A Girl Has A Name, And It's ...* Adversarial Authorship Attribution for Deobfuscation†

Wanyue Zhai
University of California, Davis

Jonathan Rusert
University of Iowa

Zubair Shafiq
University of California, Davis

Padmini Srinivasan
University of Iowa

Abstract

Recent advances in natural language processing have enabled powerful privacy-invasive authorship attribution. To counter authorship attribution, researchers have proposed a variety of rule-based and learning-based text obfuscation approaches. However, existing authorship obfuscation approaches do not consider the adversarial threat model. Specifically, they are not evaluated against adversarially trained authorship attributors that are aware of potential obfuscation. To fill this gap, we investigate the problem of adversarial authorship attribution for deobfuscation. We show that adversarially trained authorship attributors are able to degrade the effectiveness of existing obfuscators from 20-30% to 5-10%. We also evaluate the effectiveness of adversarial training when the attributor makes incorrect assumptions about whether and which obfuscator was used. While there is a a clear degradation in attribution accuracy, it is noteworthy that this degradation is still at or above the attribution accuracy of the attributor that is not adversarially trained at all. Our results underline the need for stronger obfuscation approaches that are resistant to deobfuscation.

1 Introduction

Recent advances in natural language processing have enabled powerful attribution systems[1] that are capable of inferring author identity by analyzing text style alone (Abbasi and Chen, 2008; Narayanan et al., 2012; Overdorf and Greenstadt, 2016; Stolerman et al., 2013; Ruder et al., 2016). There have been several recent attempts to attribute the authorship of anonymously published text using such advanced authorship attribution approaches.[2] This poses a serious threat to privacy-conscious individuals, especially human rights activists and journalists who seek anonymity for safety.

Researchers have started to explore text *obfuscation* as a countermeasure to evade privacy-invasive authorship attribution. Anonymouth (McDonald et al., 2012; Brennan et al., 2012) was proposed to identify words or phrases that are most revealing of author identity so that these could be manually changed by users seeking anonymity. Since it can be challenging for users to manually make such changes, follow up work proposed rule-based text obfuscators that can automatically manipulate certain text features (e.g., spelling or synonym) (McDonald et al., 2013; Almishari et al., 2014; Keswani et al., 2016; Karadzhov et al., 2017; Castro-Castro et al., 2017; Mansoorizadeh et al., 2016; Kacmarcik and Gamon, 2006; Kingma and Welling, 2018). Since then more sophisticated learning-based text obfuscators have been proposed that automatically manipulate text to evade state-of-the-art authorship attribution approaches (Karadzhov et al., 2017; Shetty et al., 2018; Li et al., 2018; Mahmood et al., 2019; Gröndahl and Asokan, 2020).

In the arms race between authorship attribution and authorship obfuscation, it is important that both attribution and obfuscation consider the adversarial threat model (Potthast et al., 2018). While recent work has focused on developing authorship obfuscators that can evade state-of-the-art authorship attribution approaches, there is little work on developing authorship attribution approaches that can work against state-of-the-art authorship obfuscators. Existing authorship attributors are primarily designed for the non-adversarial threat model and only evaluated against non-obfuscated documents. Thus, it is not surprising that they can be readily evaded by state-of-the-art authorship obfuscators

* This paper is third in the series. See (Mahmood et al., 2019) and (Mahmood et al., 2020) for the first two papers.

† Our code and data are available at: https://github.com/reginazhai/Authorship-Deobfuscation

[1] https://www.eff.org/deeplinks/2013/06/internet-and-surveillance-UN-makes-the-connection

[2] https://www.nbcchicago.com/news/politics/Science-May-Help-Identify-Opinion-Columnist-492649561.html

Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7372 - 7384

(Karadzhov et al., 2017; Shetty et al., 2018; Li et al., 2018; Mahmood et al., 2019; Gröndahl and Asokan, 2020).

To fill this gap, we investigate the problem of authorship deobfuscation where the goal is to develop adversarial authorship attribution approaches that are able to attribute obfuscated documents. We study the problem of *adversarial authorship attribution* in the following two settings. *First,* we develop attributors that filter obfuscated documents using obfuscation/obfuscator detectors and then use an authorship attributor that is adversarially trained on obfuscated documents. *Second,* we develop adversarially trained authorship attributors that does not make assumptions about whether and which authorship obfuscator is used.

The results show that our authorship deobfuscation approaches are able to significantly reduce the adverse impact of obfuscation, which results in up to 20-30% degradation in attribution accuracy. We find that an authorship attributor that is purpose-built for obfuscated documents is able to improve attribution accuracy to within 5% as without obfuscation. We also find that an adversarially trained authorship attributor is able to improve attribution accuracy to within 10% as without obfuscation. Additionally, we evaluate the effectiveness of adversarial training when the attributor makes incorrect assumptions about whether and which obfuscator is used. We find that these erroneous assumptions degrade accuracy up to 20%, however, this degradation is the same or smaller than when the attributor is not adversarially trained, which can degrade accuracy up to 32%.

Our key contributions include:

• investigating the **novel problem** of adversarial authorship attribution for deobfuscation;

• **proposing approaches** for adversarial authorship attribution; and

• **evaluating robustness** of existing authorship obfuscators against adversarial attribution.

Ethics Statement: We acknowledge that authorship deobfuscation in itself is detrimental to privacy. Our goal is to highlight a major limitation of prior work on authorship obfuscation under the adversarial threat model. We expect our work to foster further research into new authorship obfuscation approaches that are resistant to deobfuscation.

2 Related Work

Authorship attribution is the task of identifying the correct author of a document given a range of possible authors. It has been a long-standing topic, and researchers have developed a wide range of solutions to the problem. Earlier researchers focus more on analysis based on writing style features. These include the distribution of word counts and basic Bayesian methods (Mosteller and Wallace, 1963), different types of writing-style features (lexical, syntactic, structural, and content-specific) (Zheng et al., 2006), and authors' choices of synonyms (Clark and Hannon, 2007). Other researchers combined machine learning and deep learning methods with stylometric features. Abbasi and Chen (2008) combine their rich feature set, "Writeprints", with an SVM. Brennan et al. (2012) improve "Writeprints" to reduce the computational load required of the feature set. Finally, more recent research focuses on fine-tuning pre-trained models since they do not require predefined features sets. Ruder et al. (2016) tackle authorship attribution with a CNN, while Howard and Ruder (2018) introduce the Universal Language Model Fine-tuning (ULMFiT) which shows strong performance in attribution.

To the best of our knowledge, prior work lacks approaches for adversarial authorship deobfuscation. Prior work has shown that existing authorship attributors do not perform well against obfuscators. Brennan et al. (2012) present a manual obfuscation experiment which causes large accuracy degradation. Since this obfuscation experiment, much has been done in the area of authorship text obfuscation (Rao and Rohatgi, 2000; Brennan et al., 2012; McDonald et al., 2012, 2013; Karadzhov et al., 2017; Castro et al., 2017; Mahmood et al., 2019; Gröndahl and Asokan, 2020; Bo et al., 2019). We focus on state-of-the-art obfuscators, Mutant-X (Mahmood et al., 2019) and DS-PAN (Castro et al., 2017) specifically in our research. Other obfuscation methods are as vulnerable to adversarial training which is reinforced in (Gröndahl and Asokan, 2020).

Our proposed authorship attributor leverages adversarial training to attribute documents regardless of obfuscation. First described in (Goodfellow et al., 2014), adversarial training uses text produced by an adversary to train a model to be more robust. Adversarial training has seen success in other text domains including strengthening word embeddings

(Miyato et al., 2016), better classification in cross-lingual texts (Dong et al., 2020), and attacking classifiers (Behjati et al., 2019).

3 Methodology

In this section, we present our approaches for adversarial authorship attribution for deobfuscation.

3.1 Threat Model

We start by describing the threat model for the authorship deobfuscation attack. There is an arms race between an attacker (who desires to identify/attribute the author of a given document) and a defender (an author who desires privacy and therefore uses an obfuscator to protect their identity). Figure 1 illustrates the expected workflow between the defender and the attacker. The defender uses an obfuscator before publishing the documents and the attacker employs obfuscation and/or obfuscator detector as well as an adversarially trained attributor for deobfuscation.

Defender. The goal of the defender is to obfuscate a document so that it cannot be attributed to the author. The obfuscator takes as input an original document and obfuscates it to produce an obfuscated version that is expected to evade authorship attribution.

Attacker. The goal of the attacker is to use an attributor trained on documents from multiple authors to identify the author of a given document. The attacker assumes to know the list of potential authors in the traditional closed-world setting. We examine two scenarios: First, as shown in Figure 1a, the attacker assumes to know that the document is obfuscated and also the obfuscator used by the defender. In this scenario, the attacker is able to access the documents that are produced by the obfuscator and hence train an attributor for obfuscated documents from the obfuscator. Second, as shown in Figure 1b, the attacker assumes to know that the document is obfuscated and that there is a pool of available obfuscators, of which one is used by the defender. Note that the attacker does not know exactly which obfuscator from the pool was used by the defender. Thus, the attacker trains an attributor for documents that are obfuscated by any one of the pool of available obfuscators.

3.2 Obfuscation

We use two state-of-the-art text obfuscators .

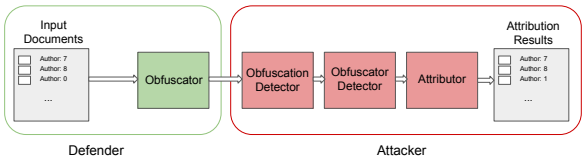
Document Simplification (DS-PAN). This approach obfuscates documents through rule-based sentence simplification (Castro et al., 2017). The transformation rules include lexical transformations, substitutions of contractions or expansions, and eliminations of discourse markers and fragments of text in parenthesis. This approach was one of the best performing in the annual PAN competition, a shared CLEF task (Potthast et al., 2017). It was also one of the few approaches that achieves "passable" and even "correct" judgements on the soundness of obfuscated text (i.e., whether the semantics of the original text are preserved) (Hagen et al., 2017). We refer to this approach as DS-PAN.

Mutant-X. This approach performs obfuscation using a genetic algorithm based search framework (Mahmood et al., 2019). It makes changes to input text based on the attribution probability and semantics iteratively so that obfuscation improves at each step. It is also a fully automated authorship obfuscation approach and outperformed text obfuscation approaches from PAN (Potthast et al., 2017) and has since been used by other text obfuscation approaches (Gröndahl and Asokan, 2020). There are two versions of Mutant-X: Mutant-X writeprintsRFC, which uses Random Forests along with Writeprints-Static features (Brennan et al., 2012); and Mutant-X embeddingCNN, which uses a Convolutional Neural Network (CNN) classifier with word embeddings. We use writeprintsRFC version because it achieves better drop in attribution accuracy and semantic preservation as compared to embeddingCNN.

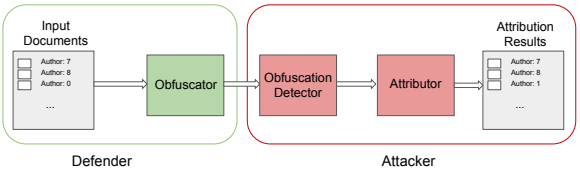
3.3 Deobfuscation

We describe the design of the authorship attributor and our adversarial training approaches for deobfuscation.

Authorship Attributor. We use writeprintsRFC as the classifier for authorship attribution. More specifically, we use the Writeprints-Static feature set (Brennan et al., 2012) that includes lexical features on different levels, such as word level (total number of words) and letter level (letter frequency) as well as syntactic features such as the frequency of functional words and parts of speech tags. It is one of the most widely used stylometric feature sets and has consistently achieved high accuracy on different datasets and author sets while maintaining a low computational cost. We then use these features to train an ensemble random forest classifier



(a) Scenario 1: Attacker knows the document is obfuscated and the obfuscator used



(b) Scenario 2: Attacker only knows the document is obfuscated

Figure 1: Deobfuscation pipeline using obfuscation and/or obfuscator detectors for adversarial training

with 50 decision trees.

Adversarial Training. The basic idea of adversarial training is to include perturbed/obfuscated inputs into the training set to improve the model's resistance towards such adversarially obfuscated inputs (Goodfellow et al., 2014). It has been widely used in various domains including text classification. In our case, obfuscated texts are texts that vary slightly from the original texts and these serve as adversarial examples. We examine how using these adversarial examples as training data influences the attributor's performance and whether it adds resilience against obfuscation. Based on our two scenarios described in Section 3.1 and shown in Figure 1, we propose two ways of adversarial training. For both cases, original texts from the list of possible authors are selected and prepared for obfuscation. For scenario 1, we train the attributor using documents obfuscated by a known obfuscator. For scenario 2, since the attacker does not assume to know the specific obfuscator used by the defender, we train the attributor using documents obfuscated by the pool of available obfuscators.

4 Experimental Setup

We describe the dataset, evaluation metrics, and experimental design to assess the effectiveness of our adversarial authorship attribution approaches for deobfuscation.

Dataset. Following previous research (Mahmood et al., 2019), we examine a publicly available dataset for evaluation of our methodology. The Blog Authorship Corpus (Schler et al., 2006) contains over 600,000 blog posts from blogger.com. These posts span 19,320 unique authors. Previous research (Narayanan et al., 2012) found that authorship attribution gets harder when more authors are included. Based on the author selection in (Mahmood et al., 2019), we select a subset of 15 each with 100 documents (compared to their 5 and 10 authors) for a more precised evaluation. These

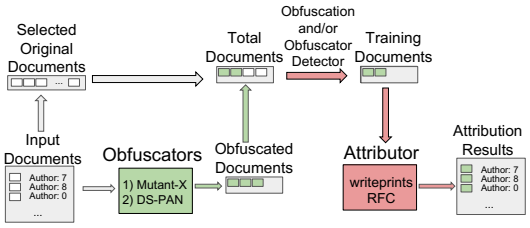


Figure 2: Generalized deobfuscation training process using adversarial training

1500 documents are divided into 80-20% split for training and testing, respectively. Specifically, 80 documents from each author are used in the training set while the rest 20 documents are used in the test set.

As shown in Figure 2, we train on various combinations of obfuscated documents. These documents are obfuscated by the obfuscators described in Section 3.2. When an attributor-dependent-obfuscator (e.g. Mutant-X (Mahmood et al., 2019)) is used, the attributor will have access to the same training documents used to train the obfuscator. Otherwise, the attributor does not assume to have access to the attributor used by the obfuscator. To control for training size, when more than 1 obfuscator is used, we sample equal amounts of documents from each set of obfuscated documents. For example, if we train against 2 obfuscators, then 600 documents are sampled from each set of respective obfuscated documents to get a training set of size 1200.

To calibrate the obfuscated texts, we use METEOR score (Banerjee and Lavie, 2005) to evaluate the soundness of documents. The score for Mutant-X ranges from 0.3 to 0.7 (mean=0.46), and the score for DS-PAN ranges from 0.24 to 0.57 (mean=0.38), which are comparable to previous studies (Mahmood et al., 2019). An in-depth analysis of the METEOR score is reported in Appendix A.

Metric. To quantify attribution performance on the

15-class problem, we calculate the accuracy as:

$$\text{accuracy} = \frac{\text{\# of correctly attributed documents}}{\text{total \# of documents}} \quad (1)$$

Attack Scenarios. Figure 3 illustrates the flow of our experimental evaluation under different attack scenarios.

0. *Baseline:* For the baseline model, we assume that there is no obfuscation in this world. The attacker is trained on original documents and is deployed on original documents.

1. *Obfuscation-unaware-attacker:* The first case we examine is when the defender actively seeks to hide author identity. Thus, the defender gains an advantage by obfuscating documents using either Mutant-X or DS-PAN in order to bypass the attacker. The attacker, however, remains unaware of obfuscation and trains the attributor only on original documents.

2. *Obfuscation-aware-attacker with obfuscation detector:* Next, we give knowledge of obfuscation to the attacker by introducing an obfuscation detector into the system. Previous research (Mahmood et al., 2020) shows that texts generated by existing obfuscators can be detected as obfuscated with high accuracy. The device for this type of detection is called an obfuscation detector. Hence, in this scenario we ask whether there is any benefit to the attacker if the text is identified as obfuscated before attribution. Since the attacker does not know which obfuscator was used by the defender, the attributor is trained on the combination of documents generated from DS-PAN and from Mutant-X. The defender is the same as in the previous scenario, i.e., it uses one of the obfuscators to generate documents.

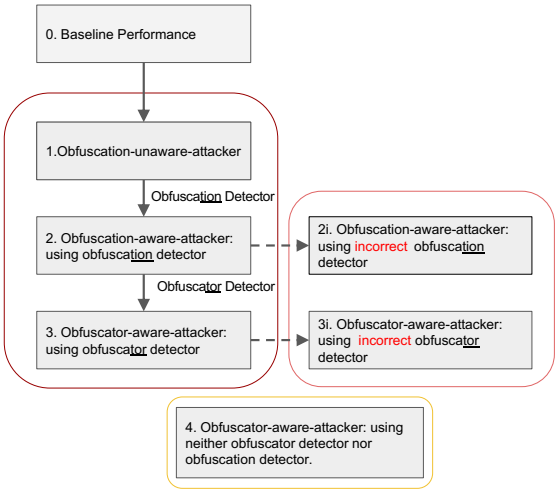


Figure 3: Progression of various attack scenarios

3. *Obfuscator-aware-attacker with obfuscator detector:* We give additional information to the attacker. Previous research has shown that embedding watermarks and fingerprints that show the identity of the model into deep neural networks is a valid approach to protect the rights of shared trained models (Uchida et al., 2017; Zhang et al., 2018). Hence, it is reasonable to assume that there will be methods in the future to identify the watermarks for specific deep neural networks. Here, we propose the concept of obfuscator detector, which can detect the specific obfuscator used. In this case, the attacker attributor is trained always on the documents generated by the same obfuscator as the defender: either documents generated from DS-PAN or from Mutant-X.

2i. *Obfuscation-aware-attacker with incorrect obfuscation detector:* Here we ask the question: what happens in scenario 2 if the obfuscation detector makes errors? The specific error addressed is that the detector classifies the text as obfuscated whereas it is actually an original. Under this condition, the attacker attributor is still trained on the combination of documents generated from DS-PAN and from Mutant-X. But the defender now presents an original document.

3i. *Obfuscator-aware-attacker with incorrect obfuscator detector:* When the obfuscator detector classifies incorrectly, it assumes that the defender uses a specific obfuscator when it actually uses a different one. The attacker attributor is trained on the documents generated by one of the obfuscators: either documents generated from DS-PAN or from Mutant-X. However, the defender uses a different obfuscator than the attacker to generate the documents.

4. *Obfuscator-aware-attacker that does not rely on an obfuscator detector or obfuscation detector:* Since the previous processes require the proposed obfuscation and obfuscator detector, it is not efficient. Hence, a simpler, more efficient solution is to train on all the documents at once. In this simplified version, the attacker attributor is trained on the combination of original documents, documents generated from DS-PAN, and documents generated from Mutant-X. Since this is the combined condition, the defender may or may not use an obfuscator, and will choose from the two possible obfuscators to generate documents.

5 Results

In this section, we present the results following the progression of various attack scenarios shown in Figure 3.

5.1 Major Scenarios

0. *Baseline:* The original authorship attributor has an accuracy of 76.7% when trained on the original documents and tested on original documents. The attribution accuracy should be higher than 6.7%, which is when we attribute the 15 authors randomly, to be considered effective.

1. *Obfuscation-unaware-attacker:* Based on the first row of Table 1, the result shows that the attribution accuracy drops from 76.7% to 50.7% with a decrease of 26% when tested on DS-PAN obfuscated documents, while the accuracy for testing on Mutant-X obfuscated documents drops from 76.7% to 44.3%. The average drop in accuracy is from 76.7% to 47.5%, which is 29.2%. Based on the results, we know that as shown by previous works (Karadzhov et al., 2017; Mahmood et al., 2019) on the performance of the obfuscators, DS-PAN and Mutant-X obfuscators can successfully decrease the attribution accuracy of original attributor.

2. *Obfuscation-aware-attacker with obfuscation detector:* The second row of Table 1 shows that attribution accuracy increases by 13.2% from 50.7% to 63.9% when tested on DS-PAN documents, and increases by 24.7% from 44.3% to 69% when tested on Mutant-X documents. The average accuracy is 66.4%, which increases from the previous 47.5% by about 19%. While the performance is still not comparable to the baseline results, the increase in accuracy from the previous scenario is significant, which suggests that the obfuscation detector would benefit the performance against obfuscated documents.

3. *Obfuscator-aware-attacker with obfuscator detector:* As shown on the third and forth row of Table 1, when trained only on DS-PAN documents, the accuracy tested on DS-PAN is 68.6%, with an increase of 17.9% from the first scenario; when trained only on Mutant-X documents, the accuracy tested on Mutant-X is 75.7%, with an increase of 31.4%. The average test accuracy is 71.1%, which increases by about 5% compared to the 66.4% in the previous case. From the results, we can see that having an obfuscator detector as well as an obfuscation detector is the most beneficial to improve the attribution accuracy from obfuscated texts.

Training set	Test set		
	DS-PAN	MutantX	Average
Original	50.7	44.3	47.5
DS-PAN+MutantX	63.9	69.0	**66.4**
DS-PAN	**68.6**	-	-
MutantX	-	**75.7**	-

Table 1: Accuracy of original attributor and different adversarially trained attributors tested against different obfuscators

5.2 Error Conditions

Although obfuscation/obfuscator detector are quite accurate, they are not perfect. Hence, we test the success of the attacker when the obfuscation detector and obfuscator detector are incorrect.

2i. *Obfuscation-aware-attacker with incorrect obfuscation detector:* Shown on the first column of row four on Table 2, the attribution accuracy decreases by 8.4% from the baseline 76.7% to 68.3%, but a higher accuracy is maintained than the average of Attack Scenario 2 (66.4%) The result shows that when the obfuscation detector produces wrong results, performance will be influenced, but still stay at a relatively high level. Thus, having an obfuscation detector is generally good for the attacker with little cost.

3i. *Obfuscator-aware-attacker with incorrect obfuscator detector:* From second and third rows of Table 2 we see that when the attacker is trained only on DS-PAN documents, the accuracy tested on Mutant-X is 57.3%, with a drop in performance of 18.4% when compared to training on only Mutant-X documents (75.7%). When the attacker is trained only on Mutant-X documents, the accuracy tested on DS-PAN is 48.5%, with a drop in performance of 20.1% as compared to training on only DS-PAN documents (68.6%). The average test accuracy is 52.9%, which is lower than training on the same obfuscator, but higher than the results in 1 of 5.1 (50.7% and 44.3%). When the obfuscator detector gives incorrect results, the attribution accuracy will not achieve its best performance, but the result is still higher than trained only on original documents. Hence, using obfuscated documents to train always tends to benefit the attribution accuracy.

5.2.1 Combined Condition

Here the attacker simply uses originals and obfuscated documents from all available obfuscators for adversarial training of the attributor.

4. *Obfuscator-aware-attacker that does not rely on an obfuscator detector or obfuscation detector:*

This result is shown on the last row of Table 2. Attribution accuracy when tested on original documents drops from 76.7% to 66.3%, but increases by 10.5% from 50.7% to 61.2% when tested on DS-PAN, and increases by 24.5% from 44.3% to 68.8% when tested on Mutant-X. The average accuracy is 65%, which increases from the average of the former three, 57.2%, by about 8%. While the attacker does not know if the document is obfuscated or not, or by which obfuscator, it is still able to achieve a high boost in attribution accuracy by adversarial training. Therefore, although the previous processes can achieve higher performances, training on a combination of these documents could be a valid approach when time and resources are limited.

6 Discussion

Next, we look more closely into the results from adversarial training to better understand them.

6.1 General Author Analysis

Figure 4 pesents the confusion matrices produced from DS-PAN obfuscated documents tested on Attack Scenario 1, 2 and 3 respectively. Rows represent the Original Authors, while the columns represent the Predicted Authors. The values in the matrices are the percentage of the original documents that are classified as a specific author.

Moving from scenario 1 to 3, we see an increase in color density and percentage on the diagonal, which signifies the general increase in accuracy when the training documents become more specific. Consistent with above, the color on the non-diagonal areas becoming more transparent also indicates reduction of classification errors. At the author level, we observe that almost all of the authors show increases in accuracy on the diagonal cells across the three scenarios. It shows that adversarial training is effective even on authors with different styles.

Looking more closely at each author, we know that Author 9 is the easiest to classify - performance is always at 100%. Author 6, on the other hand, is relatively hard to attribute. The best performance for Author 6 is only 35% from the most effective Attack Scenario 3.

Figure 6 presents another view on performance. It shows the percentage of errors made for each author out of all the errors in the three scenarios combined (note: the sum of all errors in the figure is 100%). Thus, the errors made for Author 1 under Scenario 1 is 3.18% of total errors across the three scenarios. We observe that the color is generally darker in Scenario 1, while it gradually lightens in Scenario 2 and then in Scenario 3. Again, this indicates the benefit of having more specific training data. Looking more closely within each scenario, we see that the attributor of Attacker Scenario 1 tends to misclassify Authors 5 and 8 the most. But the attributors for Scenario 2 and Scenario 3 learn more effectively for these two authors thereby reducing mistakes. For Attack Scenario 3, the most misclassified author is Author 6, where 3.76% of all errors. But this percentage is still an improvement over the 4.34% in the previous two scenarios. Motivated by the above observations, next we investigate shifts in performance for a specific author.

6.2 Individual Author Analysis

We assign labels to the 15 authors in the dataset and select Original Author 15 for more detailed analysis. The reason we choose Author 15 is that its accuracy is among the ones that increases the most, from 45% to 80%. In order to find out the reasons behind such increase, we perform PCA analysis on all of the DS-PAN documents whose original author is Author 15. We use Writeprints-Static feature set, which has a total of 555 features. In order to preserve the most significant features for attribution, we select the most important 25 features from the original writeprintsRFC and process them through PCA so that we can visualize the features into 3 dimensional graphs.

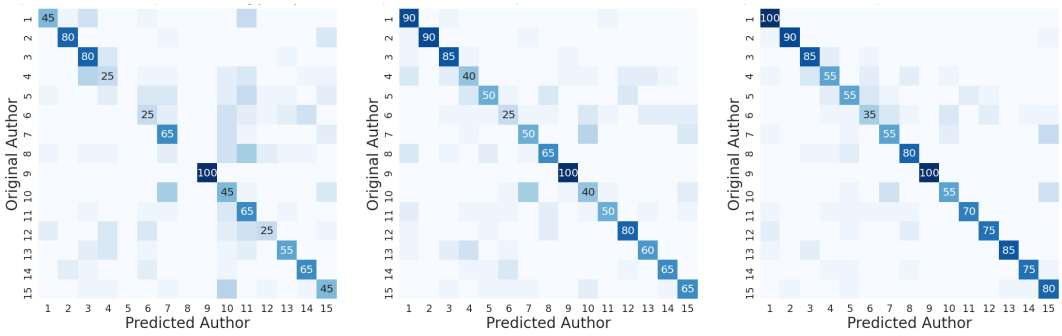
As shown in the graphs in Figure 5, each dot on the graph represents a document. The green ones are the ones that are attributed correctly while the red ones are attributed incorrectly. In Figure 5a, the incorrectly attributed ones are mainly gathered in a cluster. This suggests that the attributor has trouble discriminating the documents that are similar to each other. But as we go from left to right, the documents in the cluster are also gradually attributed correctly. The trend shows that the attributor is getting better at distinguishing between documents that are similar to each other. Hence, we can infer that adversarial training improves attribution accuracy by discriminating between the ones that are more similar to each other.

6.3 Comparing DS-PAN and Mutant-X

In Attack Scenarios 2, 3, and 4, the test sets using DS-PAN for obfuscation yield worse attribution

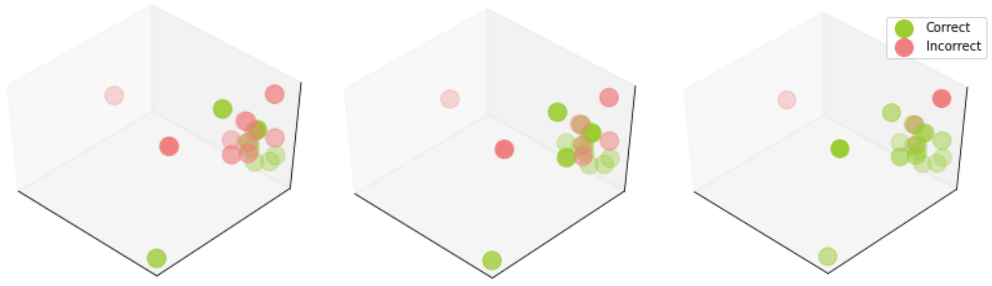
Training set	Test set			
	Original	DS-PAN	MutantX	Average of DS+MX
Original	**76.7**	50.7	44.3	47.5
DS-PAN	57.3	**68.6**	57.3	62.9
MutantX	72.0	48.5	**75.7**	62.1
DS-PAN + MutantX	68.3	63.9	69.0	**66.4**
DS-PAN + MutantX + Original	66.3	61.2	68.8	65.0

Table 2: Accuracy of adversarial training on various combinations of test documents



(a) Attack Scenario 1 (b) Attack Scenario 2 (c) Attack Scenario 3

Figure 4: Confusion matrices of different attack scenarios



(a) Attack Scenario 1 (b) Attack Scenario 2 (c) Attack Scenario 3

Figure 5: Attribution performance of Author 15 with PCA under different attack scenarios

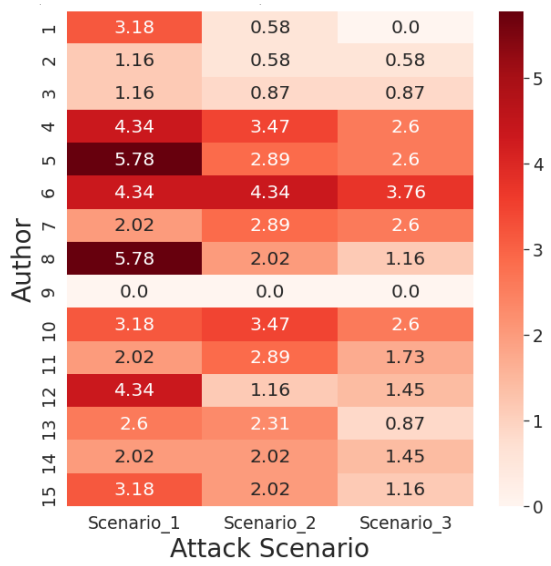


Figure 6: Percentage of misclassified document for each author across attack scenarios

accuracy than those using Mutant-X. Our analysis of obfuscated documents showed that DS-PAN makes both a greater number of changes as well as more significant changes as compared to Mutant-X. Thus, we surmise that DS-PAN results in larger degradation in attribution accuracy because the attacker's training set contains text that is less similar to the original text. However, the changes made by DS-PAN also have side effect in that they lower the soundness of obfuscated text as reflected by lower METEOR scores. The mean METEOR score for DS-PAN is 0.38 as compared to 0.46 for Mutant-X. A more detailed analysis of METEOR score and semantic similarity between obfuscated and original texts is reported in Appendix A.

6.4 Insights into Adversarial Training

The performance gain of adversarial training comes from a "noisy" training dataset comprising of obfuscated documents as well as knowledge about the obfuscator. To disentangle these two factors, we compare the accuracy improvements of the second and third rows of Table 2 against the Mutant-X obfuscated test documents. We note that the improvement in attribution accuracy is 13% when DS-PAN obfuscated documents are used for training. The improvement in attribution accuracy is further 18% (31% overall) when Mutant-X obfuscated documents are used for training. This difference (13% vs. 18%) indicates that although having a noisy dataset helps, the knowledge of the specific obfuscator is likely more crucial to improving attribution performance. This trend holds for DS-PAN obfuscated test documents.

7 Concluding Remarks

In this work, we explored the novel problem of adversarial authorship attribution for deobfuscation. We demonstrate that adversarial training is able to significantly reduce the adverse impact of existing text obfuscators on authorship attribution accuracy. We found that an adversarially trained authorship attributor improves attribution accuracy to within 5-10% as without obfuscation. While an adversarially trained authorship attributor achieved best accuracy when it is trained using the documents obfuscated by the respective obfuscator, we found that it achieves reasonable accuracy even when it is trained using documents obfuscated by a pool of obfuscators. When the adversarially trained attributor makes erroneous assumptions about the obfuscator used to obfuscate documents, we note a degradation in attribution accuracy. It is noteworthy, however, that this degradation is still similar or better than the attribution accuracy of the baseline attributor that is not adversarially trained.

Our results shed light into the future of the ensuing arms race between obfuscators and attributors. Most notably, we find that the effectiveness of adversarial training is somewhat limited if the obfuscators continue to employ new and improved methods that are not available to attributors for adversarial training. Therefore, it is important to continue development of new and improved text obfuscation approaches that are resistant to deobfuscation (Bevendorff et al., 2019; Bo et al., 2019; Gröndahl and Asokan, 2020; Hlavcheva et al., 2021). On the other hand, recent work on understanding and improving transferability of adversarial attacks can inform development of better adversarial attributors that might work well even for unknown obfuscators (Tramèr et al., 2017; Zheng et al., 2020; He et al., 2021; Mireshghallah and Berg-Kirkpatrick, 2021).

Finally, our experiments were limited to the closed-world setting where the universe of potential authors is assumed to be known by the attributor. Further research is needed to investigate whether (and how much) adversarial algorithms are effective in the open-world setting.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identi-

fication and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS)*, 26(2):7.

Mishari Almishari, Ekin Oguz, and Gene Tsudik. 2014. Fighting Authorship Linkability with Crowdsourcing. In *ACM Conference on Online Social Networks (COSN)*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and Summarization*.

Melika Behjati, Seyed-Mohsen Moosavi-Dezfooli, Mahdieh Soleymani Baghshah, and Pascal Frossard. 2019. Universal adversarial attacks on text classifiers. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7345–7349. IEEE.

Janek Bevendorff, Martin Potthast, Matthias Hagen, and Benno Stein. 2019. Heuristic authorship obfuscation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1098–1108.

Haohan Bo, Steven HH Ding, Benjamin Fung, and Farkhund Iqbal. 2019. Er-ae: differentially-private text generation for authorship anonymization. *arXiv preprint arXiv:1907.08736*.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. In *ACM Transactions on Information and System Security (TISSEC)*, volume 15, pages 12:1 – 12:22.

Daniel Castro, Reynier Ortega, and Rafael Muñoz. 2017. Author masking by sentence transformation—notebook for pan at clef 2017. In *CLEF 2017 Evaluation Labs and Workshop–Working Notes Papers*, pages 11–14.

Daniel Castro-Castro, Reynier Ortega Bueno, and Rafael Munoz. 2017. Author Masking by Sentence Transformation. In *Notebook for PAN at CLEF*.

Jonathan H. Clark and Charles J. Hannon. 2007. An Algorithm for Identifying Authors Using Synonyms. In *Eighth Mexican International Conference on Current Trends in Computer Science (ENC 2007)*, pages 99–104. IEEE.

Xin Dong, Yaxin Zhu, Yupeng Zhang, Zuohui Fu, Dongkuan Xu, Sen Yang, and Gerard De Melo. 2020. Leveraging adversarial training in self-learning for cross-lingual text classification. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1541–1544.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Tommi Gröndahl and N Asokan. 2020. Effective writing style transfer via combinatorial paraphrasing. *Proc. Priv. Enhancing Technol.*, 2020(4):175–195.

Matthias Hagen, Martin Potthast, and Benno Stein. 2017. Overview of the author obfuscation task at pan 2017: Safety evaluation revisited. In *CLEF (Working Notes)*.

Xuanli He, Lingjuan Lyu, Qiongkai Xu, and Lichao Sun. 2021. Model extraction and adversarial transferability, your bert is vulnerable! *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Yulia Hlavcheva, Victoria Bobicev, Olga Kanishcheva, et al. 2021. Language-independent features for authorship attribution on ukrainian texts. In *CEUR Workshop Proceedings*, volume 2833, pages 134–143.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Gary Kacmarcik and Michael Gamon. 2006. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 444–451. Association for Computational Linguistics.

Georgi Karadzhov, Tsvetomila Mihaylova, Yasen Kiprov, Georgi Georgiev, Ivan Koychev, and Preslav Nakov. 2017. The case for being average: A mediocrity approach to style masking and author obfuscation. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 173–185. Springer.

Yashwant Keswani, Harsh Trivedi, Parth Mehta, and Prasenjit Majumder. 2016. Author Masking through Translation. In *Notebook for PAN at CLEF 2016*, pages 890–894.

Diedrik P Kingma and Max Welling. 2018. Auto-encoding variational bayes. In *Proceedings of NAACL-HLT*, pages 1865–1874.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. TextBugger: Generating Adversarial Text Against Real-world Applications. In *Network and Distributed Systems Security (NDSS) Symposium*.

Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using x-mutant. In *Privacy Enhancing Technologies Symposium (PETS)*.

Asad Mahmood, Zubair Shafiq, and Padmini Srinivasan. 2020. A girl has a name: Detecting authorship obfuscation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2235–2245.

Muharram Mansoorizadeh, Taher Rahgooy, Mohammad Aminiyan, and Mahdy Eskandari. 2016. Author obfuscation using WordNet and language models. In *Notebook for PAN at CLEF 2016*.

Andrew WE McDonald, Sadia Afroz, Aylin Caliskan, Ariel Stolerman, and Rachel Greenstadt. 2012. Use fewer instances of the letter 'i': Toward writing style anonymization. In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 299–318. Springer.

Andrew W.E. McDonald, Jeffrey Ulman, Marc Barrowclift, and Rachel Greenstadt. 2013. Anonymouth Revamped: Getting Closer to Stylometric Anonymity. In *PETools: Workshop on Privacy Enhancing Tools*, volume 20.

Fatemehsadat Mireshghallah and Taylor Berg-Kirkpatrick. 2021. Style Pooling: Automatic Text Style Obfuscation for Improved Classification Fairness. In *EMNLP*.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.

Frederick Mosteller and David L Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *Journal of the American Statistical Association*, 58(302):275–309.

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 *IEEE Symposium on Security and Privacy (SP)*, pages 300–314. IEEE.

Rebekah Overdorf and Rachel Greenstadt. 2016. Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. In *Privacy Enhancing Technologies Symposium (PETS)*.

Martin Potthast, Francisco Rangel, Michael Tschuggnall, Efstathios Stamatatos, Paolo Rosso, and Benno Stein. 2017. Overview of pan'17. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 275–290. Springer.

Martin Potthast, Felix Schremmer, Matthias Hagen, and Benno Stein. 2018. Overview of the author obfuscation task at pan 2018: A new approach to measuring safety. In *Notebook for PAN at CLEF 2018*.

Josyula R Rao and Pankaj Rohatgi. 2000. Can pseudonymity really guarantee privacy? In *USENIX Security Symposium*, pages 85–96.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *arXiv:1609.06686*.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. 2018. A4NT: Author Attribute Anonymity by Adversarial Training of Neural Machine Translation. In *USENIX Security Symposium*.

Ariel Stolerman, Rebekah Overdorf, Sadia Afroz, and Rachel Greenstadt. 2013. Classify, but verify: Breaking the closed-world assumption in stylometric authorship attribution. In *IFIP Working Group*, volume 11, page 64.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2017. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*.

Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin'ichi Satoh. 2017. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*, pages 269–277.

Jialong Zhang, Zhongshu Gu, Jiyong Jang, Hui Wu, Marc Ph Stoecklin, Heqing Huang, and Ian Molloy. 2018. Protecting intellectual property of deep neural networks with watermarking. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security*, pages 159–172.

Haizhong Zheng, Ziqi Zhang, Juncheng Gu, Honglak Lee, and Atul Prakash. 2020. Efficient adversarial training with transferable adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1181–1190.

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *Journal of the American society for information science and technology (JASIST)*.

A Qualitative Analysis

We conduct analysis to evaluate the quality of the text. We first evaluate the semantics of the obfuscated text with respect to the original text using METEOR scores. The results show that METEOR scores of obfuscated text are comparable to those reported in prior studies. We also conduct qualitative analysis of the obfuscated text.

First, we evaluate the quality of obfuscated documents from the two obfuscators. We use METEOR score to measure the soundness of the obfuscated text in terms of the semantic similarity between the original and the obfuscated text.

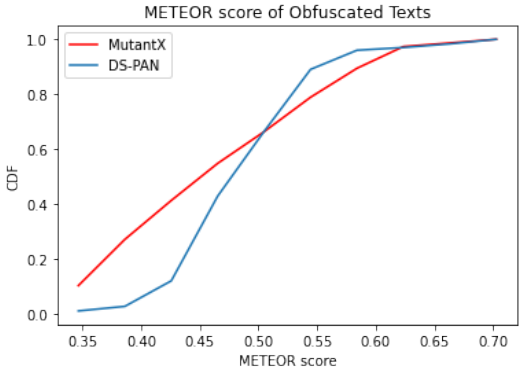


Figure 7: CDF plot of METEOR score for obfuscated texts

Figure 7 shows the distribution of the METEOR score for Mutant-X and DS-PAN. The plot shows that the METEOR scores for Mutant-X ranges from 0.3 to 0.7 (mean=0.46), and the METEOR score for DS-PAN ranges from 0.24 to 0.57 (mean=0.38). Compared to the previous METEOR score results calculated in (Mahmood et al., 2019), where the METEOR score for Mutant-X ranges from 0.48 to 0.55 (mean = 0.51), and the METEOR score for other baseline models ranges from 0.32 to 0.46 (mean = 0.38), the two obfuscators used in this work achieve similar results at preserving the semantics of the original texts.

Table 3 contains examples from the two obfuscators showing different types of changes. Synonym replacement is common in both systems. Examples of such are (street <-> sidewalk), (student <-> pupil). There are also changes in word form. (run <-> running), (waited <-> wait) preserves the morpheme, but changes the tense of the word. It is also worth noting that DS-PAN tends to change the form of abbreviations, such as (I'm <-> I am) and (to have <-> to've). In general, the transformations make sense to the readers, and preserve most of the original meanings. But there are also cases (like the last row) where the transformations change the content and break the grammar.

Index	Original	DS-PAN	MutantX
1	I'm not an expert	I'm not An expert	I am non an expert
2	What was the first print run?	What was the first print running?	What was the ane print run?
3	The New York Times ran a Styles section profile two weeks before publication	The New York Times ran a Styles editor profile two weeks before publication	the new_york_times run a styles division profile two calendar_week before publishing
4	Cornelius walks in off of the street.	Cornelius walks in off of the sidewalk	Cornelius walks in away of the street.
5	We've discovered librarians are very networked and seem to know about everything before it happens	We've found librarians are extremely networked and seem to believe about everything before it happens.	we suffer detect bibliothec are really network and appear to cognize about everything before it happen
6	Homework is minimal, but the reading load is daunting.	Homework is minor, but the reading load is daunting.	Prep is minimum, but the read load is daunt
7	Some traces of the original layout remain	Some traces of the manifest makeover remain	Some trace of the original layout stay
8	Some professors seem happy to have a visitor	Some professors seem happy to become a pilgrim	Some prof appear happy to've a visitor
9	He expects interest in the Nancy Pearl doll to be strongest in Seattle, where she is best known.	He expects grateful in the Nancy Pearl mannequin to be strongest in Seattle, where she is best known.	He expect involvement in the nancy_pearl dolly to be strongest in seattle, where she's well cognize.
10	When the sales slot came open a few months later, she applied.	When the sales position came open a few years later, she applied.	When the cut-rate_sale time_slot arrive open_up a few calendar_month she utilize.
11	Professors often mistake her for a student	Professors often mistake her for a campus	Prof frequently err her for a pupil
12	They may look sleepy, but many used-book stores are thriving.	They may look sleepy, although many used-book stores are mature	they may search sleepy-eyed, but many used-book stores are boom
13	The perfumed bear she gave to me lost his scent	The perfumed bobcat she gave to me lost his odor	The perfume bear she render to me lose his aroma
14	I suppose I would have just waited until the morning if I were her.	I reckon I will rest just waited until the afternoon if I were She.	I presuppose i'd suffer precisely wait until the morn if i were her.

Table 3: Sentences from test document showing the result of different obfuscators