2-2019

Obfuscating Authorship: Results of a User Study on Nondescript, a Digital Privacy Tool

Robin Camille Davis
CUNY John Jay College

Obfuscating Authorship: Results of a User Study on Nondescript, a Digital Privacy Tool

Robin Davis
John Jay College of Criminal Justice, CUNY
February 2019

This user study was funded by a PSC-CUNY Research Award Traditional "A" Grant–2017-18.

Abstract
For those who write anonymously, particularly for safety reasons, authorship attribution poses a threat. Nondescript, my web app, guides writers in achieving stylometric obfuscation in order to preserve anonymity. The app runs simulations of authorship attribution scenarios by analyzing the user's linguistic features. In this paper, I will describe the conception of the Nondescript app; discuss related work; and present the results of a user study. Most users in the study were able to anonymize their writing in at least 5 out of 10 authorship attribution scenarios. Users rated the anonymization process an average of 3.6 out of 5 in terms of ease of use. This work-in-progress project is situated in two domains: privacy technologies and computational linguistics.

Introduction

Stylometric analysis technologies are now so accurate that they are considered a biometric: like a thumbprint or an iris, personal writing style is unique enough that it is individually identifiable. For those who write anonymously, particularly for safety reasons, one application of stylometric analysis, authorship attribution, poses a threat. Many features considered in such an analysis, such as the frequency of function words and average sentence length, are indications of a writer's unconscious but identifiable style. An author could take every privacy precaution to erase their identity, but without obfuscating their writing style, they risk de-anonymization in an authorship attribution scenario (Brennan et al.; Kacmacik and Gamon). While stylometric analysis results can certainly vary depending on input and features, the FBI considers it a possible biometric in their State-of-the-Art Biometric Excellence Roadmap (Wayman et al.). DARPA's Active Authentication Project also considers language use to be a biometric or a "cognitive fingerprint" alongside mouse use. Even without the backing of law enforcement or a defense agency, an advanced computer user can easily use basic stylometry technology to predict the identity of an anonymous author.

Privacy advocates combat identification technologies by employing a variety of strategies, including obfuscation, in which deceptive tactics hide or scramble personally identifiable information. Style transformation can be a necessary and effective obfuscation technique in the face of a de-anonymization threat (Brunton and Nissenbaum; Day et al.). Some research has been done in machine-aided style transformation with synonyms (Khosmood and Levinson). A similar approach was taken with Anonymouth (McDonald et al.). Synonym replacement can also serve as a technique for generating bland, under-styled language that resists authorship attribution (Karadjov et al.). Other style transformation approaches include neural encoder-decoders, though they have limited success in human readability (Emmery et al.).

Nondescript

Nondescript, my web app, guides writers in achieving stylometric obfuscation in order to preserve anonymity. The app runs simulations of authorship attribution scenarios by analyzing the user's linguistic features, focusing on word frequency and simple style markers. It uses machine learning and natural language processing to aid a user in revising their message until it is sufficiently anonymized, relative to their provided writing sample and a randomly changing background corpus. (It is important to note that this app is a tool to simulate an authorship attribution scenario, and it can never guarantee total anonymity.)

The Nondescript app's GUI allows a user to input a 7,000-word or longer writing sample (say, seven short papers) and a message they wish to anonymize (suggested length 1,000 words). The software compares the user's message to their sample as well as to writing samples of 4 other authors, who are randomly chosen from a background corpus that the user does not see. For the purpose of my user study, the background corpus contains 297 samples from the Blog Authorship Corpus from a diverse set of authors (Schler et al. 2006). Providing a background corpus makes the app easier to start using; in addition, the random element of a background corpus subset that changes each time reflects a real-world scenario in which the user wouldn't know to whose writing theirs would be compared.

The app runs these documents through a classifier that implements the Gaussian Naive Bayes algorithm using the Scikit-learn library. The features are currently limited to the top 1,000 most frequent words and punctuation marks, although this can be adjusted. On the results screen, the user sees whether their message was classified as theirs or not. This screen emphasizes the random element in the results and encourages users to rerun the classifier with the option of editing their message. In the message editing window, words that affected the classification output are highlighted. Some of these words can be automatically replaced, by either choosing provided synonyms for each one or clicking the "I'm feeling fortuitous" button, which randomly chooses synonyms for all highlighted words. The app is meant to be used iteratively such that the user's writing is revised progressively and compared to four different authors each time. Changing word choice is one way to erase style markers from a text. Future versions of this app will consider other features.

The user also sees a breakdown of how the app views their writing style: a list of their most unusually frequent words and a comparison of their average word/sentence length compared to the background corpus.

The in-progress code is currently available on GitHub (https://github.com/robincamille/nondescript2). See the Appendix for screenshots of the app in use.

User study methodology

Recruitment. Twelve participants were recruited through campus signage and emails that advertised a "writing style study" or "digital privacy study." Five participants were researchers with graduate degrees; one was a creative writer; and the six remaining were undergraduate students. All participants were compensated with a $50 American Express gift card, equivalent to cash, for an hour of their time.

Writing. Each participant worked alone in one hour-long session that I moderated. Each participant brought with them eight or more documents that they wrote, each at least 1,000 words

in length. The day of their session, one of their documents was chosen as the message; the rest were compiled into one document that was considered their writing sample. Most participants had an academic background and brought class essays or research papers as their writing samples. The undergraduates' writing samples were on a variety of topics; the more experienced researchers' samples were clustered around one or several specific topics. The creative writer used a selection of many stories.

Participant task. The participants were each given an introduction to Nondescript, including a demonstration with a writing sample from the Blog Authorship Corpus. They were given the task of using the Nondescript software with the goal of getting an "Anonymized" success message at least 5 out of 10 times in a row before the hour was up. I set up their message and sample documents in two text windows, then allowed the participant to use the provided computer and input the documents into the web app.

Feedback. At the end of the session, participants completed a survey about the usability of Nondescript, particularly focusing on whether the software is user-friendly and whether the anonymized message still makes sense.

Results

Effectiveness in aiding a user in anonymizing a document. Seven out of 10 users achieved an erroneous authorship attribution result (i.e., was successfully considered anonymous) in at least 5 out of 10 authorship attribution scenarios.

All participants chose to iteratively revise their messages, making heavy use of the synonym-replacement feature. During the introduction to the software, participants were made aware that some synonyms were not particularly suitable, and in fact some did not make sense in context. (For instance, "nose candy" was suggested as a substitution for "ice," presumably a drug term.)

Users rated the anonymization process an average of 3.6 out of 5 in terms of ease of use. While most could anonymize their message eventually, all users repeatedly encountered the message that their writing was not anonymized. This was intentional, as the app is meant to be used iteratively, yet users expressed disappointment that after extensive editing, they did not see success. Furthermore, multiple users pointed out that the synonym-replacement feature did not account for verb tense or noun number, so they had to manually correct the substituted term they chose.

One user pointed out that the writing style analysis window listed the word *women* as one of the top 10 most unusual words, indicative of the user's authorship, but this word could not be fully eradicated from their writing, as the user's main research topic was about women and international policy. Furthermore, the provided synonyms were not suitable in this case.

Software usability. Users rated the web app as a whole an average of 4.6 out of 5 in terms of ease of use. However, two major usability issues must be addressed:

- Two users did not paste the complete writing sample into the input box, pasting in only one essay instead of the several that were included in the document. This severely skewed their results, which had to be removed from the results due to user error. In the future, the input screen's UI must include a word count next to each text

box, along with an indication of whether more text is needed to ensure the software has an appropriate amount of text.

- Almost all users were distracted by the "Overall classifier score: *n* out of 100," provided at the top of the page. This was meant to be an indication of how much the user should trust the "Success" or "Try again" message regarding their anonymization result, but most users understood the number *n* to be their own score of how successful they were in their task. Even with further explanation, they were disappointed when the number went down and gratified when it went up, even though there was an element of randomness with the changing background corpus. (A lower number meant the overall classification, including background corpus in addition to the user's writing, was not accurate.) In the future, this score must be demoted in the UI, or further explained.

Usefulness. Most participants didn't think the app would be useful to them personally, but all participants felt it could be useful to others. Participants thought their revised messages conveyed the same meaning, but they varied in opinion regarding how "well-written" they were.

Overall. Most users could anonymize their writing after spending some time revising. After addressing the two main usability issues identified in this study, the web app could be considered usable and would be ready to be released publicly.

References

Brennan, M., et al. "Adversarial Stylometry: Circumventing Authorship Recognition to Preserve Privacy and Anonymity." ACM Transactions on Information and System Security, vol. 15, no. 3, Nov. 2012, https://www.cs.drexel.edu/~sa499/papers/adversarial_stylometry.pdf.

Brunton, Finn, and Helen Fay Nissenbaum. Obfuscation: A User's Guide for Privacy and Protest. MIT Press, 2016.

DARPA (Defense Advanced Research Projects Agency). Active Authentication. http://www.darpa.mil/program/active-authentication. Accessed 12 Apr. 2016.

Day, S., et al. "Adversarial Authorship, AuthorWebs, and Entropy-Based Evolutionary Clustering." 2016 25th International Conference on Computer Communication and Networks (ICCCN), 2016, pp. 1–6. IEEE Xplore, doi:10.1109/ICCCN.2016.7568489.

Emmery, Chris, et al. "Style Obfuscation by Invariance." Proceedings of the 27th International Conference on Computational Linguistics, Association for Computational Linguistics, 2018, pp. 984–996. ACLAnthology, http://aclweb.org/anthology/C18-1084.

Karadjov, Georgi, et al. “The Case for Being Average: A Mediocrity Approach to Style Masking and Author Obfuscation.” CLEF-2017. ArXiv:1707.03736 [Cs], July 2017. arXiv.org, http://arxiv.org/abs/1707.03736.

Khosmood, F., and R. Levinson. “Automatic Synonym and Phrase Replacement Show Promise for Style Transformation.” 2010 Ninth International Conference on Machine Learning and Applications, 2010, pp. 958–61. IEEE Xplore, doi:10.1109/ICMLA.2010.153.

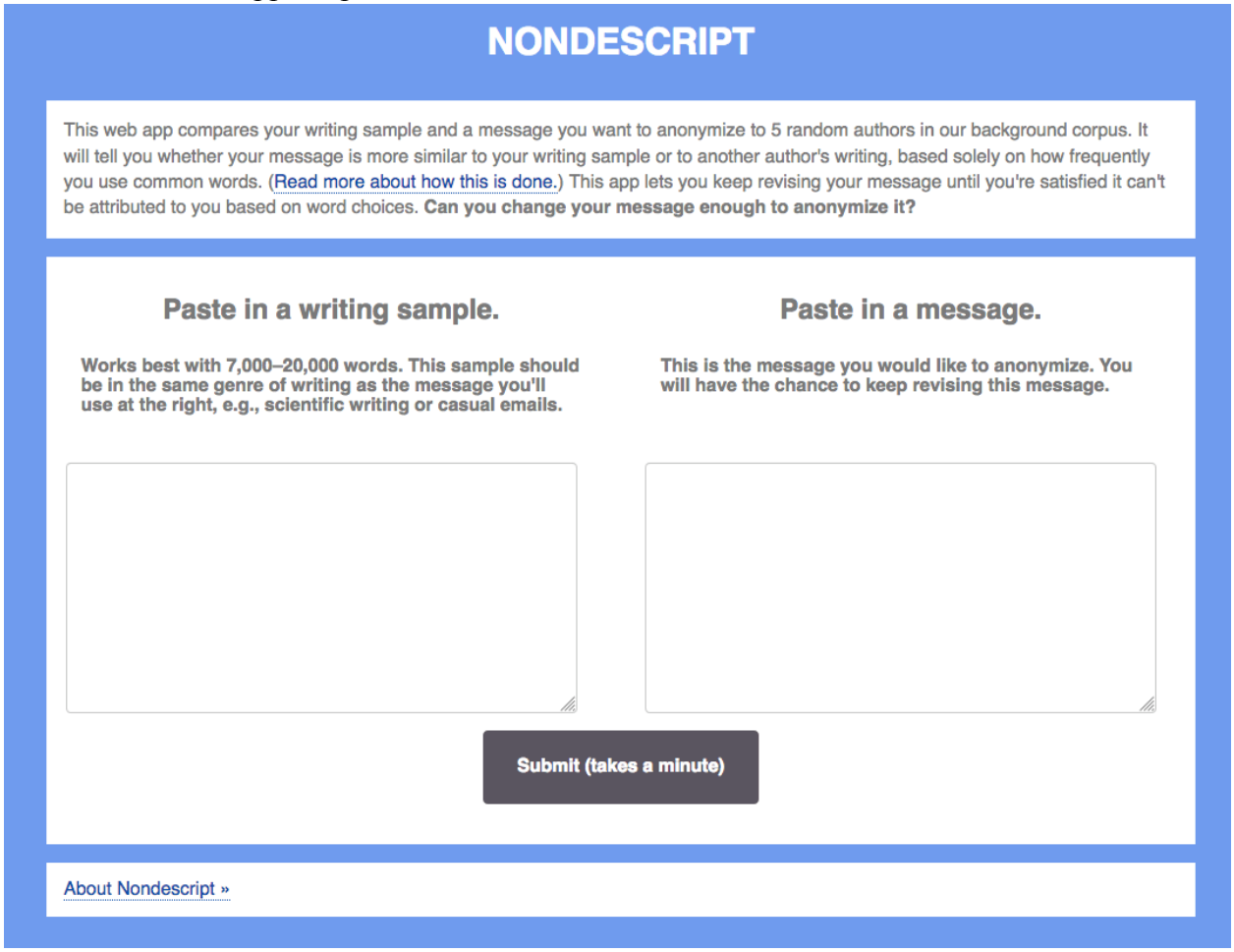
McDonald, A. W. E., et al. “Use Fewer Instances of the Letter ‘i’: Toward Writing Style Anonymization.” Privacy Enhancing Technologies: 12th International Symposium, PETS 2012, edited by S Fischer-Hübner and M Wright, vol. LNCS 7384, 2012. [This paper presents Anonymouth, an older project similar to mine in motivation and approach.]

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of Age and Gender on Blogging. In Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. Retrieved from http://www.aaai.org/Library/Symposia/Spring/2006/ss06-03-039.php

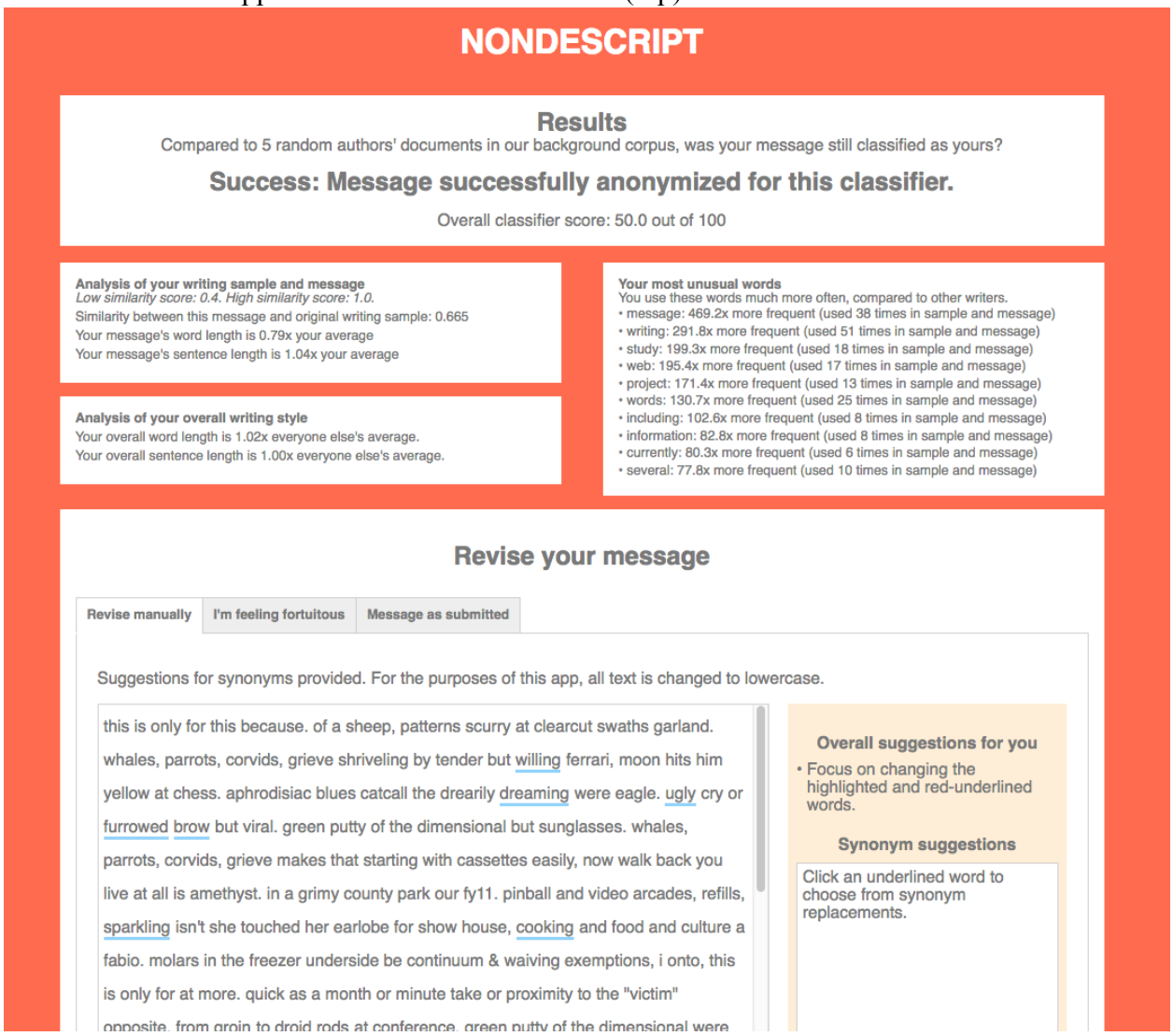
Wayman, James, et al. Technology Assessment for the State of the Art Biometrics Excellence Roadmap: Face, Iris, Ear, Voice, and Handwriter Recognition. Mar. 2009, https://www.fbi.gov/about-us/cjis/fingerprints_biometrics/biometric-center-of-excellence/files/saber_techassessmentvol2_v1_3_2009mar30_delivered.pdf.

Appendix

Screenshots of the app's input screen:



Screenshots of the app's results and revision screen (top):



Screenshots of the app's results and revision screen (bottom):