Overview of the Author Obfuscation Task at PAN 2017: Safety Evaluation Revisited

Matthias Hagen, Martin Potthast, and Benno Stein

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

Abstract We report on the second large-scale evaluation of style obfuscation approaches in a shared task on author obfuscation, organized at the PAN 2017 lab on digital text forensics. Author obfuscation means to automatically paraphrase a given text such that state-of-the-art authorship verification approaches misjudge a given pair of documents as having been written by "different authors" if in fact they would have decided otherwise without obfuscation. This year, two new obfuscators are compared to the participants from last year's task against a total of 44 authorship verification approaches. The best-performing obfuscator successfully impacts the decision-making process of the authorship verifiers significantly. However, as in the last year, the paraphrased texts are often not really human-readable anymore and have some changed context, indicating that there is still way to go to "perfect" automatic obfuscation that (1) tricks verification approaches, (2) keeps the meaning of the original, and (3) is, regarding its obfuscation, unsuspicious to a human eye.

1 Introduction

At PAN 2017 we organized the second shared task on author obfuscation in order to foster exploring the potential vulnerabilities of author identification technology. Like in the first edition, the specific task is that of *author masking* against *authorship verification*, which in turn has been a shared task at PAN 2013–2015 [11, 17, 18]. The following synopses point out the differences:

Authorship Verification	vs.	**Author Masking**
Given two documents, decide whether both have been written by the same author.		Given two documents from the same author, paraphrase the designated one such that an authorship verification will fail.

Figure 1 illustrates the setting and shows that the two tasks are diametrically opposed to each other: Success of a certain approach for one of these tasks depends on its "immunity" against the most effective approaches for the other. In our overview of last year's first author masking edition [16], we already included a survey of related work on author obfuscation. In particular, we introduced and discussed the "obfuscation impact measures" used in the evaluation, which we will quickly recap in Section 2. Section 3

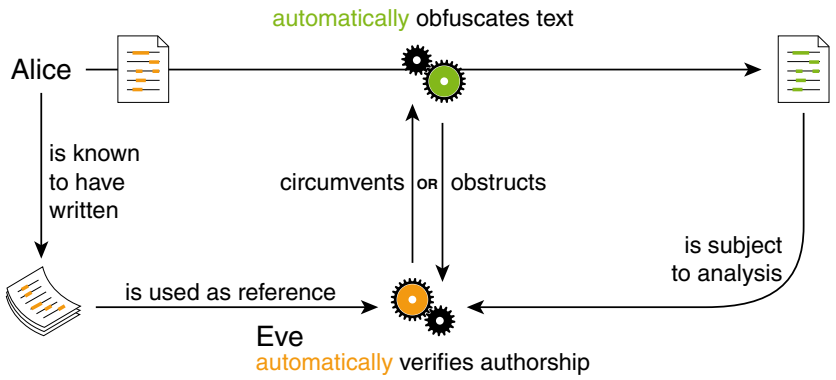

Figure 1. Schematic view of the opposed tasks of author masking and authorship verification.

reviews the obfuscation approaches that have been submitted to this year's edition of the shared task, and Section 4 reports on their evaluation against the state of the art in authorship verification.

2 Evaluating Author Obfuscation

As of last year, 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. 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 textually entailed by 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 measures for these dimensions has been included in our overview from the last year [16]. In order to analyze the safety dimension, we run the obfuscated texts against 44 authorship verification approaches and measure the impact of the obfuscation on the verifiers in form of changed verification decisions (cf. last year's overview for details on the used measures [16]). As for sensibleness and soundness we stick to manual inspection and grading of examples.

3 Survey of Submitted Obfuscation Approaches

The two approaches submitted to this year's edition of our shared task follow different strategies: sequence-to-sequence models and rule-based replacements. While a more

conservative rule-based strategy often changes the to-be-obfuscated text only slightly, the sequence-to-sequence modeling can lead to substantial differences.

Bakhteev and Khazov The approach of Bakhteev and Khazov [1] is mainly based on different sequence-to-sequence models and some small set of rules. The rules replace contractions (e.g., 'll → will), split or concatenate sentences using conjunctive words (e.g., and), and add or remove introductory phrases (e.g., anyway) to and from sentences respectively. The main idea of sequence-to-sequence modeling comes in two flavors: (1) replacing synonyms based on nearest neighbors in word embeddings from a Wikipedia dump, and (2) an encoder-decoder approach that generates some "reproduced" version of the original text, which is also based on embeddings trained on a Wikipedia dump. In both cases, the author choose from different possible variants of an obfuscated sentence that one that best matches a language model trained on Shakespeare texts.

As for the resulting texts, the strategy for combining and splitting sentences should pay more attention to the local situation, since otherwise it will quickly lead to incomplete or overlong constructions. A more detailed analysis of the text quality follows in the evaluation (cf. Section 4).

Castro et al. The approach of Castro et al. [6] focuses on simple rule- or pattern-based replacements. Using the FreeLing NLP tool for pre-processing texts (POS tagging, word sense disambiguation, etc.) several ideas are combined. Contractions are replaced based on a dictionary or the long version if it is used more often, synonyms are substituted using FreeLing functionality, and sentences are shortened by leaving out parts in parentheses, by leaving out discourse markers, or by eliminating appositions based on two simple patterns that identify explanations if named entities are introduced in the text.

The resulting text will usually be shorter than the original text, which, however, is intended by the authors. Most of the removals do not dramatically change the meaning of the text; a similar observation applies to the treatment of contractions. Still, leaving out information from the original may render parts of the resulting text hard to understand. Depending on FreeLing's synonym functionality, synonyms are often not appropriately chosen since the context seems not to be considered when selecting a replacement candidate. A more detailed analysis of the text quality follows in the evaluation (cf. Section 4).

4 Evaluation

As in the last year, we automatically evaluate the safety of the submitted obfuscation approaches against 44 authorship verifiers which have been submitted to the previous three shared tasks on authorship identification at PAN 2013–2015. Sensibleness and soundness of the obfuscated texts are assessed manually by human inspection.

The evaluation setup is the cloud-based evaluation platform TIRA [9, 15],[1] which is being developed as part of our long-term evaluation-as-a-service initiative [10]. We

[1] www.tira.io

Table 1. Safety evaluation of five obfuscators, including those submitted to PAN 2016, 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**			**Obfuscation 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.	[14]	36	PAN13	14	**-0.1066**	**-0.0759**	**-0.1030**	**-0.1389**	**-0.2778**	**0.4690**
Castro et al.	[6]	36	PAN13	14	-0.1106	-0.0545	-0.0920	-0.1248	-0.2449	0.4175
Keswani et al.	[12]	36	PAN13	14	-0.0908	-0.0695	-0.0940	-0.1148	-0.2361	0.4245
Bakhteev et al.	[1]	36	PAN13	14	-0.0518	-0.0547	-0.0631	-0.0796	-0.1667	0.2881
Mansoorizadeh et al.	[13]	36	PAN13	14	-0.0422	-0.0254	-0.0392	-0.0463	-0.0933	0.1442
Mihaylova et al.	[14]	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.	[12]	26	PAN14 EE	100	-0.1085	-0.0870	-0.0960	-0.0975	-0.1873	0.4058
Bakhteev et al.	[1]	26	PAN14 EE	100	-0.0518	-0.0453	-0.0509	-0.0631	-0.1177	0.2558
Mansoorizadeh et al.	[13]	26	PAN14 EE	100	-0.0514	-0.0463	-0.0473	-0.0577	-0.1038	0.2512
Mihaylova et al.	[14]	36	PAN14 EN	100	**-0.1613**	**-0.1050**	**-0.1260**	**-0.1456**	**-0.2456**	**0.4750**
Castro et al.	[6]	36	PAN14 EN	100	-0.1335	-0.0793	-0.1014	-0.1149	-0.1900	0.3811
Keswani et al.	[12]	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.	[13]	36	PAN14 EN	100	-0.0579	-0.0408	-0.0493	-0.0665	-0.0958	0.2345
Mihaylova et al.	[14]	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
Keswani et al.	[12]	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.	[13]	35	PAN15	250	-0.0375	-0.0339	-0.0420	-0.0502	-0.0994	0.1952

want to point out that—by using TIRA—it was 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 submitted obfuscation approaches. The outputs, in turn, were generated from the authorship verification corpora PAN13, PAN14 EE, PAN14 EN, and PAN15.

4.1 Safety

Table 1 shows the results of our safety evaluation of the two approaches from this year compared to the three approaches from last year against 44 authorship verification approaches on the aforementioned four PAN evaluation datasets. We combine the two rankings into an overall ranking of obfuscation approaches suggested so far in order to interpret the results of this year's participants in context.

The best-performing approach this year was submitted by Castro et al. [6], which achieve second rank overall across both 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 the same as that induced by all other measures, rendering the measures perfectly consistent. This consistency, however, forecloses more insights that can usually be derived from differing performance characteristics. In this regard, the qualitative assessment of sensibleness and soundness presented in the following subsection is important. Altogether, the approach of Mihaylova et al. [14] still performs best among all approaches.

4.2 Sensibleness and Soundness

As in last year's edition, a human assessor conducted an in-depth manual assessment on problem instances 6, 135, and 430. Spot checks on other instances again 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 last year were a grade 2 for Mansoorizadeh et al.'s approach [13] that does not really change a lot on a per sentence basis, a grade 4 for Mihaylova et al.'s obfuscator [14], and a grade 5 for Keswani et al.'s obfuscator [12]. This year's approaches get a grade 4 for Bakhteev's and Khazov's approach [1], since there are a lot of issues with respect to uppercasing at sentence starts as well as many grammatical problems due to problematic sentence splits and merges, and due to inappropriate use of synonyms. As for Castro et al.'s approach [6] grade 2s were assigned if only some problematically short sentences were grammatically incorrect or if spacing around punctuation marks was incorrect, while other documents got a grade 3 for too short sentences that were grammatically wrong or for synonyms not making sense in some contexts.

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 and Keswani et al.'s approaches were all judged "incorrect", while Mansoorizadeh et al.'s very conservative approach achieved "correct" and "passable" scores. This year's approaches (Bakhteev's and Khazov's, and Castro et al.'s) both got "incorrect" as judgments—but for different reasons: With regard to Bakhteev's and Khazov's approach, many parts of the resulting texts were not understandable anymore because of overly rigid changes in sentences, which completely removed the original meaning. With regard to Castro et al.'s approach, the judgment results from the fact that the obfuscated text covers only a small portion of the original text (about the first third of the original), maybe an undesired side-effect due to some pre-processing problems. The parts that are still contained in the obfuscated version often achieve at least a "passable" judgment, and they could even be judged as "correct". However, the fact that about two thirds of the original was omitted precluded a better outcome.

5 Conclusion and Outlook

In the second 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 three approaches from last year. The best-performing obfuscator flips on average about 42% of an authorship verifier's decisions towards choosing "different author" when the opposite decision would have been correct, indicating some level of safety against verification approaches. As for soundness and sensibleness, though, the approaches often produce rather unreadable text or text whose meaning is significantly changed. Still, such insights are mainly obtained from manual inspection.

The challenge of evaluating author obfuscation approaches properly and at scale would definitely benefit from new technologies that are capable of recognizing paraphrases, textual entailment, grammaticality, and style deception. However, a very important direction for future research in the authorship obfuscation domain is that on producing safe and still sound and sensible texts. So far, there are only two groups of obfuscation approaches: (1) approaches that are somewhat safe but that often produce unreadable text or text that is neither sound nor sensible, and (2) approaches that produce sound and sensible texts but that are not safe against authorship verification.

A significant improvement of current obfuscation technology requires a much better consideration and integration of the surrounding context when replacing, adding, or removing words. Note that such kind of sensible text operations can also be operationalized by applying paraphrasing rules from the PPDB [8], as is done for instance in an approach on constrained paraphrasing [19].

Acknowledgments

We thank the participating teams of the two 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-/

[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., 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

[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-/
[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] 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
[11] Juola, P., Stamatatos, E.: Overview of the Author Identification Task at PAN 2013. In: [7]
[12] 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/
[13] 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/
[14] 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/
[15] 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)
[16] 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, CLEF and CEUR-WS.org (Sep 2016), http://ceur-ws.org/Vol-1609/
[17] 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: [5]
[18] 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]
[19] 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)