuscated problem (i.e., 1 if v correctly reports "same author" and 0 if not). Then the world ranking score of the obfuscator o equals

$$\sum_{d \in \mathcal{D}^+_{\text{test}}} \sum_{v \in V} (c(v,d) - c(v, o(d))) \cdot \text{importance}(d) \cdot \text{unambiguity}(d).$$

The measure is named *world ranking score* since it allows to incorporate all available verifiers and verification problem corpora to produce a single numerical value evaluating the obfuscator's safety with respect to the given verifiers and verification problems.

2.2 Theoretical Discussion of Fairness Properties

We give the following a-priori arguments why we hypothesize that the world ranking score satisfies the three fairness axioms.

To the first axiom: We decided to quantify effectiveness using accuracy. An ineffective verifier will only obtain a small effectiveness score, such that the influence of its decision changes on the final score are not as high as the influence of an effective verifier. Moreover, it is reasonable to assume that an effective and an ineffective verifier will have small correlation, such that adding ineffective verifiers will not change the importance scores of the effective verifiers nor the unambiguity scores of the problems. Therefore, an ineffective verifier has only little influence on the final world ranking score of an obfuscator.

An ambiguous problem should get a small unambiguity score, such that the overall influence of its obfuscated version on an obfuscator's world ranking score should be small. Adding ambiguous problems will, in general, reduce the effectiveness scores and increase the coverage scores, as most verifiers will effectively guess their answer.

However, this should uniformly affect all verifiers and all obfuscators, such that the world ranking scores with respect to a fixed corpus are comparable.

Adding a variant v' of an existing verifier $v \in V$ will not change the effectiveness score of any existing verifier and not change the coverage scores of verifiers $w \in V$ which are not similar to v. This will likely leave both the problem unambiguity scores and the overall obfuscator world ranking scores mostly unchanged, or at least affect them in a way that preserves the general proportions (i.e., the most unambiguous problems should remain the most unambiguous ones, even though their actual ambiguity scores may change).

2.3 Criticism and Shortcomings of the Impact Measure

Recall the definitions of *recall* and *accuracy* of a verifier $v \in V$:

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

$$\mathrm{rec}(v, \mathcal{D}_{\mathrm{test}}) = \mathrm{acc}(v, \mathcal{D}^+_{\mathrm{test}}) = \frac{|\{d \in \mathcal{D}^+_{\mathrm{test}} : v(d) = \gamma(d)\}|}{|\mathcal{D}^+_{\mathrm{test}}|}.$$

The performance of an obfuscator can therefore be measured by the *change of recall*:

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

The *impact* of o is Δ_{rec} normalized to $[-1, 1]$:

$$\mathrm{imp}(o, v, \mathcal{D}_{\mathrm{test}}) = \begin{cases} -\frac{\Delta_{\mathrm{rec}}(o,v,\mathcal{D}_{\mathrm{test}})}{\mathrm{rec}(v,\mathcal{D}_{\mathrm{test}})} & \text{if } \Delta_{\mathrm{rec}}(o, v, \mathcal{D}_{\mathrm{test}}) < 0, \\ -\frac{\Delta_{\mathrm{rec}}(o,v,\mathcal{D}_{\mathrm{test}})}{1-\mathrm{rec}(v,\mathcal{D}_{\mathrm{test}})} & \text{otherwise,} \end{cases}.$$

If V denotes the set of available verifiers, the *average impact* of an obfuscator o is

$$\mathrm{avg\,imp}(o, V, \mathcal{D}_{\mathrm{test}}) = \frac{1}{|V|} \sum_{v \in V} \mathrm{imp}(o, v, \mathcal{D}_{\mathrm{test}}).$$

The average impact is the measure used for the obfuscation shared tasks in PAN 16 and PAN 17 to automatically evaluate an obfuscator's safety. This measure does not satisfy the earlier defined fairness criteria: We will later see that the performance of the verifiers used here differ enormously. For an obfuscator o and a verifier v such that obfuscation by o decreases the recall of v (which is usually the case), the impact is, by definition,

$$\mathrm{imp}(o, v, \mathcal{D}_{\mathrm{test}}) = \frac{\mathrm{rec}(v, \mathcal{D}_{\mathrm{test}}) - \mathrm{rec}(v, o(\mathcal{D}_{\mathrm{test}}))}{\mathrm{rec}(v, \mathcal{D}_{\mathrm{test}})} = 1 - \frac{\mathrm{rec}(v, o(\mathcal{D}_{\mathrm{test}}))}{\mathrm{rec}(v, \mathcal{D}_{\mathrm{test}})}$$

$$= 1 - \frac{\left|\{d \in \mathcal{D}^+_{\mathrm{test}} \mid v(o(d)) = \gamma(d)\}\right|}{\left|\{d \in \mathcal{D}^+_{\mathrm{test}} \mid v(d) = \gamma(d)\}\right|}.$$

If, e.g., o flips half of the correct decisions of v and leaves the wrong decisions as they were, the impact factor is $\frac{1}{2}$. Achieving an impact factor of $\geq \frac{1}{2}$ is therefore easier for weak verifiers (change few decisions of an ineffective verifier) than for more effective ones (change many decisions of an effective verifier). This is counter-intuitive and directly contradicts our first fairness principle above (the verifier effectiveness axiom). Of course, one has to be careful when using only recall (and not precision) to describe "effective verifiers", but a high recall is not a good indicator for ineffective verifiers.

Moreover, there are examples of similar verifiers: Jankowska et al. submitted a verifier in 2013 [12], and an improved variation of it in 2014 [13]. These are treated as independent verifiers, such that the influence of that single approach in the final impact score is inappropriately high. This is not so much of a problem for the set of verifiers available at the time of writing (there are few such examples), but opens the door for simple manipulation of certain obfuscators' scores by re-submitting a verifier multiple times, possibly in slight variations. However, submission of variations of already present verifiers need not be an attempt of score manipulation, it could simply be an improvement of previous work (e.g. incorporating more features, or a different machine learning algorithm). Therefore it is not fair to disallow submissions of such variants (since we want to reflect the state of the art), nor can it be fair to consider them as entirely independent of one or more related verifiers when averaging the scores. When using the average impact measure, however, one has to decide for one of these options.

Finally, we will present evidence that there are very ambiguous as well as very unambiguous problems in the test corpora, an important distinction not reflected by the average impact measure. The idea is that fooling a certain verifier in an unambiguous problem is supposed to be more difficult than fooling it in an ambiguous one, so that success in the more difficult task should get a better reward than success in the easier one. However, one could make the non-trivial, yet reasonable assumption that each obfuscator which successfully (against a particular verifier) obfuscates problems up to some degree of unambiguity will also successfully (against the same verifier) obfuscate more ambiguous problems. Under this assumption, it is not necessary unfair to simply count the number of flipped decisions independent of each problem's ambiguity (though it also would not be unfair to take the ambiguity into account). This assumption can, however, be questioned, e.g. by pointing out that there are random effects involved such that an obfuscator may by chance fail to fool a verifier in some ambiguous problems although the obfuscation is successful in some more unambiguous cases. It is therefore desirable to have a measure whose fairness can be justified without relying on this or a similar assumption.

3 Survey of Submitted Obfuscation Approaches

The two approaches submitted to this year's edition of our shared task are of a more rule-based flavor, but with different aggressiveness. The rather conservative rule-based replacements of Kocher and Savoy's approach [16] aim for sensible and sound obfuscations, while the more aggressive strategy of Rahgouy et al.'s approach [21] was inspired by Mihaylova et al.'s approach [18].

Table 2. Characteristics of the used corpora and the number of verifiers we were able to run on a corpus.

Corpus	Problem instances			Verifiers
	Same author	Different author	Total	
PAN13	14	16	30	41
PAN14 EE	100	100	200	38
PAN14 EN	100	100	200	39
PAN15	250	250	500	40

Kocher and Savoy The approach of Kocher and Savoy [16] is based on a set of 20 rules. The rules replace contractions (e.g., 'll → will) or shorten words to contracted forms, replace adjectives (e.g., very good→excellent), exchange conjunctive terms (e.g., in fact→ actually), or introduce spelling errors by repeating a repeated letter.

As for the resulting texts, the rather conservative strategy does not change a lot and aims to keep the original text quality. A rather problematic issue of the approach is that it seems as if the sentences in the obfuscated text were re-ordered (the whole second paragraph is often directly inserted after the first sentence while the second sentence comes rather late). It is unclear how reordering a text changes its style, so that this issue might be due to some implementation error, and is the main cause for the rather sub-optimal scores for soundness. A more detailed analysis of the text quality of this approach follows in the evaluation (cf. Section 4).

Rahgouy et al. The approach of Rahgouy et al. [21] mainly focuses on changing the signals of sentence length (splitting at conjunctive terms or combining sentences), usage of contractions (either contracting two words or extending contractions), and word usage in general (replacement candidates from WordNet or a word2vec model), where a potential replacement term is scored based on the similarity to the original word, the likelihood of seeing it in the author's documents, and the word mover distance of the sentence variant.

The resulting text will usually be changed more than with Kocher and Savoy's approach. Depending on the word similarities and the "correctness" of splitting / combining sentences, some obfuscated text passages read very well while others are changed to often even a contrary meaning. Again, the text quality of this approach is discussed in Section 4.

4 Evaluation

As in the past two years, we automatically evaluate the safety of the submitted obfuscation approaches against 44 authorship verifiers that participated in the previous three shared tasks on authorship identification at PAN 2013–2015. We employ the new scoring scheme to obtain the first "world ranking" for obfuscation and contrast it with the previous way of evaluating safety. Sensibleness and soundness of the obfuscated texts are assessed manually via human inspection in the same way as before.

Table 3. World ranking of seven obfuscators, including those submitted to PAN 2016 and PAN 2017, against sets of 44 authorship verification approaches submitted to PAN 2013 through PAN 2015. In addition, we provide the scores achieved with our legacy measures.

Obfuscator		**Verifier**	**Dataset**		**Legacy Measures**			**World Ranking**
Team	[Reference]	$\|Y\|$	$\mathcal{D}_{\text{test}}$	$\|\mathcal{D}^{+}_{\text{test}}\|$	Δ_{acc}	Δ_{rec}	avg imp	Score
Castro et al.	[6]	44	all corpora	464	-0.1281	-0.2387	0.4495	**474.24**
Mihaylova et al.	[18]	44	all corpora	464	-0.1104	-0.2099	0.3901	466.18
Keswani et al.	[15]	44	all corpora	464	-0.1071	-0.1990	0.3736	296.64
Rahgouy et al.	[21]	44	all corpora	464	-0.0940	-0.1771	0.3531	355.18
Bakhteev et al.	[1]	44	all corpora	464	-0.0726	-0.1322	0.2491	291.01
Mansoorizadeh et al.	[17]	44	all corpora	464	-0.0378	-0.0738	0.1523	208.62
Kocher and Savoy	[16]	44	all corpora	464	-0.0549	-0.1003	0.2180	107.85

4.1 Safety

The evaluation setup is based on the cloud-based evaluation platform TIRA [9, 19],[1] which is being developed as part of our long-term evaluation-as-a-service initiative [11]. By using TIRA, it is possible to run 44 of the 49 authorship verification approaches which have been submitted to the shared tasks at PAN 2013–2015 on the outputs of the 7 obfuscation approaches submitted to the shared tasks at PAN 2016–2018 using the authorship verification corpora PAN13, PAN14 EE, PAN14 EN, and PAN15 (cf. Table 2 for respective corpus characteristics).

Table 3 contrasts our new "world ranking" scheme described in Section 2 with the "legacy" safety measure we employed in the previous two years. It can be seen that the ranking of the approaches does not change from one measure to the other with Castro et al.'s approach [6] being the safest on both measures. Although the ranking of the so far seven submitted approaches does not change, the new measure seems a lot fairer since an obfuscator cannot simply gain a lot of average impact by fooling a few rather bad verifiers to each change one of their maybe only two correct "same author" decisions (impact of 0.5). Moreover, the point difference between of the world ranking provides for an intuitive idea of "how far" one obfuscator is ahead of another. Thus we will further employ the new "point-based" world ranking for safety evaluation since it takes the strengths (and weaknesses) of the participating verifiers and corpora into account.

Interestingly, the two obfuscator approaches from this year's shared task (green entries in Table 3) are not able to break into the top-3 safety ranks overall. But their characteristic is also a less aggressive obfuscation with sometimes no changes in a sentence, so that the resulting text quality is not changed too much from the original (cf. next section for the respective evaluation of sensibleness and soundness). It should also be noted that even the "safest" of the approaches submitted so far still leaves a lot of room for safety improvements alone. The perfect obfuscator (i.e., the one that would be able to change every correct "same author" decision of each of the 44 verifiers in the evaluation) would obtain a world ranking score of 1093.49—thus Castro et al.'s best approach so far has not even made "half the way" to perfect obfuscation. A more

[1] www.tira.io

fine-grained analysis of the old and new safety scores per corpus can be found in Appendix B while Appendix A gives some more experimental details on the robustness of the new world ranking.

4.2 Sensibleness and Soundness

As in the previous two shared task, a human assessor conducted a manual assessment on problem instances 6, 135, and 430. Spot checks on other instances indicated that the overall characteristics of the output texts are similar on other instances. 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. The sensibleness scores obtained in the previous years were a grade 2 for Mansoorizadeh et al.'s approach [17], grade 2-3 for Castro et al.'s approach [6], grade 4 for Mihaylova et al.'s [18] and Bakhteev's and Khazov's [1] obfuscators, and a grade 5 for Keswani et al.'s obfuscator [15]. Kocher and Savoy's approach [16] approach obtains a grade 1-2, since there are hardly any changes, though spurious uppercase letters occurred in the middle of a sentence (probably due to some suboptimal "stitching" of text passages). Rahgouy et al.'s approach [21] had a much wider range with one text rather left intact and obtaining grade 1, while for another a grade 4 was assigned due to a lot of punctuation problems; on average, with the additional grade 3 on the third text, grade 3 overall is assigned.

After grading the sensibleness of the obfuscated texts, the assessor read the original texts and judged the textual differences in various ways to evaluate the soundness of the obfuscated texts on a three-point scale as either "correct", "passable", or "incorrect". The obfuscated texts of Mihaylova et al.'s, Keswani et al.'s, Bakhteev's and Khazov's, and Castro et al.'s previous years' approaches were all judged "incorrect", while Mansoorizadeh et al.'s very conservative approach from 2016 achieved "correct" and "passable" scores. This year's approaches (Kocher and Savoy's and Rahgouy et al.'s) both got "passable" as their average judgments—but for different reasons: With regard to Kocher and Savoy's approach, almost everything was left as it was with the main problem that the ordering of the sentences was changed which in some passages caused a rather odd reading "flow" resulting in a "'passable" for all three checked texts. With regard to Rahgouy et al.'s approach, the judgments are again more wide-spread with one text being "incorrect" since almost all sentences were wrongly split into parts, one document being "passable" and one being "correct" since hardly anything was changed.

5 Conclusion and Outlook

In the third year of evaluating author obfuscation approaches in terms of their safety against the state of the art in authorship verification, two new approaches were added to the five approaches from the previous years. The best-performing obfuscator today achieves about 43% of the score a "perfect" obfuscator can achieve on our testbed with 44 authorship verifiers on the PAN authorship verification corpora from 2013–2015.

Such a perfect obfuscator would be able to flip any correct decision for "same author" by any verifier towards choosing "different author".

We have developed a new "world ranking of obfuscators" in terms of safety. The idea is that an obfuscator obtains points per flipped correct "same author" decision, dependent on how unambiguous a problem is (i.e., the more unambiguous the more verifiers solve it correctly before obfuscation) and on the effectiveness of a verifier on unobfuscated problems (more points if a more effective verifier is fooled). This new safety measure is fairer than the previous impact measure, where scores were computed independent of a problem's unambiguity and a verifier's effectiveness.

Still, even with the new scoring scheme, the actual safety ranking does not change but the relative differences between the verifiers are more pronounced in the sense that they now include information about the actual problem unambiguity. The safety-wise best-performing approach of Castro et al. from 2017 was not beaten by the two new obfuscators, whose main focus seems to rather be text quality (i.e., soundness and sensibleness). As in the previous years, text quality was measured by manual inspection. It became clear that sometimes even small changes can "destroy" a particular sentence or text passage (distorting its meaning or decreasing readability). Unsurprisingly, the least safe approach of Kocher and Savoy from this year's shared task obtains the best scores among all approaches with respect to soundness and sensibleness—it simply does not change a lot. While still being readable, the obfuscated texts of the safest approach (Castro et al.) are rather poor when it comes to soundness.

It is still an open problem to develop obfuscation technology that is safe (even the best one is not half the way to a perfect obfuscation safety) while not harming paraphrase soundness or readability too much. Paradigmatically, there are still more or less only two groups of obfuscation approaches: (1) the ones that are somewhat safe but that produce rather unreadable text or text that is neither sound nor sensible, and (2) the ones that produce sound and sensible texts but that are not really safe against authorship verification.

As hinted in the previous shared task editions, a significant improvement of current obfuscation technology might require a much better consideration and integration of the surrounding context when, for example, replacing, adding, or removing words, and better ways of reordering clauses in sentences. Ideas in that direction could be to apply constrained paraphrasing [24] or paraphrasing rules from the PPDB [8].

The remaining challenge of evaluating author obfuscation approaches properly and at scale does not seem to be safety. The new "world ranking" provides for a fair and robust tool that incorporates future verifiers, obfuscators, and corpora. What is missing are new and improved technologies for recognizing paraphrases, textual entailment, grammaticality, and style deception—the existing technology is not mature enough to easily replace manual inspection for evaluating soundness and sensibleness.

Acknowledgments

We thank the participating teams of the three editions of this shared task.

Bibliography

[1] Bakhteev, O., Khazov, A.: Author Masking using Sequence-to-Sequence Models—Notebook for PAN at CLEF 2017. In: [3], http://ceur-ws.org/Vol-1866/

[2] 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

[3] Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.): CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Dublin, Ireland. CEUR Workshop Proceedings, CEUR-WS.org (2017), http://www.clef-initiative.eu/publication/working-notes

[4] 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

[5] Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.): CLEF 2017 Evaluation Labs and Workshop – Working Notes Papers, 11-14 September, Avignon, France. CEUR Workshop Proceedings, CEUR-WS.org (2018), http://www.clef-initiative.eu/publication/working-notes

[6] Castro, D., Ortega, R., Muñoz, R.: Author Masking by Sentence Transformation—Notebook for PAN at CLEF 2017. In: [3], http://ceur-ws.org/Vol-1866/

[7] 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

[8] 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)

[9] 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)

[10] Hagen, M., Potthast, M., Stein, B.: Overview of the Author Obfuscation Task at PAN 2017: Safety Evaluation Revisited. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, vol. 1866. CLEF and CEUR-WS.org (Sep 2017), http://ceur-ws.org/Vol-1866/

[11] 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

[12] 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: [7]

[13] Jankowska, M., Kešelj, V., Milios, E.: Ensembles of Proximity-Based One-Class Classifiers for Author Verification—Notebook for PAN at CLEF 2014. In: [4]

[14] Juola, P., Stamatatos, E.: Overview of the Author Identification Task at PAN 2013. In: [7]

[15] Keswani, Y., Trivedi, H., Mehta, P., Majumder, P.: Author Masking through Translation—Notebook for PAN at CLEF 2016. In: [2], http://ceur-ws.org/Vol-1609/

[16] Kocher, M., Savoy, J.: UniNE at CLEF 2018: Author Masking—Notebook for PAN at CLEF 2018. In: [5]
[17] Mansoorizadeh, M., Rahgooy, T., Aminiyan, M., Eskandari, M.: Author Obfuscation using WordNet and Language Models—Notebook for PAN at CLEF 2016. In: [2], http://ceur-ws.org/Vol-1609/
[18] Mihaylova, T., Karadjov, G., Nakov, P., Kiprov, Y., Georgiev, G., Koychev, I.: SU@PAN'2016: Author Obfuscation—Notebook for PAN at CLEF 2016. In: [2], http://ceur-ws.org/Vol-1609/
[19] 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)
[20] Potthast, M., Hagen, M., Stein, B.: Author Obfuscation: Attacking the State of the Art in Authorship Verification. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, vol. 1609. CLEF and CEUR-WS.org (Sep 2016), http://ceur-ws.org/Vol-1609/
[21] Rahgouy, M., Babaei Giglou, H., Rahgooy, T., Zeynali, H., Mirza Rasouli, S.: Author Masking Directed by Author's Style—Notebook for PAN at CLEF 2018. In: [5]
[22] 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: 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 (Sep 2015)
[23] 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: [4]
[24] 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)

A Robustness of the World Ranking

We empirically test whether the parameters underlying the new world ranking are distributed as expected and how robust the new world ranking is against the addition of "random" verifiers or verifiers that are very similar to already existing ones.

A.1 Variation of the Measures Underlying the World Ranking

First, to ensure that the considerations underlying our new safety evaluation approach make sense in the given setup, we confirm that the main verifier and problem characteristics of effectiveness, coverage, and unambiguity vary sufficiently among the considered verifiers and datasets. We confirmed our findings on all PAN verification datasets but only give plots and explanations on the basis of the PAN 15 data since the observations on the other datasets are similar.

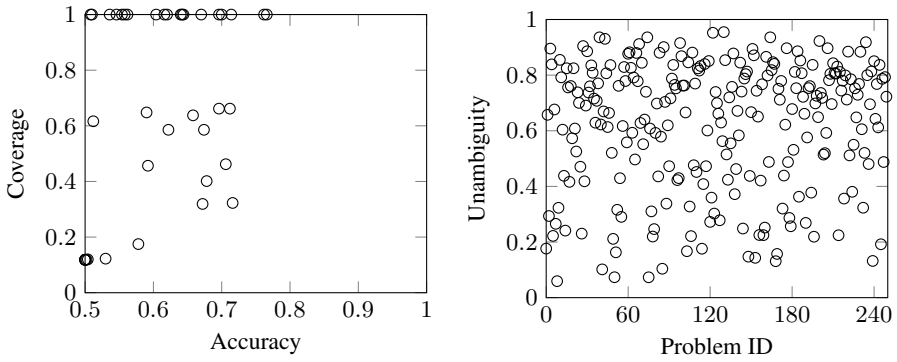

Figure 1. Left: Distribution of accuracy and coverage for the 40 verifiers that we were able to run on the PAN 15 corpus. Right: Distribution of the unambiguity scores for the PAN 15 corpus (by definition, only of the positive problems). The x-axis gives the internal IDs of the problems.

Figure 1 (left) shows the accuracy and coverage scores of the 40 verifiers that we were able to run on the original PAN 15 data. Many of the verifiers seem to follow rather unique approaches such that 19 of the 40 verifiers have a coverage score of 1. Note that some verifiers perform so poorly on the test data that our "best threshold" method chooses a threshold of $+\infty$ or $-\infty$, forcing them to always respond "different authors" or "same author", respectively. This behavior is mainly due to employing verification models that were trained on a different PAN corpus that are now tested on PAN 15 data (we did not re-train models that were submitted in another year), leading to multiple verifiers giving identical answers, which in turn then have very low coverage scores (between 0.11 and 0.123, with accuracies between 0.5 and 0.55). The two implementations of Jankowska et al. [12, 13] also seem to be similar to each other and have a coverage score of 0.59.

Figure 1 (right) exemplify the unambiguity scores for the PAN 15 test data. Following the definition of unambiguity, a clear problem (i.e., easy to identify the author) has a high unambiguity score whereas an obscure one has a low score. The observation that clear and obscure problems are rather evenly spread does not only hold for PAN 15 but also for the other datasets.

Our inspection of the different PAN authorship verification corpora shows that the effectiveness, similarity, and unambiguity scores vary in the expected scope such that the definitions reasonably capture differences.

A.2 Influence of Random or Copied Verifiers on the World Ranking

Since the PAN shared tasks are intended to continue even after the respective conferences, everyone can submit new approaches for authorship verification at any time. The measures underlying the "world ranking" are directly able to incorporate any such new approaches and thus are able to always include the state of the art in authorship verification. Also creators of obfuscation approaches, of course, may submit new verifiers. This

possibility could in principle also be exploited to submit verifiers with the intention of boosting a specific obfuscator's world ranking score:

- One could let the obfuscator leave a certain "watermark" in the text that a newly submitted verifier could then detect to turn all its decisions on such watermarked cases to "different authors".
- Somewhat similarly, one could develop a verifier just focused on those aspects of the texts that the desired obfuscator will manipulate (e.g., counting the occurrences of "it's" vs. "it is"). Such a verifier might work with some success on unobfuscated texts but then not at all on the obfuscated texts of the specific obfuscator since the obfuscator probably will have removed most of the features used by the verifier.
- One could also submit a slight variation of an existing verifier which is particularly vulnerable to the to-be-boosted obfuscation approach by altering the answers only in low-confidence cases. The verifier variant then should easily be fooled by the obfuscator while retaining reasonable performance on unobfuscated texts.

Such and similar deliberate attempts to boost an obfuscator's score should be considered "unfair play" and unscientific. In case of being detected—probably rather difficult to do automatically—, such verifiers should probably be removed entirely instead of just being ignored in the world ranking.

The second option could still yield a reasonable and well-performing approach in authorship verification, in which case it is a valid contribution. If the verifier, however, does not perform better than guessing, its effectiveness score should become too low to make any difference. If the verifier significantly resembles an existing one (as in the third option), the corresponding coverage scores would be lowered accordingly such that the overall score of all obfuscators ultimately should remain stable.

These assertions can be tested (effectiveness and coverage for some potentially adversarial submissions). We perform two experiments, in which we add 40 mock verifiers to the existing ones and look how the characteristics of the verifiers and the problems behave, and whether the overall scores for the different obfuscators and test corpora change. In the first experiment, the *guessing attack*, the mock verifiers just choose their confidence scores randomly in $[0, 1]$ (uniformly distributed). We expect that those mock verifiers get poor effectiveness and high coverage scores, whereas the characteristics of the original verifiers and the problems remain stable, as well as the obfuscators' scores. In the second experiment, the *variation attack*, each mock verifier is obtained from an existing verifier, replacing 10% of its decisions (on the original and the obfuscated problems) by random confidence scores in $[0, 1]$. We expect that the effectiveness scores of the new verifiers are a bit lower than those of their originals, that these mock verifiers get low coverage scores in general, and that the coverage scores of the existing verifiers decrease overall, in particular for those which have been copied. Note that in both experiments, the effectiveness scores of the original verifiers remain stable by definition.

The left column of Figure A.2 (Experiment 1: Guessing Attack) shows how the characteristics change after attacking the ranking by adding 40 mock verifiers which only guess their responses. The effectiveness scores of the mock verifiers are very low—as expected since they are just guessing. Note that the effectiveness scores of the original verifiers remain stable (by definition). The coverage scores of the badly performing verifiers, which, after choosing optimal thresholds, give constant answers, shrink a bit

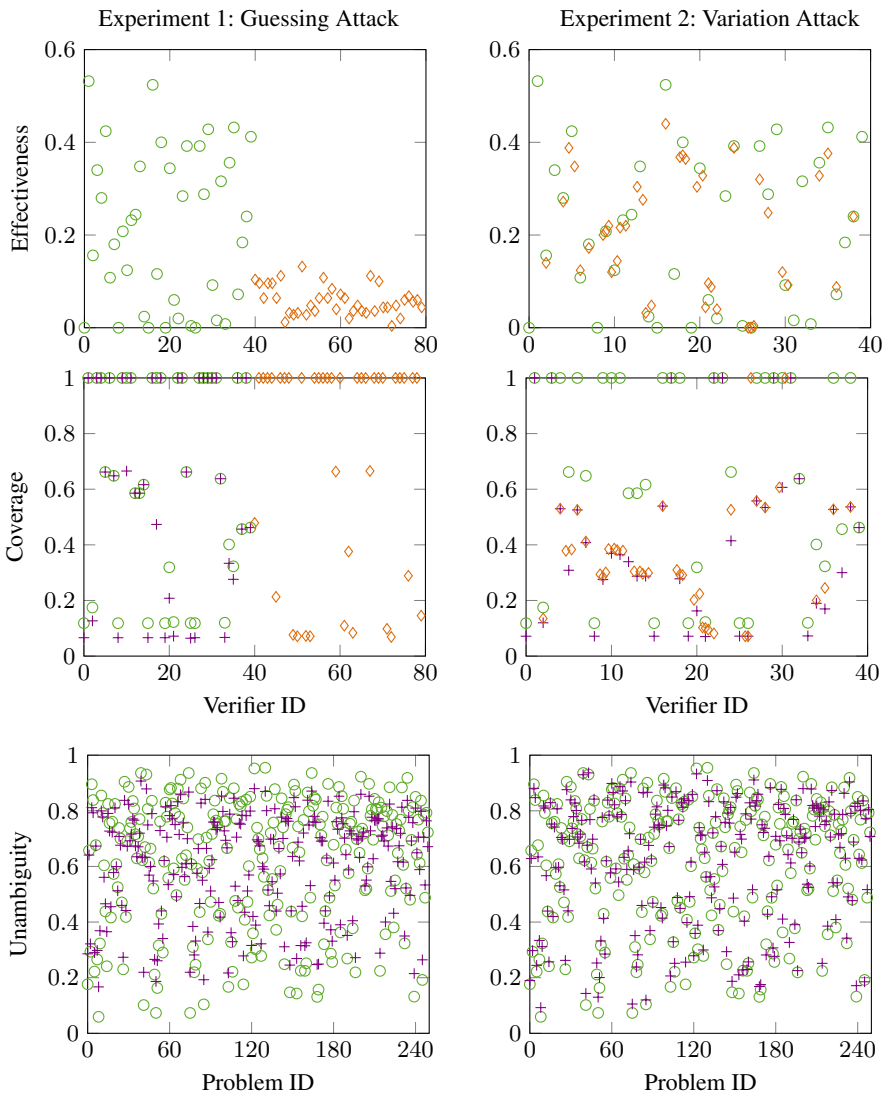

Figure 2. Results of two robustness experiments with regard to effectiveness, coverage, and unambiguity. The x-axes of the plots show IDs of verifiers and problems, respectively. Circles (○) denote scores of the original verifiers or problems, crosses (+) denote scores of original verifiers or problems after adding 40 mock verifiers, and diamonds (◇) denote scores of the 40 mock verifiers. Since effectiveness is unaffected by adding verifiers and since no mock problems have been added, there are no crosses in the first row of plots, and no diamonds in the last one. Dependent on the experiments, the 40 mock verifiers guess at random (verifier IDs 40-79 in the left plots) or are variants of the original ones (IDs 0-39 in the right plots shared with their respective original verifiers). The problem IDs refer to the subset of positive problems of the PAN 15 dataset.

since also some of the mock verifiers also perform very badly. The other coverage scores remain mostly stable—as expected. The problem unambiguity scores move towards 0.5 since about half of the guessing verifiers guess the correct answer—if every verifier guessed its answers on an instance, half of them will guess correctly, resulting in a unambiguity score of exactly 0.5 for that instance.

The right column of Figure A.2 (Experiment 2: Variation Attack) shows the behavior of the three characteristics after attacking the ranking by copying (with a 10% random change) the 40 original verifiers. Similar to the first experiment, the observations are close to our expectations. The mock verifiers have roughly the same effectiveness as their originals, sometimes diminished due to the 10% random choices. Those verifiers which have not been copied mostly have the same coverage scores (except some few side effects), whereas those which have been copied have significantly reduced coverage scores. Except some few random outliers, these reduced coverage scores are also observed for the copies of the verifiers (often slightly higher which may be explained by the "unpredictable" 10% random choices). Similar to the first experiment, the random 10% choices explain the slight drift of the unambiguity scores towards 0.5.

The obfuscators' scores in the two experimental setups from above (random mock verifiers and verifier variants) are given in Table 4. It can be seen that adding the mock verifiers does change the obfuscators' scores, though most changes are small compared to the actual differences between the obfuscators under consideration. In particular, the ordering of the obfuscators for each dataset remains stable, with the two exceptions that Keswani et al. and Castro et al. change their ranks for PAN13 and that Mansoorizadeh et al. and Castro et al. change their ranks for PAN14-EE for the setup with the 40 random mock verifiers. In both experiments, some scores are increased whereas some are reduced; however, a general trend towards higher scores is observable in both experiments. This trend may be explained statistically, noting that those random verifiers with an accuracy $\gg 0.5$ on the original training data will still have an expected accuracy of 0.5 on the obfuscated data, thus changing probably some correct positive decisions to negative ones while having a moderate effectiveness score. Those random verifiers which perform poorly on the original data (i.e., that have an accuracy of ≈ 0.5), are more likely to correct wrong decisions by chance, but get a smaller effectiveness score such that their final influence is not as high.

Other than that, the changes in the obfuscators' scores are probably best explained as just random effects, since adding 40 random mock verifiers to ≈ 40 real verifiers will induce them. The remarkable thing here is that our proposed measure retains relatively stable output under such heavy modifications of the input data.

B Safety Evaluation According to the Legacy Evaluation Measures

Table 5 shows the results of our safety evaluation of the two obfuscators from this year compared to the five obfuscators from the last two years against 44 authorship verification approaches on the aforementioned four PAN evaluation datasets. Although we combined the rankings into an overall score to arrive a our ranking of obfuscation approaches, a per-dataset inspection allows for a more in-depth interpretation of the results of this year's participants in context.

Table 4. World ranking of seven obfuscators, including those submitted to PAN 2016 and PAN 2017, against sets of 34–39 authorship verification approaches submitted to PAN 2013 through PAN 2015. We note the scores for the original data and for the two experiments, each with additional 40 mock verifiers.

Obfuscator		**Verifier**	**Dataset**		**Experiments**		**World Ranking**
Team	[Reference]	$\|Y\|$	$\mathcal{D}_{\text{test}}$	$\|\mathcal{D}^{+}_{\text{test}}\|$	Guessing	Variation	Score
Castro et al.	[6]	36	PAN13	14	28.92	21.56	**15.96**
Mihaylova et al.	[18]	36	PAN13	14	28.66	22.58	15.70
Keswani et al.	[15]	36	PAN13	14	20.97	18.84	13.84
Rahgouy et al.	[21]	36	PAN13	14	22.25	15.71	9.63
Bakhteev et al.	[1]	36	PAN13	14	21.00	15.19	8.74
Kocher and Savoy	[16]	36	PAN13	14	12.72	5.86	4.43
Mansoorizadeh et al.	[17]	36	PAN13	14	13.92	9.10	3.83
Castro et al.	[6]	34	PAN14 EE	100	63.71	69.59	**61.80**
Mihaylova et al.	[18]	34	PAN14 EE	100	57.02	60.34	54.47
Keswani et al.	[15]	34	PAN14 EE	100	49.90	48.87	46.40
Rahgouy et al.	[21]	34	PAN14 EE	100	41.47	39.90	35.98
Bakhteev et al.	[1]	34	PAN14 EE	100	30.77	31.65	28.62
Mansoorizadeh et al.	[17]	34	PAN14 EE	100	27.51	25.82	23.00
Kocher and Savoy	[16]	34	PAN14 EE	100	22.20	18.44	17.44
Mihaylova et al.	[18]	37	PAN14 EN	100	99.83	119.93	**102.29**
Castro et al.	[6]	37	PAN14 EN	100	73.50	90.83	72.90
Keswani et al.	[15]	37	PAN14 EN	100	56.83	58.51	56.20
Rahgouy et al.	[21]	37	PAN14 EN	100	55.89	60.97	52.47
Bakhteev et al.	[1]	37	PAN14 EN	100	40.35	47.23	44.32
Mansoorizadeh et al.	[17]	37	PAN14 EN	100	33.35	38.43	36.10
Kocher and Savoy	[16]	37	PAN14 EN	100	25.67	19.31	17.41
Castro et al.	[6]	39	PAN15	250	318.07	335.58	**323.58**
Mihaylova et al.	[18]	39	PAN15	250	295.14	305.07	293.72
Rahgouy et al.	[21]	39	PAN15	250	253.17	271.10	257.09
Bakhteev et al.	[1]	39	PAN15	250	208.16	218.35	209.32
Keswani et al.	[15]	39	PAN15	250	177.68	194.26	180.04
Mansoorizadeh et al.	[17]	39	PAN15	250	149.44	151.66	145.70
Kocher and Savoy	[16]	39	PAN15	250	74.85	72.94	68.58

The best-performing approach this year was submitted by Rahgouy et al., which achieves 4th rank overall across the three years as per average impact; the average impact quantifies the averaged ratio of true positive decisions turned false negative. However, this result must be taken with a grain of salt since this approach basically removed large parts of the original text. The approach of Bakhteev and Khazov [1] performs second-best this year, and ranks fourth out of five overall. The ranking induced by average impact is inconsistent with those induced by AUC difference or C@1 difference on some datasets. The penultimate approach of Kocher and Savoy [16] achieves best performance for these measure on the PAN 13 and the PAN 14 EN datasets. We hypothesize that is is due to the ability of this approach to effectively lower confidence values without actually bringing them below the threshold. Nevertheless, the approach of Mihaylova et al. [18] still performs best in most situations.

Table 5. Safety evaluation of seven obfuscators, including those submitted to PAN 2016 and PAN 2017, against sets of 26-36 authorship verification approaches submitted to PAN 2013 through PAN 2015. The 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. The four row groups belong to the four English PAN test datasets; the rows within the row groups are ordered by average impact (avg imp, see the last column).

Obfuscator		**Verifier**	**Dataset**		**PAN Measures**			**Legacy Measures**		
Team	[Reference]	$\|Y\|$	$\mathcal{D}_{\text{test}}$	$\|\mathcal{D}^{+}_{\text{test}}\|$	Δ_{AUC}	$\Delta_{\text{C@1}}$	Δ_{final}	Δ_{acc}	Δ_{rec}	avg imp
Mihaylova et al.	[18]	36	PAN13	14	-0.1066	-0.0759	**-0.1030**	**-0.1389**	**-0.2778**	**0.4690**
Keswani et al.	[15]	36	PAN13	14	-0.0908	-0.0695	-0.0940	-0.1148	-0.2361	0.4245
Castro et al.	[6]	36	PAN13	14	-0.1106	-0.0545	-0.0920	-0.1248	-0.2449	0.4175
Rahgouy et al.	[21]	36	PAN13	14	-0.1116	-0.0640	-0.0800	-0.1088	-0.2248	0.3952
Bakhteev et al.	[1]	36	PAN13	14	-0.0518	-0.0547	-0.0631	-0.0796	-0.1667	0.2881
Kocher and Savoy	[16]	36	PAN13	14	**-0.1353**	**-0.1149**	-0.0912	-0.0452	-0.1020	0.1910
Mansoorizadeh et al.	[17]	36	PAN13	14	-0.0422	-0.0254	-0.0392	-0.0463	-0.0933	0.1442
Mihaylova et al.	[18]	26	PAN14 EE	100	**-0.1305**	**-0.1088**	**-0.1144**	**-0.1229**	**-0.2304**	**0.4891**
Castro et al.	[6]	26	PAN14 EE	100	-0.1287	-0.1093	-0.1142	-0.1217	-0.2273	0.4328
Keswani et al.	[15]	26	PAN14 EE	100	-0.1085	-0.0870	-0.0960	-0.0975	-0.1873	0.4058
Rahgouy et al.	[21]	26	PAN14 EE	100	-0.1221	-0.1090	-0.1002	-0.0954	-0.1756	0.3859
Bakhteev et al.	[1]	26	PAN14 EE	100	-0.0518	-0.0453	-0.0509	-0.0631	-0.1177	0.2558
Mansoorizadeh et al.	[17]	26	PAN14 EE	100	-0.0514	-0.0463	-0.0473	-0.0577	-0.1038	0.2512
Kocher and Savoy	[16]	26	PAN14 EE	100	-0.1152	-0.1071	-0.0749	-0.0457	-0.0783	0.1646
Mihaylova et al.	[18]	36	PAN14 EN	100	**-0.1613**	-0.1050	**-0.1260**	**-0.1456**	**-0.2456**	**0.4750**
Rahgouy et al.	[21]	36	PAN14 EN	100	-0.1269	-0.0967	-0.0988	-0.1032	-0.1788	0.3945
Castro et al.	[6]	36	PAN14 EN	100	-0.1335	-0.0793	-0.1014	-0.1149	-0.1900	0.3811
Keswani et al.	[15]	36	PAN14 EN	100	-0.1020	-0.0704	-0.0845	-0.1074	-0.1783	0.3769
Bakhteev et al.	[1]	36	PAN14 EN	100	-0.0700	-0.0475	-0.0599	-0.0776	-0.1129	0.2354
Mansoorizadeh et al.	[17]	36	PAN14 EN	100	-0.0579	-0.0408	-0.0493	-0.0665	-0.0958	0.2345
Kocher and Savoy	[16]	36	PAN14 EN	100	-0.1509	**-0.1320**	-0.0871	-0.0289	-0.0561	0.1867
Mihaylova et al.	[18]	35	PAN15	250	**-0.1074**	**-0.0927**	**-0.1090**	**-0.1050**	**-0.2009**	**0.3649**
Castro et al.	[6]	35	PAN15	250	-0.0899	-0.0647	-0.0793	-0.1018	-0.1973	0.3087
Rahgouy et al.	[21]	35	PAN15	250	-0.0721	-0.0592	-0.0744	-0.0912	-0.1736	0.2899
Keswani et al.	[15]	35	PAN15	250	-0.0599	-0.0468	-0.0612	-0.0645	-0.1298	0.2543
Bakhteev et al.	[1]	35	PAN15	250	-0.0593	-0.0572	-0.0651	-0.0701	-0.1314	0.2172
Mansoorizadeh et al.	[17]	35	PAN15	250	-0.0375	-0.0339	-0.0420	-0.0502	-0.0994	0.1952
Kocher and Savoy	[16]	35	PAN15	250	-0.0704	-0.0661	-0.0508	-0.0304	-0.0676	0.1137