UniNE at CLEF 2018: Author Masking

Notebook for PAN at CLEF 2018

Mirco Kocher and Jacques Savoy

Computer Science Dept., University of Neuchâtel, Switzerland
{Mirco.Kocher, Jacques.Savoy}@unine.ch

Abstract. This paper describes and evaluates an author masking model to obfuscate the writer of a document. The suggested strategy works in English with different text genres (e.g., essays, novels, poems) and various text sizes (e.g., from less than 500 to 4,000 tokens). The approach mainly focuses on retaining high soundness and sensibleness in the obfuscated texts with the reduced set of modifications. To improve the safety, rules with a high probability of correctness are applied by attacking the feature frequencies. Depending on the writing style in the comparable documents of an author, a feature is either increased or decreased in the masked text. The evaluations are based on 205 training and 464 test problems (PAN AUTHOR OBFUSCATION task at CLEF 2018).

1 Introduction

Stylometry is an interesting problem in computational linguistics but also in applied areas such as criminal investigation and historical studies where knowing the author of a document (such as a ransom note) may be able to save lives [14]. With the Web 2.0 technologies, the number of anonymous or pseudonymous texts is increasing, and in many cases, one person writes in different places about different topics (e.g., multiple blog posts written by the same author). Therefore, proposing an effective algorithm to the authorship identification problem presents a real interest. Detecting the author style has been studied for years and different approaches have been explored.

However, the reverse process of obfuscating the style of an author is less studied. There are many challenges in different directions. Of course, the author style must be hidden, but also, the text needs to remain syntactically correct, and the semantics of the original document should be retained. The challenge is to use the information in a provided set of documents to mask the original document. Therefore, by analyzing someone's usual writing style, a text must be transformed to obfuscate the writer.

This paper is organized as follows. After the presentation of the related works, the next section presents the evaluation methodology and test collection used in the experiments. The fourth section explains our proposed masking algorithm. Then, we evaluate the proposed scheme and compare it to the other participants. A conclusion draws the main findings of this study.

2 Related Work

Author identification is a well-studied topic and was explored in the PAN lab for years. Juola et al. [5] created JGAAP (Java Graphical Authorship Attribution Program) that can use distinctive features, e.g., words, parts of speech, and characters or word *n*-grams, to solve author identification problems. The PAN 2017 task overview paper [14] summarizes the approaches and features used for author identification by different participants. Among the most used features are the lengths of words, sentences, or paragraphs, type-token ratios, and frequencies of hapax legomena, *n*-grams, words, punctuation marks, or parts of speech.

Kacmarcik and Gamon [6] masked an author's text by detecting the most used words and tried to change them. They also mention the application of machine translation as a possible approach for author obfuscation. Machine translation was also used as a means for author obfuscation ([6], [12]) where passages of text from English were translated to at least one other language and then back to English. The main advantage of this method is a strong modification of the original text, but the disadvantage is that there are many untranslated words and it can result in weak semantic coherence of the obfuscated text.

Brennan et al. [2] investigated three different approaches for adversarial stylometry, namely obfuscation (masking author style), imitation (trying to copy another author's style), and machine translation. They have summarized the features people use most when trying to obfuscate their own writing style [9]. Another approach used in [8] is to synonymize the most frequent words of the original text. This method keeps the meaning of the text in most of the cases but gives a small number of modifications of the original text. The best result of the metrics used in the PAN lab can be achieved by combining strong context modifications and preserving the original sense of the text [9] by using several types of text obfuscation.

Juola et al. [5] experimented with different techniques for author obfuscation. Their system consists of three main modules, i.e., canonization (unifying cases, normalizing whitespaces, spelling correction, etc.), event set determination (extraction of events significant for author detection, such as words, parts of speech *n*-grams, etc.), and statistical inference (measures that determine the results and confidence in the final report). The same authors used this approach [4] to detect deliberate style obfuscation. Some other features used for author recognition are personal pronouns, sentence length, unique words, and parts of speech [1].

Statistical and context features are used in modern detecting authorship approaches, for example, in GLAD [3]. In our participation, we studied the feature frequencies to mask the author style, i.e., to address the Author Obfuscation task.

3 Evaluation Methodology and Test Collection

The evaluation was performed using the *TIRA* platform, which is an automated tool for deployment and evaluation of the software [10]. The data access is restricted such that during a software run the system is encapsulated and thus ensuring that there is no data

leakage back to the task participants. This evaluation procedure also offers a fair evaluation of the time needed to produce an answer.

For each obfuscation problem, there was one document that had to be obfuscated and a set of other documents from the same author. The goal was to mask one document such that its writing style is different from the others [13]. In this context, the task is defined as follows:

> *Given a document, paraphrase it so that its writing style does not match that of its original author, anymore.*

The organizers have proposed the following parameters for the evaluation of the author masking task. The quality of all submitted systems is assessed based on the following three questions:

1. Safeness: does forensic analysis reveal the original author of its obfuscated texts?
2. Soundness: are the obfuscated texts textually entailed with their originals?
3. Sensibleness: are the obfuscated texts inconspicuous to a human reader?

These dimensions are orthogonal; an obfuscation software may meet any of them to various degrees of perfection. If no modification is performed at all, the obfuscation would be sound and sensible but not safe. To assess the performance in soundness and sensibleness, the obfuscations are sampled and handed out to participants for manual peer-review [11].

The task organizers evaluated safeness. The obfuscated texts were tested with four authorship verification models, namely *Caravel* (the best-performing verification approach at PAN 2015), *GLAD* (Groningen Lightweight Authorship Detection), *Authorid* (model using Bayes, imposters, and sparse representation), and *AuthorIdentification-PFP* (a universal background approach based on random forest with increased generalization). Furthermore, all participants were invited to submit automatic performance measures in the corresponding task called Obfuscation Evaluation.

All the texts were written in English, and for each problem, there were between one and five documents from the same author in addition to the original document that had to be obfuscated. The text length varied considerably between problems. In approximation, we saw three different sections, the first 105 problems had less than 1,000 word tokens (<5,000 characters) per text, the next 50 problems contained almost 4,000 words (>20,000 characters), and the last 50 problems were small again with <500 tokens (<2,000 characters) in each document. Different problems also originated from different genres. There were extracts from scientific books, personal and topical essays, self-evaluations, reviews, passages from novels, poems, and plays.

As test collections, the data sets from the previous years were used, namely 14 problems from PAN13, 100 problems each English essays and novels from PAN14, and 250 from PAN15.

4 Masking Algorithm

We applied an obfuscation system with simple conditions, search objects, and replacement rules. Our method focused on attacking frequency features to trick verification systems based on the bag of word approach while leaving out, for instance,

the average sentence length or Boolean features. For each problem, we have a document that must be masked (called *original*) and a set of similar texts from the author (called *same*). Therefore, we compare the frequency of a feature in *original* and *same*. If the feature is more frequent in *original*, then we try to increase it even more in the masked text to make is more dissimilar from *same*. If no condition of any rule is met, then the obfuscated text is the same as the *original*.

An overview of our obfuscation rules can be seen in Table 1. In the first rule, as an example, if the abbreviation of "*to be*" and "*not*" is more common in *same*, we expand all occurrences in *original* for our obfuscated text. The second rule would do the reverse and contract those versions to the "*to be*" and "*n't*" version. Besides a potential increase in safety, those rules do not infer with soundness and should also be sensible.

From the dataset with rule 3, we have "[…] a lot of stress *is being put* on language skills […]" which we transform to "[…] a lot of stress *is put* on language skills […]". The reverse in rule 4 would be "All this information *is stored* for each customer […]" which we transform to "All this information *is being stored* for each customer […]".

In the rules 5 and 6 we use a dictionary look-up with 142 entries of the format "*very X*" and for each, we have one or two synonyms. Table 3 in the Appendix shows all the word pairs. As an example, "*very good*" would be randomly replaced by either "*excellent*" or "*superb*" if the comparable texts contain the word "*very*" frequently (rule 5). In rule 6, the inverse procedure is performed, and both "*excellent*" and "*superb*" are replaced by "*very good*" to increase the frequency of "*very*" even more in *original*.

For rule 7, we have "[…] Marie *started introducing* them." in the dataset, which is transformed to "[…] Marie *introduced* them.". Due to the transformation of the verb, it is possible that this rule does not produce perfect a sensible obfuscation. As an example, "spinning" would be "spinned" and "reading" would be "readed".

The phrase " *in order to* " is usually redundant and we replace it with a simple " to " in case the phrase is less common in *original*. Rules 9 to 14 are simple replacement of two semantically equal strings depending on its appearances in *original* and *same*. The obfuscation is punctuated according to the original text, meaning that if the Oxford comma is found in the search phrase in rule 13 it is also used in the replacement part.

In rule 15, we reorder part of a sentence if it contains " *of the* " and if it is less common in the *original*. As an example, "[…] which *the author of the editorial* seems to imply." is transformed to "[…] which *the editorial author* seems to imply.". This rule may decrease the sensibleness in case where the second part is not a single word, as in "[…] New York at *the beginning of the 20th* century and […]" which would be obfuscated as "[…] New York at *the 20th beginning* century and […]". In retrospective, a Part of Speech tagger could have helped reducing the error rate in this case.

The rules 16 to 20 introduce some improper spellings with a fixed probability. The exclamation mark and question mark can be repeated up to three times or left as is. For words with repeated characters, e.g., "*excellences*", we add spelling mistakes by either adding the repeated letter once more, i.e., "*excelllences*", or removing one of its occurrences, i.e., "*excelences*". This is only done for 5% of the matches and randomly decided, which means that this part of the obfuscation is not deterministic.

Table 1. Obfuscation rules.

Rule	Condition	Search	Replace	Notes
1	more "*n't*" than "*not*" in *same*	"*isn't*" "*don't*" "*doesn't*" "*didn't*" "*wasn't*" "*weren't*" "*couldn't*" "*hasn't*" "*haven't*" "*can't*"	"*is not*" "*do not*" "*does not*" "*did not*" "*was not*" "*were not*" "*could not*" "*has not*" "*have not*" "*can not*"	-
2	more "*not*" than "*n't*" in *same*	↗	↖	vice versa from Rule 1
3	more "*is|are|was|were being*", than "*is|are|was|were X+ed*" in *same*	"*is being*" "*are being*" "*was being*" "*were being*"	""	-
4	more "*is|are|was|were X+ed*", than "*is|are|was|were being*" in *same*	"*is X+ed*" "*are X+ed*" "*was X+ed*" "*were X+ed*"	"*is being X+ed* " "*are being X+ed* " "*was being X+ed* " "*were being X+ed* "	-
5	more "*very X*" in *original*	" *Y* "	" *very X* "	list of 142 *X* and 161 *Y*
6	less "*very X*" in *original*	" *very X* "	" *Y* "	
7	less "*started X+ing*" in *original*	"*started X+ing*"	"*X+ed*"	-
8	less " *in order to* " in *original*	" *in order to* "	" *to* "	-
9	more " *in fact,?* " than " *actually,?* " in *same*	" *in fact,?* "	" *actually,?* "	optional comma
10	more " *actually,?* " than " *in fact,?* " in *same*	↗	↖	vice versa from Rule 9
11	more " *However,?* " than " *On the contrary,?* " in *same*	" *However,?* "	" *On the contrary,?* "	optional comma
12	more " *On the contrary,?* " than " *However,?* " in *same*	↗	↖	vice versa from Rule 11
13	more " *X, Y,? and Z*" than " *as well as* " in *same*	" *X, Y,? and Z*"	" *X, Y,? as well as Z*"	optional comma
14	more " *as well as* " than " *X, Y,? and Z*" in *same*	" *as well as* "	" *and* "	optional comma
15	less " *of the* " in *original*	" *the X of the Y*"	" *the Y X*"	-
16	more "*!*" in *original*	"*!*"	"*!!|!!!*"	only for 50%
17	less "*!*" in *original*	"*!*"	"."	
18	more "*?*" in *original*	"*?*"	"*?|??|???*"	
19	repeated character	" *XYYZ* "	" *XYZ* "	only for 5%
20			" *XYYYZ* "	

When inspecting the coverage of those rules in the training set, we see that not all texts were obfuscated with the same intensity. There are texts which are only modified by the probabilistic rules (rule 16 – 20) because none of the conditions was satisfied or because the search was not able to do any valid replacements. However, some texts met many of the conditions and many opportunities for changes were found.

5 Evaluation

A human assessor conducted an in-depth manual sensibleness assessment on a subset of the data who assigned school grades (on a scale from 1 (excellent) to 5 (fail)). With the limited number of implemented changes, we obtained the grade 1-2, depending on the inspected problem. This was planned and expected for our system.

Afterward, the assessor read the original texts and judged the textual differences in several ways to evaluate the soundness of the obfuscated texts on a three-point scale as either "1 = correct", "2 = passable", or 3 = "incorrect". Our approach was marked "passable" because our system changed the ordering of the sentences resulting in passages that were not clear. This modification was not intended, and the source of this reformation is not clear at the time of writing.

In Table 2 we can see a summary of all evaluation results as a macro average over the distinct types of data sets. Specifically, we reported the average of the safety performance over four data sets (PAN13, PAN14 essay, PAN14 novel, and PAN15). The *AUC*, *C@1*, and *final* scores are performance measures from the Author Identification [14] task. The goal was to reduce those values, meaning the verifiers were not able to confirm a shared authorship anymore and lower (more negative) scores were better. The *acc*, *rec*, and *imp* refer to the performance measures specifically from the Author Masking task. In the first column, the reference indicates the team reference as used in the overview paper [11]. Our approach is in the highlighted row with the number 12. In this year, they only had one other participant, that is Rahgouy et al. with reference 17. The table includes results from the previous years to have a better overview of the different approaches. In the year 2017, two teams participated, namely Bakhteev and Khazov (ref. 1) and Castro et al. (ref. 5). In the year 2016, there were another three teams, specifically Keswani et al. (ref. 11), Mansoorizadeh et al. (ref. 13), and Mihaylova et al. (ref. 14).

Table 2. Macro average of evaluation results ordered by *final*.

Team	Safety						Sensibleness	Soundness
	AUC	C@1	final	acc	rec	imp		
14	-0.1265	-0.0956	-0.1131	-0.1281	-0.2387	0.4495	4.0	3.0
5	-0.1157	-0.0770	-0.0967	-0.1158	-0.2149	0.3850	2.5	3.0
17	-0.1082	-0.0822	-0.0884	-0.0997	-0.1882	0.3664	3.0	2.0
11	-0.0903	-0.0684	-0.0839	-0.0961	-0.1829	0.3654	5.0	3.0
12	-0.1180	-0.1050	-0.0760	-0.0376	-0.0760	0.1640	1.5	2.0
1	-0.0582	-0.0512	-0.0598	-0.0726	-0.1322	0.2491	4.0	3.0
13	-0.0473	-0.0366	-0.0445	-0.0552	-0.0981	0.2063	2.0	1.5

When comparing the safety between the participants, we see some variations. Our approach achieves the highest drop in the *C@1* performance measure and the second highest reduction of the *AUC* measure. However, the combination of those two performance statistics (fourth column labeled *final*), puts us in the fifth position. The organizers from PAN then further inspected all the approaches and calculated the

accuracy and recall for each obfuscation system. The impact of every author masking approach is further used for a normalized comparison. In all those additional measures, our approach is ranked last, which is understandable due to its underlying simplicity. Interesting to see is that all participants had a strong correlation between any two performance values, except for us. We achieved a good reduction in the PAN verification measures but had significantly lower Obfuscation scores [11].

6 Conclusion

This paper proposes a light technique to solve the author masking problem with focus on soundness and sensibleness. Depending on the writing style in the comparable documents of an author, a feature is either increased or decreased in the masked text.

We achieved good grades in the sensibleness and soundness by a human assessor who was one of the goals of our system. The safety measurements gave unusual results where we achieved great scores in the verification performance and significantly lower scores in the obfuscation impact. Further evaluation results, including in-depth comparisons with other participants and the official valuation, are available in the overview paper from the organizers [11]. Based on a personal assessment with our author verification system [7], we saw that the safety of the author masking was slightly increased over the baseline, but cases remain where forensic analysis can reveal the original author of its obfuscated texts. This deduction was expected and is according to the reduced set of modifications.

The author obfuscation method has lots of opportunities for improvement. Instead of simply focusing on feature frequencies we could also adjust the average sentence length or changing Boolean features.

Acknowledgments. The author wants to thank the task coordinators for their valuable effort to promote test collections in author identification. This research was supported, in part, by the NSF under Grant #200021_149665/1.

References

1. Afroz, S., Brennan, M., & Greenstadt, R. 2012. Detecting hoaxes, frauds, and deception in writing style online. In: *Proceedings of the 2012 IEEE Symposium on Security and Privacy*. pp. 461–475. Washington, DC, USA.
2. Brennan, M., Afroz, S., & Greenstadt, R. 2012. Adversarial stylometry - Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur*. 15(3), 12:1–12:22.
3. Hürlimann, M., Weck, B., van den Berg, E., Šuster, S., & Nissim, M. 2015. GLAD - Groningen Lightweight Authorship Detection - Notebook for PAN at CLEF 2015. In: Cappellato, L., Ferro, N., Jones, G., & San Juan, E. (eds.) *CLEF 2015 Labs Working Notes*, Toulouse, France, September 8-11, Aachen: CEUR.
4. Juola, P. 2012. Detecting stylistic deception. In: *Proceedings of the Workshop on Computational Approaches to Deception Detection*. pp. 91–96. Avignon, France.

5. Juola, P. & Vescovi, D. 2011. Advances in Digital Forensics VII - 7th IFIP WG 11.9 International Conference on Digital Forensic, chap. *Analyzing Stylometric Approaches to Author Obfuscation*. pp. 115–125. Orlando, FL, USA.
6. Kacmarcik, G, & Gamon, M. 2006. Obfuscating Document Stylometry to Preserve Author Anonymity. In: Calzolari, N., Cardie, C., & Isabelle, P. (eds) *ACL Conference on Computational Linguistics*, Sydney, Australia, July 17-21.
7. Kocher, M. & Savoy, J. 2017. A Simple and Efficient Algorithm for Authorship Verification. *Journal of the American Society for Information Science and Technology*, 68(1), 259-269.
8. Mansoorizadeh, M., Rahgooy, T., Aminiyan, M., & Eskandari, M. 2016. Author Obfuscation using WordNet and Language Models. In: Balog, K., Cappellato, L., Ferro, N., & Macdonald, C. (eds.) *CLEF 2016 Labs Working Notes*, Évora, Portugal, September 5-8, Aachen: CEUR.
9. Mihaylova, T., Karadjov, G., Kiprov, Y., Georgiev, G., Koychev, I., & Nakov, P. 2016. SU@PAN'2016 - Author Obfuscation. In Balog, K., Capellato, L., Ferro, N., & Macdonald, C. (Eds), *CLEF 2016 Labs Working Notes*, Évora, Portugal, September 5-8, Aachen: CEUR.
10. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., & Stein, B. 2014. 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.
11. Potthast, M., Schremmer, F., Hagen, M., & Stein, B. 2018. Overview of the Author Obfuscation Task at PAN 2018 - A New Approach to Measuring Safety. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) *Working Notes Papers of the CLEF 2018 Evaluation Labs*. CEUR-WS.org.
12. Quirk, C., Brockett, C., & Dolan, W. 2004. Monolingual machine translation for paraphrase generation. In: *Proceedings of EMNLP 2004*. pp. 142–149. Barcelona, Spain.
13. Stamatatos, E., Rangel, F., Tschuggnall, M., Kestemont, M., Rosso, P., Stein, B., & Potthast, M. 2018. Overview of PAN-2018 - Author Identification, Author Profiling, and Author Obfuscation. In: Bellot, P., Trabelsi, C., Mothe, J., Murtagh, F., Nie, J., Soulier, L., Sanjuan, E., Cappellato, L., Ferro, N. (eds.) *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. 9th International Conference of the CLEF Initiative (CLEF 18). Springer, Berlin Heidelberg New York.
14. Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M. 2017. Overview of the Author Identification Task at PAN 2017: Style Breach Detection and Author Clustering. In *Working Notes Papers of the CLEF 2017 Evaluation Labs*, CEUR-WS.org.

Appendix

Table 3. Synonyms for " very *X* "

very *X*	*Y*	very *X*	*Y*	very *X*	*Y*
accurate	exact	frightening	terrifying	risky	perilous
afraid	fearful, terrified	funny	hilarious	roomy	spacious
angry	furious	glad	overjoyed	rude	vulgar
annoying	exasperating	good	excellent, superb	sad	sorrowful
bad	atrocious, awful	great	terrific	scared	petrified
beautiful	exquisite	happy	ecstatic, jubilant	scary	chilling
big	immense, massive	hard	difficult	serious	grave, solemn
boring	dull	hard-to-find	rare	sharp	keen
bright	dazzling, luminous	heavy	leaden	shiny	gleaming
busy	swamped	high	soaring	short	brief
calm	serene	hot	scalding, sweltering	shy	timid
careful	cautious	huge	colossal	simple	basic
capable	accomplished	hungry	ravenous, starving	skinny	skeletal
cheap	stingy	hurt	battered	slow	sluggish
clean	spotless	important	crucial	small	petite, tiny
clear	obvious	intelligent	brilliant	smart	intelligent
clever	brilliant	interesting	captivating	smelly	pungent
cold	freezing	large	colossal, huge	smooth	sleek
colorful	vibrant	lazy	indolent	soft	downy
competitive	cutthroat	little	tiny	sorry	apologetic
complete	comprehensive	lively	vivacious	special	exceptional
confused	perplexed	long	extensive	strong	forceful, unyielding
conventional	conservative	long-term	enduring	stupid	idiotic
creative	innovative	loose	slack	sure	certain
crowded	bustling	loud	thunderous	sweet	thoughtful
cute	adorable	loved	adored	talented	gifted
dangerous	perilous	mean	cruel	tall	towering
dear	cherished	messy	slovenly	tasty	delicious
deep	profound	neat	immaculate	thin	gaunt
depressed	despondent	necessary	essential	thirsty	parched
detailed	meticulous	nervous	apprehensive	tight	constricting
different	disparate	nice	kind, lovely	tiny	minuscule
difficult	arduous	noisy	deafening	tired	exhausted
dirty	filthy, squalid	often	frequently	ugly	hideous
dry	arid, parched	old	ancient	unhappy	miserable
dull	tedious	old-fashioned	archaic	upset	distraught
eager	keen	open	transparent	valuable	precious
easy	effortless	painful	excruciating	warm	hot
empty	desolate	pale	ashen	weak	feeble, frail
excited	thrilled	perfect	flawless	well-to-do	wealthy
exciting	exhilarating	poor	destitute	wet	soaked
expensive	costly	powerful	compelling	wicked	villainous
fancy	lavish	pretty	beautiful	wide	expansive
fast	quick, swift	quick	rapid	wiling	eager
fat	obese	quiet	hushed, silent	windy	blustery
fierce	ferocious	rainy	pouring	wise	sagacious, sage
friendly	amiable	rich	wealthy	worried	anxious, distressed
frightened	alarmed				