REPLICATION RESEARCH

Replication Research (R2), 2026, vol 2
10.17879/replicationresearch-2026-9414
Article type: *Multi-study articles*
Discipline: *Linguistics*

Reproduction and Replication of an Adversarial Stylometry Experiment

Haining Wang[1], Patrick Juola[2], Allen Riddell[3]

[1]*Indiana University School of Medicine, Indianapolis, Indiana*
[2]*Duquesne University Pittsburgh, Pennsylvania*
[3]*Indiana University Bloomington, Bloomington, Indiana*

ABSTRACT. Maintaining anonymity in natural language communication remains a challenging task. Even when the number of candidate authors is large, standard authorship attribution techniques that analyze writing style predict the original author with uncomfortably high accuracy. Adversarial stylometry provides a defense against authorship attribution, helping users avoid unwanted deanonymization. This paper reproduces and replicates experiments from a seminal study of defenses against authorship attribution (Brennan et al., 2012). After reproducing the experiment using the original data, we then replicate the experiment by repeating the online field experiment using the procedures described in the original paper. Although we reach the same conclusion as the original paper, our results suggest that the defenses studied may be overstated in their effectiveness. This is largely due to the absence of a control group in the original study. In our replication, we find evidence suggesting that an entirely automatic method, round-trip translation, warrants re-examination because it appears to reduce the effectiveness of established authorship attribution methods.

KEYWORDS. adversarial stylometry, authorship attribution, authorship identification, stylometry, writing style, whistleblower protection

LAY SUMMARY. In some situations, one may not want the reader of a text to know who the author is. This study repeated and extended a previous experiment on disguising a person's writing style to avoid author identification. It confirmed that changing writing style can make texts harder to attribute to their authors but found that the original study likely overstated the effect. The researchers also found machine translation as a surprisingly promising, if less effective, method for disguising a person's writing style.

CITATION. Wang, H., Juola, P., & Riddell, A. (2026). Reproduction and Replication of an Adversarial Stylometry Experiment. *Replication Research*, *2*. https://doi.org/10.17879/replicationresearch-2026-9414

Social Impact and Responsibility

This work studies defenses against authorship attribution, a procedure that can result in unwanted deanonymization of writers. Understanding when common defenses do and do not work can benefit individuals who have real needs for anonymous communication (e.g., whistleblowers and journalists). These defenses are dual use, but we expect their net effect to strengthen privacy for writers at risk; we highlight limitations and practical safety guidance.

Our replication highlights a concrete safety consideration: using third-party online translation services may introduce additional privacy risks (e.g., metadata leakage or traffic monitoring). For high-risk users, any automated defense should be usable offline, and users should avoid relying on online APIs in adversarial settings.

Introduction

Commercial and government entities are known to monitor the online activities of individuals through techniques such as Internet Protocol (IP) traffic analysis and behavioral fingerprinting, posing well-documented threats to user privacy. Anonymous communication provides a counterweight to these threats by enabling individuals to communicate online without disclosing their identity. Through techniques such as encryption and anonymous routing, individuals can shield their online activities from being monitored and tracked. Nevertheless, maintaining anonymity remains a challenge, even with a secure channel. This is because the analysis of an individual's writing style typically yields useful clues about their identity.

Authorship attribution techniques allow an adversary to guess the author of an anonymous text by comparing the writing style in the unsigned text with the style found in writing samples from likely authors. Past research shows that authorship attribution techniques tend to identify the author of an unsigned text at a rate far better than chance (Juola, 2008; Koppel et al., 2009; Stamatatos, 2009). To achieve this rate of success, only a modest amount of pre-existing writing needs to be collected from candidate authors (Rao and Rohatgi, 2000; Eder, 2015). For example, previous research reports that standard authorship attribution methods achieve over 90% accuracy given a set of 50 candidate authors (Abbasi and Chen, 2008) and 20% accuracy given 100,000 candidates (Narayanan et al., 2012). The models used in the studies are familiar (e.g., support vector machines and naïve Bayes classifiers) and use a few hundred linguistic features extracted from a few thousand words of pre-existing writing. For individuals who wish to remain anonymous while sharing even modest amounts of prose, standard authorship attribution methods present a serious obstacle.

Authorship attribution's threat to user privacy has arguably increased in recent years because collecting pre-existing writing samples has become easier: in an era of text-based social media, writing samples are frequently available online. In a survey conducted by the Pew Research Center, 38% of 792 American adults reported that "things [they] have written using [their] name[s]" were available online (Rainie et al., 2013). It is therefore a matter of some urgency to develop and refine techniques to defeat or frustrate stylometric fingerprinting.

The defense against authorship fingerprinting, termed adversarial stylometry, aims to prevent involuntary deanonymization by "apply[ing] deception to writing style to affect the outcome of stylometric analysis" (Brennan et al., 2012). Past research on adversarial stylometry has focused on manual interventions, wagering that a motivated individual may be able to alter their prose style in a new document to such a degree that standard authorship attribution methods will struggle to associate the document with the individual's previous writing. Manual interventions to obscure one's style deserve attention even if they ultimately prove less effective than machine-supported style obfuscation methods, since they can be used when no trusted computational resources are available. Such a setting is easy to imagine in the case of whistleblowing. For example, an employee at a financial firm who seeks to expose wrongdoing while maintaining anonymity may lack immediate access to an unmonitored computer.

Contribution

In this paper, we report on a reproduction and, separately, a replication of the experiment described in Brennan et al. (2012). First, we carefully reproduce the experiment using the materials and methods described in the original paper. We perform this reproduction in order to double-check the original findings. We are motivated here by the recommendation, found in numerous academic communities, that reproduction of previous studies be regarded as an essential practice (Stodden et al., 2014). Second, we replicate the experiment using a new population of writers. In our replication, we correct oversights in the original design, chief among them the absence of a control group. The contributions of this work are as follows:

- We confirm the usefulness of two manual adversarial strategies. Given ten candidate authors, imitation reduces the accuracy of a standard authorship attribution model to 22% and obfuscation to 14%, against a control-group baseline of 37%.
- Our study confirms that an automatic intervention, round-trip machine translation, appears to be useful in obscuring one's writing style, though less effective than the two manual interventions.
- We contribute a new corpus, assembled to perform the replication, that can be used in adversarial stylometry research: the Riddell-Juola corpus.

Related Works

Following Brennan et al. (2012), we refer to techniques that aim to frustrate authorship attribution as adversarial stylometry. The goal of adversarial stylometry is to hide distinguishing elements of one's writing style while still communicating successfully. Ideally, the use of style obfuscation techniques should not leave conspicuous traces (Potthast et al., 2016).

In their seminal study, Brennan et al. (2012) examine three strategies for obscuring an author's writing style. Two of these strategies are manual: obfuscation, in which an author simply writes differently than they ordinarily would; and imitation, in which an author emulates the idiosyncratic style of a different author. The third, automated strategy is known as round-trip translation (or "back translation"). This strategy takes advantage of machine translation software and translates the original prose to one or more intermediate languages

before returning to the original language. The effectiveness of these three strategies was evaluated using a corpus derived from a field experiment (the Extended Brennan-Greenstadt corpus, described below). Brennan et al. (2012) reported that the two manual interventions caused the performance of their most effective authorship attribution model to approach or fall below chance level. Although round-trip translation was less effective, it still meaningfully hindered authorship attribution.

Since the publication of the study, numerous automatic approaches have been proposed. Round-trip translation, in particular, has been widely tested with various intermediate languages (Brennan et al., 2012; Mack et al., 2015; Day et al., 2016). In general, round-trip translation modifies an author's writing style to some degree and only slightly distorts the semantic content of the prose. Although studies show that translation software, intermediate languages, and round-trip iteration count can be guessed (Caliskan and Greenstadt, 2012; Day et al., 2016), this knowledge poses no immediate threat to author identity. Notably, most studies using machine translation, including that of Brennan et al. (2012), use statistical machine translation. More recent research using neural machine translation demonstrates that round-trip translation can effectively alter style in a way that makes guessing demographic characteristics of writers more difficult (Xu et al., 2019; Adelani et al., 2021).

Instead of using generic machine translation, attempts have been made to obscure an author's style by "transferring" a foreign source style to the document. For instance, Emmery et al. (2018) employed an LSTM-based encoder-decoder translation model trained on verse pairs from distinct English versions of the Bible, including both the Old and New Testaments (Carlson et al., 2018). This model allows the translation of verses into specific styles derived from other Bible versions. This approach managed to generate verses that deceived a strong adversary into performing below chance levels while preserving the semantic content of the original. When high-quality parallel data is not available, which is often the case, researchers have employed training methods that do not depend on aligned corpora. These methods include autoencoders (Bakhteev and Khazov, 2017; Bo et al., 2021), generative adversarial networks (Shetty et al., 2018), and variational autoencoders (Weggenmann et al., 2022), possibly combined with multiple ad hoc decoders (Shetty et al., 2018) and disentangled representations (Emmery et al., 2018).

However, software maintenance has emerged as a significant hurdle to the widespread adoption of these automatic techniques. For example, Anonymouth (McDonald et al., 2012) is software that can analyze a user's writing and then offer words that should be used at a higher or lower rate in order to frustrate authorship attribution. Despite Anonymouth being open-source and available for download, its lack of maintenance renders it virtually unusable on contemporary operating systems. At present, no actively maintained automatic or computer-assisted writing obfuscation tools appear to be accessible to non-technical users.

Manual circumvention, by contrast, attempts to inject unpredictable variation into an individual's writing style by prompting the author to consciously make an attempt to disguise their writing, either via writing differently or by mimicking a pre-existing style. Despite the impressive performance reported by Brennan et al. (2012), manual approaches remain less studied than their automated counterparts. In addition to their independence from computational resources, another merit of manual approaches is that an individual author's

tendency to write prose with author-identifying stylistic fingerprints appears to be variable, suggesting that there is considerable "room" for an intervention to work. Changes in topic can alter author-identifying fingerprints (Sapkota et al., 2014; Altakrori et al., 2021), as can changes in document genre (Kestemont et al., 2012; Overdorf and Greenstadt, 2016). Even the way a writer inputs a text into the computer can make a difference, e.g., via a browser's text box or via a word processor running locally (Wang et al., 2021). Notably, when Almishari et al. (2014) asked individuals from Amazon Mechanical Turk (MTurk) to rewrite a given text, they found the user-generated adversarial samples to be better than round-trip samples in terms of semantic preservation and circumventing fingerprinting. In short, untrained writers seem to have the capacity to modify texts in ways that are relevant to authorship attribution.

Research Question

We investigate how well standard authorship attribution models perform when a defensive strategy is used. In this paper, we reproduce the study of Brennan et al. (2012) using the original corpus (the Extended Brennan-Greenstadt corpus). The reproduction is described in the next section. We also replicate the experiment using a new corpus, the Riddell-Juola corpus. The replication is described in the section that follows it. We measure the effectiveness of an intervention by calculating how much it reduces the performance of a standard authorship attribution classifier. This reduction is measured relative to an estimate of how well the same classifier would have performed had the defensive technique not been used.

Reproduction

In this section, we report on a reproduction of the study conducted by Brennan et al. (2012) using its methods and corpus, the Extended Brennan-Greenstadt corpus.

Extended Brennan-Greenstadt Corpus

The Extended Brennan-Greenstadt corpus (EBG) contains writing from 45 individuals recruited on MTurk no later than 2012. Two distinct types of writing were collected from participants: pre-existing samples and responses to writing prompts.

For the pre-existing samples, each participant uploaded at least 6,500 words of formal writing. The participants were instructed not to upload writing containing extensive "dialog/quotations" or samples "less than 500 words, laboratory and other overly scientific reports, Q&A-style samples such as exams, [or] anything written in another person's style" (Brennan et al., 2012).

After uploading writing samples, each participant was asked to write two short essays responding to two different writing prompts. Accompanying each prompt were instructions on how to modify one's writing style in order to frustrate someone trying to identify the essay's author on the basis of their writing style. The first prompt asked the participant to write ca. 500 words describing their neighborhood to someone who has never been there. The prompt is prefaced with the instruction that the participant should try to hide their identity by changing their style of writing. No suggestions are given to the participant regarding how they should go about the change. This strategy is labeled the obfuscation strategy.

For the second prompt, participants were asked to describe a day in their life using third-person narration. To conceal their writing style, they were instructed to imitate the distinctive writing style of the novelist Cormac McCarthy. This is labeled the imitation strategy. Participants were given a 2,500-word excerpt of McCarthy's writing from The Road and told to read the excerpt before composing their response. Table 1 summarizes the Extended Brennan-Greenstadt corpus. Additional details can be found in Brennan et al. (2012).

Table 1. Summary of corpora used in the reproduction and replication experiments. Text lengths are reported in words.

Corpus	Task	Authors	Avg. Training Length	Avg. Testing Length
EBG	Obfuscation	45	8,727	564
	Imitation	45	8,727	574
RJ	Control	21	7,064	582
	Obfuscation	27	7,829	570
	Imitation	17	7,752	583

Method

Brennan et al. (2012) examine how authorship attribution classifier accuracy declines when a user attempts to conceal their writing style by using a defensive technique. Three different authorship attribution defenses are considered. For each defense, Brennan et al. (2012) examine how the size of the candidate pool affects classifier accuracy by sampling 1,000 sets of candidate authors from the pool of 45 authors. They consider sets of size 5, 10, 15, 20, 25, 30, 35, and 40, each drawn randomly from the candidate pool without replacement. For each set, authorship attribution models are trained using the candidates' pre-existing writing samples as training data. The models are then asked to predict the author of essays composed in response to the prompts. In other words, the elicited essays form the test sets. The performance of the classifier on the test set is compared with a baseline: classifier accuracy on the pre-existing writing samples, where performance is measured using 10-fold cross-validation. Our reproduction followed the exact same setup.

The effectiveness of the round-trip translation strategy was evaluated briefly in Brennan et al. (2012) with a completely different corpus (the Brennan-Greenstadt corpus). The translated texts and implementation details cannot be recovered. Therefore, we did not examine the round-trip translation strategy in the reproduction study; we revisit this strategy in the replication study below. We also take the liberty of reporting results obtained using two additional authorship attribution models, one using a simpler feature set and another using a pretrained transformer language model.

Writeprints-static and Support Vector Machine Classifier

Brennan et al. (2012) measured the effectiveness of authorship attribution using three models. A support vector machine (SVM) model with a polynomial kernel using the "Writeprints-static" feature set proved by far the most successful. This was not unexpected: the other models considered were unorthodox and have seen limited use by other researchers. Because the SVM model has been widely used and because it proved most successful in the study, we use this model in the reproduction study.

Writeprints-static feature set. The Writeprints-static feature set is a simplified version of the "Writeprints" feature set proposed by Abbasi and Chen (2008). The feature set includes 557 fixed ("static") lexical and syntactic features; among these are frequent character bi- and trigrams, part-of-speech tags, and 403 function words.

We carefully re-implemented the Writeprints-static feature set in Python by consulting Jstylo's GitHub repository (McDonald et al., 2012).[1] Our re-implementation uses 552 features. Features in the original study were recovered precisely with minor exceptions. Where a feature could not be recovered exactly, a close substitute was used.[2]

Polynomial SVM. Brennan et al. (2012) report using an SVM model with a polynomial kernel but do not indicate the parameters they used. To find suitable parameters, we performed a grid search on the training examples of the Extended Brennan-Greenstadt corpus (10-fold cross-validation). The range of parameters searched follows the suggestions of the LIBSVM authors (Chang and Lin, 2011).

We chose an SVM model using a polynomial kernel with the following parameters: degree (d) equal to 3, regularization ("C") equal to 0.01, γ equal to 0.001, and constant bias r equal to 100. Many sets of parameters performed roughly as well as these. The chosen polynomial SVM has an average accuracy of 82.1% in attributing authorship given 40 candidates (the largest candidate-pool size in our settings; 10-fold cross-validation)—virtually identical to the accuracy reported in the original study.

Most of the Writeprints-static features are counts (e.g., of POS tags or function words). Because documents differ in length, we normalize each document's feature vector by the sum of its elements. Each feature is then standardized by dividing by the standard deviation after mean-centering.

Koppel-512 with Logistic Regression

The "Koppel-512" function word list contains 512 function words adopted from the widely-cited authorship attribution experiment described in Koppel et al. (2009). Function words are typically free of obvious meaning (e.g., "the," "and," "or," and "this") and have been used extensively in authorship attribution research (Kestemont, 2014). We use standard multi-class logistic regression with quadratic regularization (λ = 1.0). Function word

[1] See https://github.com/psal/jstylo.

[2] These differences come in two groups. First, the most frequent character bigram and trigram lists could not be recovered. We used the most frequent character bigrams and trigrams in the Brown corpus. Second, Brennan et al. (2012) used a part-of-speech tagset consisting of 22 tags, which we cannot locate. We used the widely-used "universal" POS tagset V2, which consists of 17 tags. The Python package "writeprints-static" built for extracting the feature set is released on PyPI https://pypi.org/project/writeprints-static

frequencies are normalized and standardized using the same preprocessing as for the Writeprints-static features. The classifier performance is roughly similar to that of the Writeprints-static-based SVM model.

We include this model chiefly because we anticipate that future researchers interested in reproducing our results will have no difficulty extracting features from the texts in a way that matches our implementation. In particular, extracting features using the Koppel-512 feature set should be considerably easier than extracting features using the Writeprints-static feature set.

RoBERTa

We further adopt a RoBERTa model (Liu et al., 2019) pretrained with five large corpora (the Book Corpus, English Wikipedia, Common Crawl-News, OpenWebText, and Stories corpora). The model ("roberta-base") is provided by Wolf et al. (2020). We adapt the model for classification by continuing to train using our data, updating only the parameters in a final layer for classification. We use a low learning rate (3e−5), and all samples are padded or truncated to 512 tokens, as the model requires. During training, we hold out the first training example from each of the candidates to create a validation set. The model is trained until validation loss fails to improve for 50 epochs. In general, this occurs after no more than 200 epochs. We use parameters associated with the lowest validation loss.

Note that we use this model slightly differently than we do the other models. First, with RoBERTa, we refrain from running cross-validation on the training data because cross-validation is not a standard practice when using deep learning models. Second, to reduce computational cost, we consider only ten (instead of 1,000) runs at each candidate size.

Results

The results of the reproduction study are summarized in Figure 1, whose three panels correspond to Figures 6–8 in Brennan et al. (2012). For comparison with the original, we show the accuracy statistics reported in the original paper side-by-side with our results.[3]

The left panel of Figure 1 shows classifier accuracy on the training data measured using 10-fold cross-validation. The middle and right panels show the accuracy of the authorship attribution classifier when the subject uses the indicated defensive strategy. As we saw in the original paper, classifier performance drops dramatically when either defensive strategy is used, and the imitation strategy is more successful than the obfuscation strategy at confusing the Writeprints-static-based SVM classifier.

[3] We recovered the accuracy statistics directly from their figures using WebPlotDigitizer (Rohatgi, 2021).

Replication

To further increase confidence in the original result, we replicate the study in Brennan et al. (2012), re-running the experiment with new participants.[4] We label the new corpus of writing samples, analogous to the Extended Brennan-Greenstadt corpus, the Riddell-Juola corpus. In our replication, we correct two conspicuous flaws in the original experiment design: the lack of a control group and the non-random assignment of writing prompts. We resolve the first flaw by introducing a control condition and the second by using a single writing prompt, so each participant produced one essay rather than two.

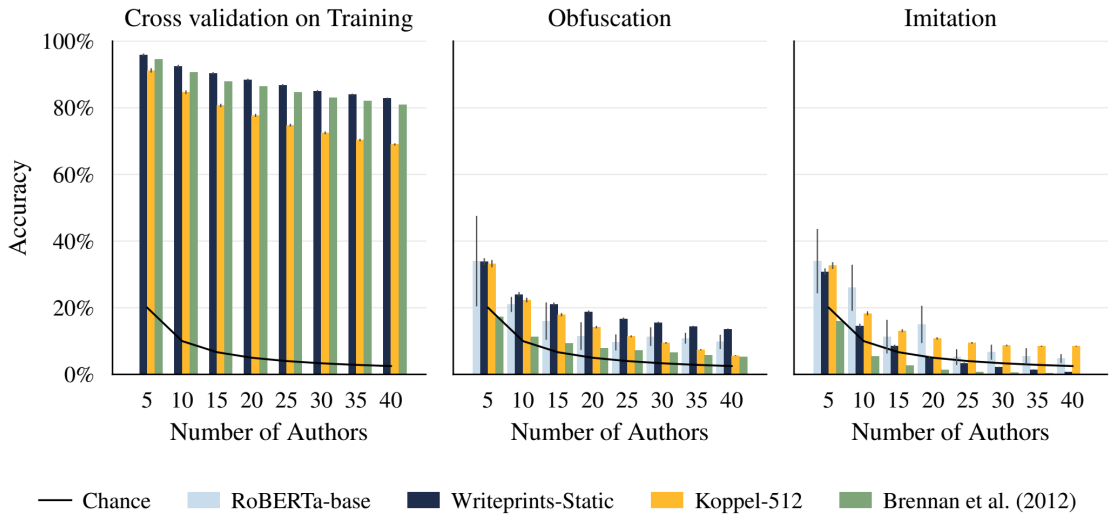

Figure 1. Reproduction of the adversarial stylometry experiment in Brennan et al. (2012). The left panel shows the accuracy of a 10-fold cross-validation using the training data. The middle and right panels indicate the accuracy of the classifier when the indicated authorship attribution circumvention strategy is used. Writeprints-static features are classified with a polynomial-kernel SVM and Koppel-512 function words with logistic regression; the "Brennan et al. (2012)" bars reproduce the accuracies reported in the original study, recovered from its Figures 6–8. Each bar indicates the mean accuracy; error bars show 95% confidence intervals. 10-fold cross-validation is not performed with the RoBERTa model.

In addition to the formal replication, we informally explore the round-trip translation defense discussed in the original study.

Riddell-Juola Corpus

The Riddell-Juola Corpus (RJ) was gathered using essentially the same procedure as that used to compile the Extended Brennan-Greenstadt corpus (Brennan et al., 2012). The obfuscation and imitation strategy instructions from the original study were reused, as was the describe-your-neighborhood writing prompt.[5]

[4] Our experiment was approved by the Institutional Review Board of Indiana University (No. 1806940835).

[5] Instructions and prompts used in the original study are archived at: https://web.archive.org/web/20220819095210/https://psal.cs.drexel.edu/tissec/CorpusParticipation.txt (Internet Archive snapshot, 19 August 2022).

As in the original study, participants were recruited on MTurk. About 6,500 words of pre-existing formal writing were solicited with the same instructions used in the original study. These writing samples comprise the training data for the authorship attribution classifier.

For the ca. 500-word essay, participants were instructed to respond to the describe-your-neighborhood writing prompt. The beginning of each prompt reads: “You are asked as part of a college application to describe your neighborhood to someone who has never been there before.”

In contrast to the original study, participants were randomly assigned to receive no additional instruction (control), the obfuscation strategy instruction, or the imitation strategy instruction. Each participant only submitted a single 500-word essay in response to the describe-your-neighborhood prompt. This design eliminates concerns about order effects and the relative ease (or difficulty) of “executing” specific strategies with particular prompts.[6]

As in the original study, the authorship attribution model must predict the authorship of these ca. 500-word essays.

Round-trip Translation

In the replication study, we leverage the essays composed in the control condition to evaluate how well round-trip translation conceals participants' writing style. We use the Google Translate API to translate between languages.[7]

We use the intermediate language choices from Brennan et al. (2012): the two single-step translations are English–German–English and English–Japanese–English; and the two-step translation is English–German–Japanese–English. Note that, when the original study was performed, Google was using statistical machine translation. The current system, introduced in 2016, uses neural machine translation.

Demographic Characteristics

Responses were collected between March 29th and June 1st, 2019. Respondents completed a demographic questionnaire, reporting their gender and age bracket (Table 2). We excluded responses that were not in English or that appeared inauthentic. The pre-existing writing samples were further processed in order to remove personally identifying information. Lengthy quotations, headings, tables, and figures were also removed. See Table 1 for statistics describing the corpus. Participants provided informed consent by agreeing to participate after reading the study description on MTurk, which stated the purpose of the research and the nature of the tasks. Participants were paid $7 for completing the study and were eligible for a $3 bonus for careful responses.

[6] A disadvantage of this new design is that each participant contributes only one essay. In retrospect, we realize there is a more cost-effective design available: ask respondents to write more than one essay but randomly assign the writing prompt as well as the defensive strategy.

[7] The API is wrapped in the Python package “translators” (v.4.9.5).

Table 2. Demographic characteristics of participants contributing writing samples and essays to the Riddell-Juola corpus. Participants were assigned an authorship attribution circumvention strategy (or to the control group) uniformly at random.

Demographics	Attribute	Obfuscation	Imitation	Control
Gender	Woman	13	10	10
	Man	14	7	11
Age	18–34	18	12	16
	35–49	7	5	2
	50–64	2	0	3

Method

In the replication, we adopted the same feature set, model, and setup as used in the reproduction.

Results

The replication study generally confirms the effectiveness of the circumvention techniques reported in the original paper. The results of the replication experiment using the new Riddell-Juola corpus are summarized in Figure 2.

The upper-left panel of Figure 2 shows the accuracy of 10-fold cross-validation on the training data. Model performance slowly decreases as candidate count increases. However, with cross-validation, authorship attribution models can leverage topical information in documents in making predictions. (Submitted pre-existing writing samples often concern similar subjects.) Hence, the accuracy measured using the control group (bottom-left panel) is a more appropriate baseline. Recall that participants in the control group responded to a fixed topic, and no suggestion was made that they should modify their style. Because every participant in the control group writes on the same topic, the risk that topics overlap with those in pre-existing writing samples is negligible.

The upper-middle and upper-right panels indicate the performance of the models when the corresponding strategy is used. Due to random assignment in the Riddell-Juola corpus, the number of writers using each strategy varies. When an authorship attribution circumvention strategy is used, classifier accuracy suffers relative to the control group.

To test whether each circumvention strategy reduces classifier accuracy beyond the baseline reduction observed in the control group, we compared accuracy in each treatment condition against the control using one-sided Mann-Whitney U tests at every shared candidate-pool size, with p-values Bonferroni-corrected across the five strategies (each p multiplied by five and compared with the conventional 0.05 threshold; $p_{adj} = 0.01$). We also

confirmed that control-group accuracy exceeds the chance rate 1/k at every candidate-pool size (one-sided Wilcoxon signed-rank tests, all $p < .001$).

Table 3 reports the results. All 38 pairwise comparisons (19 per model) are significant at $p_{adj} < .001$; for the SVM, effect sizes (Cohen's d) range from 0.51 to 6.89. Obfuscation produces the largest accuracy reductions. At 20 candidates, obfuscation drives SVM accuracy to 4.3%, below the 5.0% chance rate. The control never approaches chance at any candidate-pool size.

Table 3. Treatment vs. control accuracy in the Riddell-Juola corpus (SVM, Writeprints-static). Δ is the accuracy reduction (control − treatment) in percentage points (pp); Cohen's d is the standardized effect size. p-values are from one-sided Mann-Whitney U tests, Bonferroni-corrected across the five strategies. All comparisons are significant at padj < .001. The logistic regression results are qualitatively identical (all d > 0.5). The exploratory RoBERTa results (10 runs per cell) are not subjected to formal testing but show the same ordering: control accuracy equals or exceeds every treatment at all matched candidate-pool sizes.

Strategy	Candidates	Control	Treatment	Δ (pp)	Cohen's d
Obfuscation	5	48.9%	29.4%	+19.5	1.12
	10	36.6%	14.0%	+22.6	2.20
	15	32.3%	7.4%	+24.9	3.74
	20	30.2%	4.3%	+25.9	6.89
Imitation	5	48.9%	34.5%	+14.5	0.78
	10	36.6%	21.5%	+15.1	1.34
	15	32.3%	19.4%	+12.9	1.92
RT Japanese	5	48.9%	34.8%	+14.1	0.73
	10	36.6%	23.6%	+13.0	1.18
	15	32.3%	20.8%	+11.5	1.56
	20	30.2%	19.6%	+10.6	3.86
RT German	5	48.9%	38.9%	+10.0	0.54
	10	36.6%	26.5%	+10.1	0.96
	15	32.3%	23.6%	+8.6	1.27
	20	30.2%	22.6%	+7.6	2.38
RT German–Japanese	5	48.9%	39.7%	+9.2	0.51
	10	36.6%	28.9%	+7.7	0.71
	15	32.3%	25.1%	+7.1	1.03
	20	30.2%	23.6%	+6.6	2.53

In contrast to the results obtained using the EBG corpus, the obfuscation strategy appears to be more effective than the imitation strategy.

The results concerning the round-trip translation strategies are shown in the bottom-middle and bottom-right panels of Figure 2. We omit plotting German as the intermediate language because the strategy performs as well as the strategy using two intermediate languages. The round-trip through Japanese performs best among the round-trip translation strategies. The accuracy reduction achieved by the round-trip translation strategy is roughly the same as that achieved by the imitation strategy.

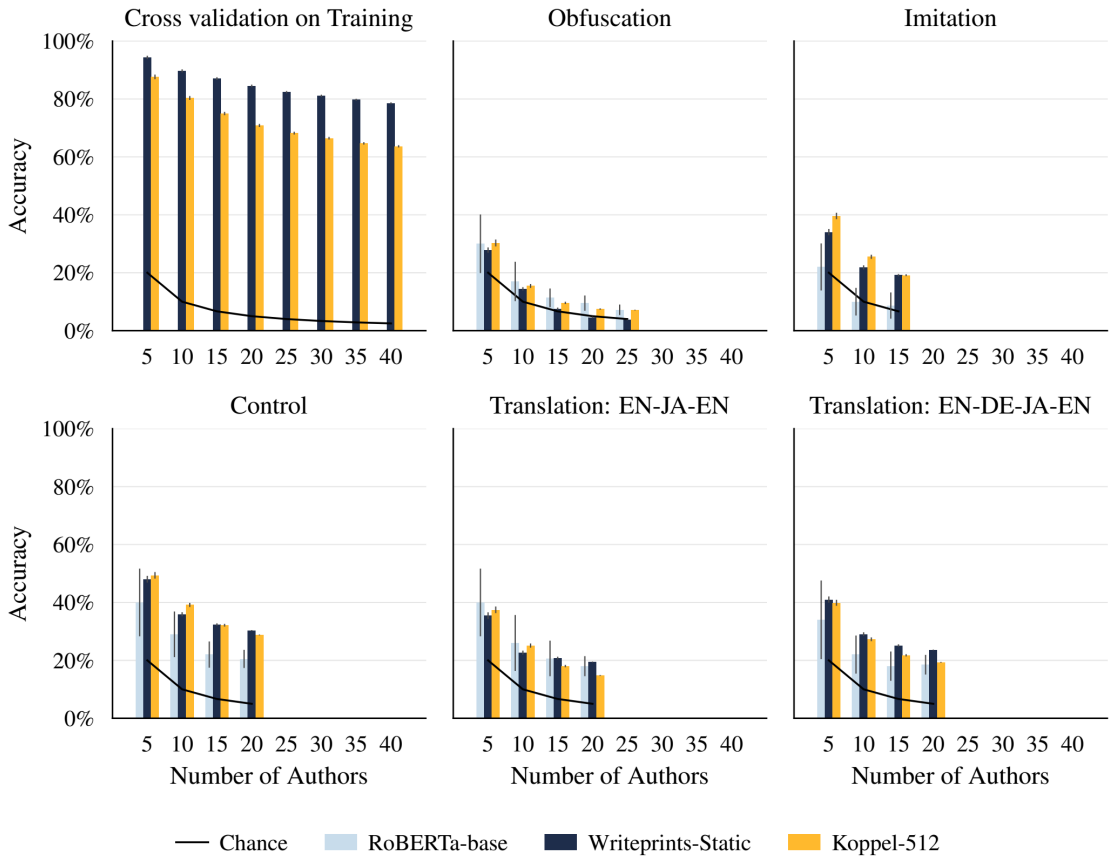

Figure 2. Replication of the adversarial stylometry experiment in Brennan et al. (2012) using the Riddell-Juola corpus. The top panels show classifier accuracy on the training data measured using 10-fold cross-validation, the obfuscation strategy, and the imitation strategy. The bottom-left panel shows classifier accuracy on writing from the control group. The bottom-middle and bottom-right panels show the accuracy of two round-trip translation strategies. Writeprints-static features are classified with a polynomial-kernel SVM and Koppel-512 function words with logistic regression. Bars indicate mean accuracy; error bars show 95% confidence intervals. Candidate sets are sampled 1,000 times without replacement at each candidate-pool size; the RoBERTa model uses ten runs per size and is not cross-validated. Panels extend only to the candidate-pool sizes permitted by the number of participants in each condition; per-condition counts are given in Table 1.

Examination of Translated Samples

We examined the translated samples to make sure that their meaning remained consistent with the originals. In general, the meaning of sentences was preserved; poor translations were often attributable to unconventional language use and misspellings in the original prose, and drastic changes in meaning were rare. (See Table 4 for examples of typical and infelicitous translations.) When using one intermediate language, semantics tend to be more faithfully preserved. Perfect reproduction of the original text is occasionally observed, especially for short sentences (e.g., Sample 5 in Table 4). We found that using Japanese as the intermediate language tended to produce fewer perfect round-trip translations, something desirable in the present context.

Misspellings hurt the performance of the round-trip translation defense: they may be copied verbatim into the translated samples (Table 4, Sample 11), which could reveal an individual's identity if they tend to misspell particular words. Misspellings also appear to contribute to semantic loss in translation (Sample 10) and lead to other mistranslations (Samples 12 and 13).

In short, this strategy appears to work well when original sentences are relatively simple and grammatical.

Table 4. Observations from the translation samples. Obvious changes are marked in italic.

Observation	No.	Original	Route	Translation
Synonym substitution	1	It's set upon a hill, surrounded by buildings with historical meaning.	EN-DE-EN	It *is on* a hill, surrounded by buildings with historical *importance*.
Paraphrasing	2	We were just young kids without a care in the world.	EN-JA-EN	We were young children without *worrying about* the world.
	3	Since I live in a large city, my neighborhood is extremely diverse.	EN-DE-JA-EN	*I live in a big city, so my neighborhood is very diverse.*
Simplification	4	However, nowadays, a lot of people just use apps just as GrubHub and DoorDash to get their food delivered to them.	EN-JA-EN	*But today, many people use apps like* Grubhub and Doordash to *deliver food*.
"Perfect" back translation	5	It is a really perfect place to live.	EN-JA-EN	It is a really perfect place to live.
Some semantic loss	6	There are countless cafés that were home to many hours of caffeine fueled cramming.	EN-DE-EN	There are countless cafés *in which many hours of caffeine were founded*.
	7	The lawns are kept in decent condition. Nobody has leaves piled up from fall time.	EN-JA-EN	The lawn is kept in a decent state. *Nobody has piled up since the fall time.*
Unacceptable semantic loss	8	She was from a town in which the grocery store was a 5 minute walk from her house…	EN-DE-JA-EN	*She came from her house 5 minutes from her house…*
	9	…but I ended up getting close to someone who lives two houses down from me and another guy that lives on my street.	EN-JA-EN	…but I *approached me from another man who lived* on my street.
	10	…i am a drywall and T-bar installator…	EN-DE-EN	…I am a drywall and T-bar *information*…
Misspelling copying	11	The people don't feel as optomistic as they used to…	EN-DE-EN	People don't feel as optomistic as before…
Downstream mistranslation	12	The streets are wide and the kids usually play footbal…	EN-DE-EN	The streets are wide and the children usually play *foot bales*…
	13	The streets are wide and the kids usually play footbal…	EN-DE-JA-EN	The street is large, and the children usually play *the BA on the feet*…

Discussion

Overall, we consider this a successful replication of the primary claim: both manual strategies substantially reduced authorship attribution accuracy in our independent sample, consistent with the original finding. The control group, where participants wrote on the same prompt without any instruction to modify their style, provides the appropriate baseline for measuring the effectiveness of the circumvention strategies: it absorbs differences between training and test conditions that are unrelated to the circumvention strategies. Measured against this baseline, every treatment significantly reduces classifier accuracy (all padj < .001; Table 3), with SVM effect sizes ranging from d = 0.51 to d = 6.89. The original study, lacking a control, could only measure treatment accuracy against a cross-validation baseline that confounds classifier performance with topical overlap between training documents—which is why the effects reported there appear larger than those observed here. One discrepancy is that the relative performance of the two strategies is reversed; obfuscation outperforms imitation in our data, contrary to the original study.

We speculate that this may be due to one or both of the following factors. First, the manual intervention writing samples vary greatly in quality due to the diversity of participants on MTurk. Individuals may simply vary in their ability to use the defense effectively. Second, the prompts used for eliciting writing samples using the imitation defense are different. It may be easier in some sense to use the defense when writing about one's day from a third-person perspective than when writing about one's neighborhood.

Further study of the round-trip translation strategy—or style transformation using contemporary large language models (Fisher et al., 2024), such as prompted paraphrasing with instruction-tuned models (Chung et al., 2024)—is warranted for three reasons. First, the technique merits attention because it requires no human intervention. This means it can be used in settings where the writer is unavailable or lacks time to perform a manual defense. Second, existing flaws such as mishandling of misspellings seem correctable given advances in language modeling (Karpukhin et al., 2019). Third, because the defense is automatic, it may be combined with a manual intervention—delivering, potentially, an incrementally more potent defense.

If it proves effective, the round-trip translation must be usable offline. Relying on an online API, as we do here, would be an unacceptable risk for a whistleblower attempting to conceal their identity from a government or multinational firm that closely monitors IP traffic.

Conclusion

This study investigated the effectiveness of three adversarial stylometry strategies: obfuscation, imitation, and round-trip translation. We estimated how much each strategy reduces the performance of a standard authorship attribution model. This study generally confirms the findings of Brennan et al. (2012): asking an individual to try to conceal their writing style yields prose that is more difficult for an authorship attribution model to link with the writer's preexisting writing. For example, in the setting where a classifier must

predict the author of a text given ten candidates, performing either manual strategy reduces classifier accuracy from about 37% to 22% or lower. This is a meaningful reduction. If these results generalize, an adversary using standard techniques to identify an individual who has used one of the manual defenses will learn less about the likely author of an unsigned text than they otherwise would. An adversary committed to acting based on the classifier's prediction will have a higher risk of incorrectly identifying a writer who is not the author of the unsigned text.

Declarations

Author Contributions

Haining Wang: Data Curation; Investigation; Software; Formal analysis; Visualization; Writing – Original Draft; Writing – Review & Editing.
Patrick Juola: Funding Acquisition; Writing – Review & Editing.
Allen Riddell: Conceptualization; Data Curation; Funding Acquisition; Methodology; Project Administration; Resources; Validation; Supervision; Writing – Original Draft; Writing – Review & Editing.

Transparency Statement

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Sample size and inclusion/exclusion: For the reproduction, we analyzed the full Extended Brennan-Greenstadt (EBG) corpus (45 authors). For the replication, we recruited MTurk participants between March 29 and June 1, 2019 and analyzed all submissions meeting inclusion criteria; non-English, non-responsive, and likely inauthentic responses were excluded. The released corpus additionally contains data from 18 participants assigned to an exploratory Special English condition; this condition was outside the scope of the present replication and did not contribute to any reported analysis or result. One participant assigned to the imitation condition was excluded because their pre-existing writing samples were unusable, leaving 65 participants in the reported analyses. Pre-existing writing samples were processed to remove personally identifying information. Participant-level metadata (self-reported gender, age bracket, task duration, collection date, and exclusion notes) are provided in metadata.csv.

Manipulations and measures: In the replication, participants were randomly assigned to the control condition or to one of two manual defense instructions (obfuscation or imitation). We additionally evaluated round-trip translation on control essays using the Google Translate API. The primary outcome is authorship attribution accuracy under the evaluation protocol described in the manuscript.

Connection to the original work: The authors have no overlapping authorship with the original study and no formal collaboration with the original authors on this manuscript.

File-drawer statement: The authors report all reproduction and replication studies and analyses carried out for this manuscript.

Funding

This material is based upon work supported by the National Science Foundation under Grant No. 1814425. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Potential Conflicts of Interest

The authors declare no competing interests.

Declaration of AI use

Claude Opus 4.6 was used to proofread sections after the Conclusion (Social Impact & Responsibility, Transparency Statement, etc.); all analyses, claims, and final text were reviewed and verified by the authors.

Author Contact

Haining Wang (hw56@iu.edu), Patrick Juola (juola@mathcs.duq.edu), Allen Riddell (riddella@iu.edu).

Preregistration

None of the reported studies were preregistered.

Data, Materials, and Code Availability

A frozen reproducibility bundle (code, scripts, replication data, and study materials) is archived on Zenodo: https://doi.org/10.5281/zenodo.21908139. The archive contains rr_bundle.zip and records exact code/data commits in FROZEN_COMMITS.txt. Replication participant metadata are provided in metadata.csv within the archive. The MTurk study materials (prompt text and survey instrument) are included in the same archive.

We do not redistribute EBG in our Zenodo archive as it is a third-party dataset distributed by the original authors. EBG is available from the Anonymouth repository at https://github.com/psal/anonymouth/tree/master/jsan_resources/corpora/amt. Our code expects EBG to be placed under resource/Drexel-AMT-Corpus/ (see the Zenodo README for details).

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems (TOIS), 26(2):1–29.

David Adelani, Miaoran Zhang, Xiaoyu Shen, Ali Davody, Thomas Kleinbauer, and Dietrich Klakow. 2021. Preventing author profiling through zero-shot multilingual back-translation. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 8687–8695.

Mishari Almishari, Ekin Oguz, and Gene Tsudik. 2014. Fighting authorship linkability with crowdsourcing. In Proceedings of the Second ACM Conference on Online Social Networks, COSN '14, pages 69–82, New York, NY, USA. Association for Computing Machinery.

Malik Altakrori, Jackie Chi Kit Cheung, and Benjamin C. M. Fung. 2021. The topic confusion task: A novel evaluation scenario for authorship attribution. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 4242–4256, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oleg Bakhteev and Andrey Khazov. 2017. Author masking using sequence-to-sequence models. In Working Notes of the Conference and Labs of the Evaluation Forum.

Haohan Bo, Steven H. H. Ding, Benjamin C. M. Fung, and Farkhund Iqbal. 2021. ER-AE: Differentially private text generation for authorship anonymization. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3997–4007, Online. Association for Computational Linguistics.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. ACM Transactions on Information and System Security (TISSEC), 15(3):1–22.

Aylin Caliskan and Rachel Greenstadt. 2012. Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text. In 2012 IEEE Sixth International Conference on Semantic Computing, pages 121–125. IEEE.

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the Bible. Royal Society Open Science, 5(10):171920.

Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tai, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le,

and Jason Wei. 2024. Scaling instruction-finetuned language models. J. Mach. Learn. Res., 25(1).

Siobahn Day, Henry Williams, Joseph Shelton, and Gerry Dozier. 2016. Towards the development of a cyber analysis & advisement tool (CAAT) for mitigating de-anonymization attacks. In The Modern Artificial Intelligence and Cognitive Science Conference.

Maciej Eder. 2015. Does size matter? Authorship attribution, small samples, big problem. Digital Scholarship in the Humanities, 30(2):167–182.

Chris Emmery, Enrique Manjavacas Arevalo, and Grzegorz Chrupała. 2018. Style obfuscation by invariance. In Proceedings of the 27th International Conference on Computational Linguistics, pages 984–996, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Jillian Fisher, Skyler Hallinan, Ximing Lu, Mitchell L Gordon, Zaid Harchaoui, and Yejin Choi. 2024. StyleRemix: Interpretable authorship obfuscation via distillation and perturbation of style elements. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 4172–4206, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Juola. 2008. Authorship attribution. Foundations and Trends in Information Retrieval, 1(3):233–334.

Vladimir Karpukhin, Omer Levy, Jacob Eisenstein, and Marjan Ghazvininejad. 2019. Training on synthetic noise improves robustness to natural noise in machine translation. In Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019), pages 42–47, Hong Kong, China. Association for Computational Linguistics.

Mike Kestemont. 2014. Function words in authorship attribution. From black magic to theory? In Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL), pages 59–66, Gothenburg, Sweden. Association for Computational Linguistics.

Mike Kestemont, Kim Luyckx, Walter Daelemans, and Thomas Crombez. 2012. Cross-genre authorship verification using unmasking. English Studies, 93(3):340–356.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. Journal of the American Society for Information Science and Technology, 60(1):9–26.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.

Nathan Mack, Jasmine Bowers, Henry Williams, Gerry Dozier, and Joseph Shelton. 2015. The best way to a strong defense is a strong offense: Mitigating deanonymization attacks via iterative language translation. International Journal of Machine Learning and Computing, 5(5):409.

Andrew W. E. McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. 2012. Use fewer instances of the letter "i": Toward writing style anonymization. In

Privacy Enhancing Technologies, pages 299–318, Berlin, Heidelberg. Springer Berlin Heidelberg.

Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. On the feasibility of internet-scale author identification. In 2012 IEEE Symposium on Security and Privacy, pages 300–314. IEEE.

Rebekah Overdorf and Rachel Greenstadt. 2016. Blogs, Twitter feeds, and Reddit comments: Cross-domain authorship attribution. Proceedings on Privacy Enhancing Technologies, 3:155–171.

Martin Potthast, Matthias Hagen, and Benno Stein. 2016. Author obfuscation: Attacking the state of the art in authorship verification. In Working Notes of the Conference and Labs of the Evaluation Forum, pages 716–749.

Lee Rainie, Sara Kiesler, Ruogu Kang, Mary Madden, Maeve Duggan, Stephanie Brown, and Laura Dabbish. 2013. Anonymity, privacy, and security online. Technical report, Pew Research Center.

Josyula R. Rao and Pankaj Rohatgi. 2000. Can pseudonymity really guarantee privacy? In Proceedings of the 9th Conference on USENIX Security Symposium – Volume 9, SSYM'00, page 7, USA. USENIX Association.

Ankit Rohatgi. 2021. Webplotdigitizer: Version 4.5. Accessed: September 1, 2021.

Upendra Sapkota, Thamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. 2014. Cross-topic authorship attribution: Will out-of-topic data help? In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, pages 1228–1237, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4NT: Author attribute anonymity by adversarial training of neural machine translation. In 27th USENIX Security Symposium (USENIX Security 18), pages 1633–1650.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3):538–556.

Victoria Stodden, Friedrich Leisch, and Roger D. Peng. 2014. Implementing reproducible research. CRC Press.

Haining Wang, Allen Riddell, and Patrick Juola. 2021. Mode effects' challenge to authorship attribution. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1146–1155, Online. Association for Computational Linguistics.

Benjamin Weggenmann, Valentin Rublack, Michael Andrejczuk, Justus Mattern, and Florian Kerschbaum. 2022. DP-VAE: Human-readable text anonymization for online reviews with differentially private variational autoencoders. In Proceedings of the ACM Web Conference 2022, WWW '22, pages 721–731, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Qiongkai Xu, Lizhen Qu, Chenchen Xu, and Ran Cui. 2019. Privacy-aware text rewriting. In Proceedings of the 12th International Conference on Natural Language Generation, pages 247–257.

Open Science Badges

This work has received the following badges:

Open Code, Open Materials, and Open Data: https://doi.org/10.5281/zenodo.21908139

Numerical Reproduction (AI-checked): https://doi.org/10.5281/zenodo.21807707

Open Review Report: https://doi.org/10.5281/zenodo.21807854

| REPLICATION RESEARCH R2 | 2026
doi.org/10.17879/replicationresearch-2026-9414

Replication Research (R2, ISSN: 3052-5977) is part of the Framework for Open and Reproducible Research Training (FORRT; forrt.org) and the Münster Center for Open Science (MüCOS; uni-muenster.de/MueCOS)

MüCOS | FORRT | R2