Evaluating Safety, Soundness and Sensibleness of Obfuscation Systems
Notebook for PAN at CLEF 2016

Matthias Liebeck, Pashutan Modaresi, and Stefan Conrad

Institute of Computer Science
Heinrich Heine University Düsseldorf
D-40225 Düsseldorf, Germany
{liebeck, modaresi, conrad}@cs.uni-duesseldorf.de

Abstract Author masking is the task of paraphrasing a document so that its writing style no longer matches that of its original author. This task was introduced as part of the 2016 PAN Lab on Digital Text Forensics, for which a total of three research teams submitted their results. This work describes our methodology to evaluate the submitted obfuscation systems based on their safety, soundness and sensibleness. For the first two dimensions, we introduce automatic evaluation measures and for sensibleness we report our manual evaluation results.

1 Introduction

Author masking is the task of paraphrasing a document so that its writing style no longer matches that of its original author. Due to the advances in fields such as *authorship attribution* and *author verification*, it is not clear whether authors (particularly in the age of the Internet and social media) can assure their anonymity anymore [20]. While in some scenarios, such as verifying the authorship of disputed novels or revealing the author of harassing messages in social media [19], author unmasking might be useful, there are situations where authors have the right to protect their privacy, among them the desire to avoid retribution from an employer or government agency [10].

The task of author masking was introduced as part of the 2016 PAN Lab on Digital Text Forensics [5], for which a total of three research teams, namely Mansourizade et al. [13], Keswani et al. [11] and Mihalvoya et al. [14] (called Team A, B and C respectively in the rest of this work) submitted their results. The evaluation was completely anonymous and the identities of the teams were revealed after the submission of our evaluation results.

Together with the task of author masking, *obfuscation evaluation* has been introduced as another task to evaluate the performance of the author masking submissions. Three dimensions have been defined by the task organizers for the performance evaluation of the obfuscation systems: *safety* to ensure that a forensic analysis does not reveal the original author of an obfuscated text; *soundness* to evaluate if the obfuscated texts are textually entailed with their originals; and *sensibleness* to ensure that the obfuscated texts are inconspicuous [18].

In this work, we describe our methodology to evaluate the performance of the submitted systems based on the aforementioned dimensions. In section 2 we define the problem of author masking more concretely and describe the provided training data. The evaluation results of the dimensions safety, soundness and sensibleness are reported in sections 3, 4 and 5, respectively. Finally, we conclude our work in section 6.

2 Problem Definition

Given a document, an author masking software has to paraphrase it so that its writing style no longer matches that of its original author. Although the organizers of the author masking task do not directly define this task as a supervised machine learning problem, a training set is provided so that the participant can evaluate their designed algorithms based on this dataset. The same dataset is also used as the test dataset for the final evaluation.

The provided dataset is a collection of 205 problems selected from author verification tasks from PAN2013 [9], PAN2014 [22] and PAN2015 [21]. Each problem is a collection of at most five known documents (written by the same author) and a questioned document. Normally in author verification problems, the author of the questioned document is unknown and the task of an author verifier is to figure out whether the questioned document has the same author as the known documents or not. But in the training dataset of the author masking task, all problems are selected from positive instances, meaning all questioned documents have the same author as the known documents. The language of all provided problems is English.

The participants were asked to develop a software that outputs a detailed list, how each piece of the original text has been paraphrased. For a detailed description of the desired system output, the reader is referred to the official task page[1].

3 Safety

An obfuscation software is called safe, if a forensic analysis does not reveal the original author of the obfuscated texts. We evaluate the safety of the obfuscation software using an automatic author verifier called GLAD [8]. The idea behind this automatic evaluation measure is that if an obfuscation system successfully masks the authors of the *questioned* documents in the training set (remember that all problems in the training set belong to the positive class), the author verifier will classify the problems as negative (meaning that the obfuscated document no longer has the same author as the other documents).

The GLAD algorithm was one of the top ranked systems at PAN2015 and treats the author verification problem as an intrinsic binary classification machine learning task. GLAD uses SVM [7] as a learning algorithm and makes use of simple feature classes such as N-Grams, tokens, sentences, visual, compression, entropy and syntactic features [8].

[1] http://pan.webis.de/clef16/pan16-web/author-obfuscation.html

To train the GLAD algorithm we used the English problems from the training datasets from PAN2013 to PAN2015. The statistics of the training dataset used are shown in Table 1.

Table 1. Statistics of the dataset used to train GLAD

	#Problems	#Documents	Avg. #Known Documents	Avg. #Tokens
PAN2015	100	200	1.0	366
PAN2014 (Essays)	200	725	2.6	848
PAN2014 (Novels)	100	200	1.0	3137.8
PAN2013	10	42	3.2	1037

Notice that the training dataset from PAN2014 consisted of novels and essays and we took both categories to train our model.

Using the trained model, we measure the performance of the GLAD algorithm once using the original documents from the training set of the author masking problem (labels are all considered to be true), and once on the submissions of each team (labels are all considered to be true). As the evaluation measure we use $c@1$ [17], as defined in Equation 1. The $c@1$ measure is a modified version of accuracy, where $c@1$ rewards approaches that maintain the same number of correct answers and decrease the number of incorrect answers by leaving some problems unanswered.

$$c@1 = \frac{1}{n} \cdot \left(n_c + \frac{n_u n_c}{n}\right) \tag{1}$$

In the definition of $c@1$ measure, n is the number of problems, n_c is the number of correct answers and n_u represents the number of unanswered problems.

Table 2 summarizes the evaluation results of the safety dimension. On the original dataset (the training dataset of author masking), the GLAD algorithm has a $c@1$ score of 0.656, meaning that the algorithm correctly classifies roughly 65% of the problems as positive. Notice that the gold standard labels are all set to be true and that teams having lower $c@1$ scores are more successful at the masking task than the other teams. It is also observable that none of the participants has a $c@1$ score below 0.5. This means that the portion of successful obfuscations for all submissions is below 50%. At the same time it can be seen that all submissions have a $c@1$ score below the baseline 0.656, meaning that all teams were partly successful at the author masking task.

Table 2. Evaluation results of the safety dimension

	Team A	Team B	Team C	Original
C@1	0.585	0.532	0.522	0.656

Although in previous PAN competitions, AUC (Area Under the Curve) [6] was also used to evaluate the automatic verifiers, the use of this measure was not possible in our scenario as the test datasets contain either only positive or only negative instances.

Another interesting analysis is to investigate the relation between true positives and false negatives. The idea behind this analysis is to figure out the portion of documents classified as positive before obfuscation, and the ones classified as negative after obfuscation. For this we select true positives from the original dataset and count the ones that have been classified as negative by the GLAD algorithm. Table 3 summarizes the results.

Table 3. Evaluation results of the safety dimension

	Team A	Team B	Team C
FN / TP	0.159	0.254	0.290

Notice that higher values in Table 3 are preferred. Team C has the highest score among the teams and has managed to obfuscate roughly 30% of the true positive problems to false negative ones. These results are consistent with the results shown in Table 2.

4 Soundness

We assume that the goal of author masking is to reword a text segment into a paraphrased one while retaining as much semantic similarity as possible. Therefore, we propose to quantify soundness by measuring the semantic textual similarity (STS) between the original text segment and its corresponding obfuscation.

The prediction of semantic textual similarity has been a recurring task in SemEval challenges since 2012 [1–4]. The aim of the STS task is to determine the semantic similarity of two sentences in the continuous interval $[0, 5]$ where 0 represents a complete dissimilarity and 5 denotes a complete semantic equivalence between the sentences. The task organizers provide sentence pairs with gold standards from different categories. The task is evaluated by calculating the Pearson correlation between the predicted values and a crowdsourced gold standard.

In this paper, we use the unsupervised semantic similarity approach called *Overlap* [12] to automatically determine the semantic similarity between the original segment and its paraphrase. There are two advantages of using an unsupervised approach: (i) human annotators can only annotate a subset of the paraphrases within a reasonable amount of time. An automatic approach can evaluate all original-paraphrase pairs and (ii) we do not need labeled training data as compared to a supervised approach.

The idea of the *Overlap* method is simple since it measures the overlap between the tokens in the original segment s_1 and the tokens in the paraphrase s_2 by aligning tokens to the best match in the other text segment. The authors first define a similarity function for two tokens which uses synsets from *WordNet* [16] and word embeddings from word2vec [15], as denoted in Equation 2.

$$\mathrm{sim}(t_1,t_2) := \begin{cases} 1 & \text{if } t_1\text{.lemma == } t_2\text{.lemma} \\ 1 & \text{if } t_1 \text{ and } t_2 \text{ have the same most common synset} \\ 0.5 & \text{if } t_1 \text{ and } t_2 \text{ share any other synset} \\ \cos(t_1,t_2) & \text{if } t_1 \text{ and } t_2 \text{ have \textit{word2vec} embeddings} \\ \textit{0.15} & \text{otherwise} \end{cases} \tag{2}$$

Afterwards, the similarity score between two text segments in $[0, 5]$ is defined as follows:

$$\mathrm{STS}(s_1,s_2) := 5 \cdot \left(\frac{\sum\limits_{t_1 \in s_1} \max\limits_{t_2 \in s_2} \mathrm{sim}(t_1,t_2)}{2 \cdot |s_1|} + \frac{\sum\limits_{t_2 \in s_2} \max\limits_{t_1 \in s_1} \mathrm{sim}(t_2,t_1)}{2 \cdot |s_2|} \right) \tag{3}$$

Since we assume the obfuscations to be semantically as close as possible to the originals, the STS score between both segments should be 5. We predict the semantic similarity for all pairs for each team. Afterwards, we average over the predicted scores for each team. Table 4 summarizes the results for the soundness dimension.

Table 4. Evaluation results of the soundness dimension

	Team A	Team B	Team C
Mean STS	4.87	4.04	4.48

For the soundness dimension, the best semantic paraphrases were created by team A with an average STS score of 4.87. This is not surprising since team A only substituted a few words and often kept the original segment as a paraphrase. Therefore, the paraphrases are semantically very close or even identical to the original. Team C achieved a mean STS score of 4.48 and team B had the lowest score with 4.04. Since the *Overlap* approach from [12] is independent of the word order, the results of team B cannot be explained by changing the word order of the phrases. One factor that definitely influenced the semantic similarity is the appearance of German words in the paraphrases, which cannot be matched to the English tokens in the original texts.

5 Sensibleness

The dimension sensibleness describes the language quality of the obfuscations and whether it allows us to understand them. An author masking software might mask the author of a text at the cost of its comprehension. Therefore, it is also crucial to evaluate the quality of the produced obfuscations.

We observed that teams A and C used dictionaries to perform simple substitutions and team B usually changed the order of phrases. It is surprising to see that the paraphrases by team B sometimes contain random German words, as in the following example: "*it is difficult to across, Once the Mitbürgers unschön is faint, odor street, on the village so massed mold Verfalls and centuries.*"

Although there are approaches to automatically predict the grammatical quality of text, we chose to manually evaluate the sensibleness because portions of the text have a low language quality but still allow for a limited understanding of the content. For example, this can be compared to a non-native speaker who asks in an online forum a question that is poorly worded but still comprehensible.

After a manual inspection of a subset of the paraphrases from all three teams, we decided to annotate each pair with a score $s \in \{0, 1, 2\}$ to measure the language quality. We then drew a small sample and discussed annotation guidelines. Our three labels and their definitions are described in Table 5.

Table 5. Labels for the sensibleness dimension

Score	Name	Definition
2	*comprehensible*	The paraphrase can be understood immediately. Example: "*These things are deeply rooted in the Swedish people.*"
1	*partially comprehensible*	The paraphrase can be understood with some restrictions. It can contain smaller errors or some smaller parts that are incomprehensible. Example: "*they him. But ignored*"
0	*incomprehensible*	The language quality of the paraphrase is too low to allow any understanding of the content. Example: "*I a In certain years in a bookstore can help , than English , French English. French*"

In our evaluation, sensibleness is only evaluated by looking at the obfuscated text. This is due to the fact that only the paraphrased text after author masking is used in a real world scenario. Therefore, it is reasonable to only evaluate the output of the system. We ignore spacing and line breaks during the annotation process. Furthermore, we also ignore the substitutions of the words "*oof*" and "*tto*" from team C because they do not impact the understanding of the text.

We randomly drew a subset of 20 problems. For each team, we then drew three obfuscations per problem. All of these obfuscations were manually annotated by three annotators. In order to report a single value per team, we averaged all the scores from the annotators. Table 6 summarizes the results for the sensibleness dimension.

Table 6. Evaluation results of the sensibleness dimension

	Team A	Team B	Team C
Average score	1.94	0.57	1.20

Team A achieved the best results in the sensibleness dimension with an average score close to 2. The paraphrases from team B allow for the lowest understanding of all three teams with an average score of 0.57 which is between *partially comprehensible* and *incomprehensible*.

We should note that there are at least two problems for the evaluation of the sensibleness dimension: (i) it is difficult to formalize language quality and understanding and (ii) the sensibleness dimension is subjective. Although we observed a high agreement on the category *incomprehensible*, we had a lower agreement on whether a paraphrase is fully or partially comprehensible. This is plausible since one annotator might perfectly understand a text segment while another annotator may have some troubles with it.

6 Conclusion

In this work, we discussed our methodology to evaluate the performance of the obfuscation systems submitted to the PAN2016 Author Masking shared task. More concretely, submissions were evaluated based on their safety (Section 3), soundness (Section 4), and sensibleness (Section 5). The scripts for our evaluation are available on GitHub[2].

An automatic author verifier was used to measure the safety of the submissions. The ranking of the teams in terms of safety is as follows: team C, B, and A

We proposed to quantify soundness by automatically measuring the semantic text similarity between the original text fragments and their obfuscations. The best score was achieved by team A, followed by teams C and B.

Unlike the first two dimensions, the sensibleness of the submissions was evaluated manually. As sensibleness is subjective and difficult to formally define, we consider its measurement a nontrivial task. Regarding sensibleness, teams A, C and B were ranked first, second, and third, respectively.

Acknowledgments

This work was partially funded by the PhD program *Online Participation*, supported by the North Rhine-Westphalian funding scheme *Fortschrittskollegs* and by the German Federal Ministry of Economics and Technology under the ZIM program (Grant No. KF2846504). We would like to thank Daniel Braun for his help in the evaluation of the sensibleness dimension.

References

1. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., Wiebe, J.: SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In: Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). pp. 252–263. Association for Computational Linguistics (2015)
2. Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Mihalcea, R., Rigau, G., Wiebe, J.: SemEval-2014 Task 10: Multilingual Semantic Textual Similarity. In: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). pp. 81–91. Association for Computational Linguistics and Dublin City University (2014)

[2] https://github.com/pasmod/obfuscation

3. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A.: SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity. In: *SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012). pp. 385–393. Association for Computational Linguistics (2012)
4. Agirre, E., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W.: *SEM 2013 shared task: Semantic Textual Similarity. In: Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity. pp. 32–43. Association for Computational Linguistics (2013)
5. Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.): CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers. CEUR Workshop Proceedings, CEUR-WS.org (2016)
6. Bradley, A.P.: The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recogn. 30(7), 1145–1159 (Jul 1997)
7. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
8. Hürlimann, M., Weck, B., van den Berg, E., Suster, S., Nissim, M.: GLAD: Groningen Lightweight Authorship Detection. In: Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum (2015)
9. Juola, P., Stamatatos, E.: Overview of the Author Identification Task at PAN 2013. In: CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers (2013)
10. Kacmarcik, G., Gamon, M.: Obfuscating Document Stylometry to Preserve Author Anonymity. In: Proceedings of the COLING/ACL on Main Conference Poster Sessions. pp. 444–451. COLING-ACL '06, Association for Computational Linguistics (2006)
11. Keswani, Y., Trivedi, H., Mehta, P., Majumder, P.: Author Masking through Translation—Notebook for PAN at CLEF 2016. In: CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2016)
12. Liebeck, M., Pollack, P., Modaresi, P., Conrad, S.: HHU at SemEval-2016 Task 1: Multiple Approaches to Measuring Semantic Textual Similarity. In: Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016). pp. 607–613. Association for Computational Linguistics (2016)
13. Mansourizade, M., Rahgooy, T., Aminiyan, M., Eskandari, M.: Author Obfuscation using WordNet and Language Models—Notebook for PAN at CLEF 2016. In: CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2016)
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: CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2016)
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. ICLR Workshop (2013)
16. Miller, G.A.: WordNet: A Lexical Database for English. Communications of the ACM 38(11), 39–41 (1995)
17. Peñas, A., Rodrigo, A.: A Simple Measure to Assess Non-response. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1. pp. 1415–1424. HLT '11, Association for Computational Linguistics (2011)
18. 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)
19. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2015)
20. Rao, J.R., Rohatgi, P.: Can Pseudonymity Really Guarantee Privacy? In: Proceedings of the 9th USENIX Security Symposium. pp. 85–96. USENIX (2000)

21. 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: CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (2015)
22. 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: CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers. CEUR-WS.org (Sep 2014)