A Multifaceted Framework to Evaluate Evasion, Content Preservation, and Misattribution in Authorship Obfuscation Techniques

Malik H. Altakrori
School of Computer Science
McGill University / Mila
Montreal, Canada
malik.altakrori@mail.mcgill.ca

Thomas Scialom
Meta AI
Paris, France
tscialom@meta.com

Benjamin C. M. Fung
School of Information Studies
McGill University / Mila
Montreal, Canada
ben.fung@mcgill.ca

Jackie Chi Kit Cheung
School of Computer Science
McGill University / Mila
Montreal, Canada
jcheung@cs.mcgill.ca

Abstract

Authorship obfuscation techniques have commonly been evaluated based on their ability to hide the author's identity (evasion) while preserving the content of the original text. However, to avoid overstating the systems' effectiveness, evasion detection must be evaluated using competitive identification techniques in settings that mimic real-life scenarios, and the outcomes of the content-preservation evaluation have to be interpretable by potential users of these obfuscation tools. Motivated by recent work on cross-topic authorship identification and content preservation in summarization, we re-evaluate different authorship obfuscation techniques on detection evasion and content preservation. Furthermore, we propose a new information-theoretic measure to characterize the misattribution harm that can be caused by detection evasion. Our results[1] reveal key weaknesses in state-of-the-art obfuscation techniques and a surprisingly competitive effectiveness from a back-translation baseline in all evaluation aspects.

1 Introduction

Authorship obfuscation is the task of masking the writing style of an author of a document to prevent authorship identification techniques from using stylistic patterns to reveal the author's identity (Kacmarcik and Gamon, 2006). The motivation for this task is to protect the public from the misuse of authorship identification techniques to suppress freedom of speech or to persecute whistle-blowers.

When a new authorship obfuscation technique is proposed, it is crucial to compare its effectiveness to state-of-the-art obfuscation tools in settings that accurately depict the real-life environment where such a tool may be used. One important assumption that should be made is that a de-anonymizer is likely to use the most competitive authorship identification tool available to identify the author of a document. Using an inferior or brittle authorship attribution technique, or a weak obfuscation baseline will overstate the performance of such obfuscation tools. For example, previous work on obfuscation has evaluated obfuscation techniques against an identification tool that uses the exact same features and classifier that were used to obfuscate it in the first place (Mahmood et al., 2019). This could lead to misleadingly high impression of the system's effectiveness.

Similarly, obfuscation techniques must convey the same intended message both before and after obfuscation. Therefore, when a new obfuscation tool is proposed, it is evaluated on both the quality of obfuscation and its ability to preserve the content. With the recent development in language models and their ability to generate text, many automatic measures have been introduced to evaluate the quality of this generated text (Novikova et al., 2017), and some of these measures have been used in obfuscation techniques.

The problem with existing content-preservation measures, however, is that they only provide an abstract, numerical score that limits the user's ability to pinpoint the part of the text that suffered from loss of information and requires re-modification. Recently, question answering-based approaches were proposed and shown to provide meaningful feedback in the form of questions that tell the user which information in the text has been changed and in which part (Durmus et al., 2020).

The concept of evasion in authorship obfusca-

[1] The code is available on https://malikaltakrori.github.io/papers/EMNLP2022/.

Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pages 2391–2406
December 7-11, 2022 ©2022 Association for Computational Linguistics

tion has a critical, potential harmful side-effect that we raise for the first time. In a classification setting, the typical setting in which evasion of obfuscation techniques are evaluated, a classifier has to pick one author from the set of candidate authors. An obfuscation technique may obfuscate a document by imitating another potential author, in effect unfairly "framing" another person. To investigate this behavior, we use an information-theoretic approach to evaluate the potential for misattribution of the obfuscating technique. Specifically, we propose a new evaluation measure; namely, misattribution harm where the goal is to characterize the confidence in the attribution algorithm rather than its output.

In this work, we highlight a number of issues with the existing work on obfuscation with respect to two dimensions: obfuscation effectiveness, and content preservation. We further propose a new evaluation dimension namely, misattribution. Our key contributions are the following:

- We show that a carefully selected baseline can outperform state-of-the-art obfuscation techniques.
- We use question answering as an evaluation measure for content preservation instead of token- and embedding-based approaches.
- Using information theory, we conduct a detailed analysis of the harming effect of misattributing a document to a different author to achieve detection evasion.

2 Background

Authorship obfuscation (Brennan et al., 2012) techniques aim to hide an author's writing style which can be used by authorship identification tools to reveal the true identity of that author. Here, the assumption is that the author has already taken the precautions to hide their identity by removing any identifying information such as their name or address from the text. By using obfuscation techniques, users aim to hide their writing habits which may or may not be known to them. With that in mind, it is important that the obfuscated text contains the same conveyed message after obfuscation.

2.1 Obfuscation

Obfuscation tools can be divided into two groups: generic off-the-shelf tools, and application-specific obfuscation tools.

Generic Tools. Examples of off-the-shelf tools include machine translation and data augmentation approaches (Mansoorizadeh et al., 2016). These tools have been adapted for the purpose of generating a slightly modified version of a document. Commonly, these tools are used as baselines to be compared against obfuscation-specific techniques (Brennan et al., 2012; Keswani et al., 2016) because they are easy to use, require no further training or extra data from the user, and need minimal knowledge about the obfuscation process. Table 1 is an example of these tools where translating a sentence into different languages and then back to the original one creates a modified version of the original sentence.

Text	Language	
How is it going bro	-	En
Wie geht es dir, Bruder?	En	De
Comment vas-tu mon frére?	De	Fr
How are you my brother?	Fr	En

Table 1: Back translation is a technique used to paraphrase a sentence by translating it to different languages and then back to original language.

When machine translation approaches were initially used, only statistical machine translation (SMT) methods such as Moses (Koehn et al., 2007) and Google's previous Translate API (Wu et al., 2016) were available. They were shown to suffer from low obfuscation effectiveness and generated text with poor linguistic fluency compared to obfuscation techniques. In contrast, more recent neural machine translation approaches are able to generate higher-quality translations compared to SMT approaches according to some evaluation metrics. This development warrants re-evaluating their performance, especially as both Brennan et al. (2012) and Keswani et al. (2016) used SMT approaches.

In this work, we use well-tuned baselines that are expected to be competitive with obfuscation techniques as opposed to using simple and primitive ones. An example of excluded baselines is Random Replacement which tries to obfuscate a document by replacing words in that document with a random word from the author's vocabulary set, or with a synonym from a dictionary. Such baselines have been explored heavily in the literature and are known for their poor obfuscation performance and incoherent output.

Obfuscation-specific Tools. By contrast, obfuscation tools are built specifically to hide the author's identity and are tested against state-of-the-art authorship attribution techniques. While these tools require further training and/or additional data, they have been shown to be more effective than generic tools.

In this work, we evaluate two different approaches that focus specifically on obfuscation. Mutant-X (Mahmood et al., 2019) is a genetic algorithm that utilizes GloVE (Pennington et al., 2014) word embeddings to replace words in a document with similar ones to create a modified version of a document. This technique requires knowledge of the authorship attribution classifier, specifically, the probability of each author, to do the obfuscation.

Heuristic Obfuscation Search (Bevendorff et al., 2019, 2020) was initially developed as an imitation approach to obfuscation. The algorithm requires a target author profile which is the tri-grams frequency and the goal is to generate a document with a similar author profile. This is a rule-based approach where changes to the text—based on different rules—are associated with costs, and the goal is to generate a document with high similarity to the target profile with the minimum cost; i.e., by making the smallest number of changes.

There exists another category of approaches where the obfuscation is done on the feature representation of the document, e.g., the n-gram vector representation, and not on the actual document. This category of obfuscation is used to protect the identity of the author while performing another task, such as sentiment analysis. Since the original text remains intact, we consider the literature on this category out of the scope of this work. An example of this work is Weggenmann and Kerschbaum (2018).

Finally, neural-based, obfuscation-specific approaches, e.g., (Emmery et al., 2018; Bo et al., 2021), are still deemed impractical for the authorship obfuscation domain where researchers would attribute this impracticality to the lack of large training datasets which these neural approaches require (Bevendorff et al., 2020).

2.2 Identification

As mentioned earlier, it is important to use a state-of-the-art authorship identification approach to evaluate evasion in authorship obfuscation. In the authorship attribution domain, it is well-established that a cross-topic authorship identification tool should have a realistic performance that mimics real-life applications (Goldstein-Stewart et al., 2009; Sundararajan and Woodard, 2018; Stamatatos, 2017, 2018; Custódio and Paraboni, 2019; Barlas and Stamatatos, 2020, 2021; Altakrori et al., 2021). Because of that, we use a state-of-the-art (Altakrori et al., 2021) cross-topic, authorship identification technique to evaluate evasion of obfuscation techniques namely, (Stamatatos, 2018).

2.3 Evaluating Content Preservation

Evaluating content preservation in text is important even if we value safety (Potthast et al., 2016). This is because people want to maintain their privacy while sharing their opinions freely. Besides obfuscation, content evaluation techniques are applicable to other NLG tasks such as machine translation and summarization, and these techniques fall within one of three groups.

Token-based evaluation metrics depend on token overlap between a source and a target document. Examples of these metrics are METEOR (Banerjee and Lavie, 2005), BLEU (Papineni et al., 2002), and ROUGE-L (Lin, 2004). While these metrics were among the early ones to be used, they have been shown to have a lower correlation with human scores for fact preserving in text summarization (Maynez et al., 2020; Honovich et al., 2021).

With recent advances in representation learning, particularly in word and sentence embeddings, new model-based metrics were adopted where a smaller change in the sentence embedding indicates higher content preservation. Examples of such metrics are the Universal Sentence Encoder (USE) (Cer et al., 2018) and BERTScore (Zhang et al., 2020).

More recently, the summarization community proposed a new, question-answering-based approach to evaluate the content preservation in summarization. The argument for this work is that the content is considered preserved if we can give the same answer to a particular question both before and after summarization. Examples of this work are (Wang et al., 2020) and (Scialom et al., 2021). Here, using such a system, providing feedback to the users of obfuscation techniques would become easier since the unanswered questions and the spans from which the questions are taken can be shown.

3 A Multifaceted Evaluation Framework

Ideally, authorship obfuscation should only modify the author's writing style in a document while retaining all the original information. However, due to the topic–writing style entanglement, modifying the document is likely to cause information loss; i,e., some content is not preserved. Based on that, obfuscation techniques are evaluated in two dimensions: evasion, and content preservation.

In the following subsections, we formally describe obfuscation, evasion and content preservation and we discuss the state of the tools used to evaluate them. Finally, we propose a novel evaluation dimension to characterize a potential side effect of a successful detection evasion namely, misattribution.

3.1 Obfuscation

Let d be a document written by author a^*. To hide their identity, $a*$ uses an obfuscation technique $O : d \rightarrow \hat{d}$ that takes a document d as an input, modifies it, and outputs an obfuscated version of this document, namely, $\hat{d}$ such that $d \neq \hat{d}$.

For example, suppose we have a document d, where d = "The decision caused the team a big loss!", which was written by author $a^* = $ "Q". Next, "Q" uses an obfuscation technique O that modifies the document d by changing it to $\hat{d}$, where $O(d) = \hat{d}$ = "The advice caused the team a huge loss".

3.2 Evading Detection

We use an authorship identification technique to evaluate the performance of authorship obfuscation. If the identification technique was able to identify the original author before obfuscation but failed to identify that author after obfuscation then the obfuscated document has evaded detection.

The evaluation process is as follows. We start by training and tuning an authorship identification tool on the training and validation documents, respectively. Then, we record the identification accuracy on the original test documents. Next, we use an obfuscation technique to modify the test documents to hide the authors' writing styles in theses document. Finally, without further training/fine-tuning to the identification tool, we measure the authorship identification performance on the obfuscated test documents. The effectiveness of an obfuscation technique is quantified by the difference in identification performance before and after obfuscation over all the test documents in the investigated dataset.

Formally, let I be an authorship identification technique $I : (d, T) \rightarrow a_i$ that takes a document d and a set of candidate authors of this document T as input, and outputs a_i as the most plausible author of this document d. Let $T = [a_1, a_2, \ldots, a_n]$ and $n =\mid T \mid$. We say that author a^* has evaded detection using the obfuscation tool O if $I(d, T) = a*$, $I(O(d), T) = a_i$, where $a^* \neq a_j$; and $a^*, a_j \in T$. Note that, if $I(d, T) \neq a*$ then d does not require obfuscation against I.

To evaluate the obfuscation performance over a whole test dataset D, let $S : (a_i, a^*) \rightarrow \mathbb{Z} \in [0, 1]$ be an indicator function given by Eq. 1. Finally, let $Accuracy = \sum_i^m S(I(d_i, T), a_i^*)/m$, where $m =\mid D \mid$ is the number of test documents.

$$S(a_i, a^*) = \begin{Bmatrix} 1, & \text{if } a_i = a^* \\ 0, & \text{otherwise} \end{Bmatrix} \quad (1)$$

Continuing from the example in Sec. 3.1, let T be ["G", "Q", "B", "M", "W"], the predicted author before obfuscation, i.e., I(d, T) be Q, and the predicted author after obfuscation, i.e., $I(O(d), T)$ be G. Here, the obfuscation tool O has evaded detection successfully.

3.3 Preserving the Content

After evaluating evasion, content preservation is evaluated to investigate whether loss of information has occurred due to obfuscation. An authorship obfuscation technique should maximize the content preservation, or equally minimize the loss of information. After evaluation, the result of this evaluation is communicated to the author to decide whether to accept the obfuscation outcome, or reject it if the information loss is drastic.

Formally, let $P : (d, O(d)) \rightarrow \mathbb{R}$ be a content-preservation evaluation tool that takes an original document d and an obfuscated document $O(d)$ as input, compares their content and outputs a content-preservation score that represents the amount of information preserved from the original document d after obfuscation.

For example, suppose that the content-preservation tool of choice is based on the word-level, uni-grams overlap between the original document d, and the obfuscated document $O(d)$, where d = "The decision caused the team a big loss!" and $O(d)$ = "The advice caused the team a huge loss". Suppose that splitting d, and $O(d)$ into word-level uni-grams yields [“The”, “decision”, “caused”, “the”, “team”, “a”, “big”, “loss!”] and [“The”, “advice”, “caused”, “the”, “team”, “a”, “huge”, “loss”], respectively. Here, the goal is to maximize content-preservation score where $P : (d, O(d)) = 5$.

3.4 Fairness, and the Potential of Misattribution Harm

In real-life applications of authorship identification, misattribution can have severe outcomes. For example, if the obfuscated text is a threatening message, then it is important to identify the real culprit to avoid persecuting an innocent person.

Confidence in the identification outcome is a core concept in authorship identification that has not been emphasized in the obfuscation literature.

Instead of imitating a writing style of one of the candidate authors, an obfuscation technique may output a generic writing style that is difficult to attribute to a specific author. In that case, the identification technique will still provide a candidate author, but its confidence in this output would be low. As a result, if an obfuscation technique can lower the confidence of an identification method, then its outcome will be in a position of doubt, hence, neither the original author nor the identified one will have to suffer.

Formally, let C:(I,d,T) $\rightarrow \mathbb{R}^n$ be a tool that takes as input an authorship identification tool I, a document d, and a set of candidate authors for that document T and outputs a probability distribution over the candidate authors $[c_1, c_2, \ldots, c_n]$, where $c_i = P_I(a_i|d)$ is the likelihood of author a_i being the original author of document d when authorship identification tool I is used, $0 \leq c_i \geq 1$, and $\sum_i^n c_i = 1$.

In this work, we consider the model's confidence to be a high when the probability distribution for the one author is much higher compared to the other authors. For example, a model is the most confident when it predicts author A_t with probability 1. In contrast, a model is the least confident, or rather clueless, when the probability distribution over all the authors is uniform, i.e. when the probability of each author is $\frac{1}{T}$, where T is the number of authors.

Note that the model can predict the wrong author and have high confidence in its prediction. Luckily, this confidence can be measured by computing the entropy (Eq. 2) for the attribution model, and the effect of misattribution harm can be characterized by the difference in entropy before and after obfuscating a document (d). While there exists a number of approaches to measure the difference between the two entropy values, such as cross-entropy or KL-divergence, we chose the difference in entropy for simplicity. Other measure could potentially be explored in future work.

$$H(X) = -\sum_{t=1}^{n} P(x_t) \log_2 P(x_t) \qquad (2)$$

Furthermore, this approach can provide a more fine-grained measure of performance than the identification accuracy. For example, let us assume that the attribution model had to identify the most plausible authors from a set of three authors: a_1, a_2, and a_3. Before obfuscation, The model correctly identifies a_2 as the most plausible author with the maximum confidence in its prediction, i.e., the probability distribution over the authors was 1 for a_2 and 0 otherwise.

After obfuscation using technique A, the model identifies a_1 as the most plausible authors, i.e., a_2 has successfully evaded detection. The model, however, outputs a probability distribution of 0.7 for a_2, 0.2 for a_1 and 0.1 a_3 with a high confidence in its prediction. Alternatively, after obfuscation using technique B, the identification model also identifies a_1 as the most plausible author, outputs a probability distribution of 0.4 for a_2, 0.3 for a_1 and finally, 0.3 a_3.

Clearly, both techniques generated the same author prediction, and so, both techniques evaded detection. However, technique B would be considered better because it caused the attribution model to have lower confidence in its prediction.

4 Experimental Setup

Our overall evaluation procedure is as follows. We started by establishing the authorship identification accuracy on the original datasets. Note that, the training and testing split is predefined for each dataset as shown in Table 2. For validation, however, we shuffled the training set and took 20% of the samples for validation.

We followed that by creating different obfuscated copies of the test sets, one for each obfuscation technique. Next, we evaluated the detection evasion and misattribution on each obfuscated copy in one step. We concluded our evaluation with content preservation. We provide below the details of each step separately.

4.1 Corpora

For this work, we use two different corpora: the Extended Brennan–Greenstadt Corpus (EBG) dataset (Brennan et al., 2012) and the Reuters Corpus Volume 1 (RCV1) (Teahan, 2000; Khmelev, 2000; Kukushkina et al., 2001), commonly referred to as C50 dataset. For each dataset, we use two authors configurations: five authors, and 10 authors. We provide corpus statistics in Table 2.

	C50		EBG	
Authors	**5**	**10**	**5**	**10**
Training set				
Docs	75	150	55	110
Docs / authors:	15	15	11	11
Avg. doc len (W)	478	452	496	494
Avg. doc len (C)	3007	2861	3157	3120
Testing set				
Docs	75	150	55	110
Docs / authors:	15	15	7	6
Avg. doc len (W)	480	479	496	497
Avg. doc len (C)	3032	3036	3068	3046
Total docs	150	300	90	169

Table 2: Corpus statistics. (Doc: Document, W: Words, C: Characters) numbers are reported using the rounded mean. SD reported in the appendix, in Table 8)

4.2 Authorship Obfuscation

The evasion performance of an obfuscation technique is compared to a set of baselines as well as state-of-the-art obfuscation techniques. Here, the role of a baseline is to set a lower bound on the performance while requiring little knowledge about the problem, and a fairly low effort to use.

In this work, we use a neural machine translation model in the back-translation baseline to replace statistical models that were used in previous studies (Brennan et al., 2012; Keswani et al., 2016). Additionally, we use a contextual language model namely, BERT to replace words based on their context, instead of replacing them with synonyms or random words from the author's vocabulary set.

Back Translation (BT) uses Facebook's many-to-many translation model (El-Kishky et al., 2020; Fan et al., 2021; Schwenk et al., 2021) implemented by the HuggingFace (Wolf et al., 2020) library [2]. This model has two advantages. Firstly, it is open-source and its results can be replicated in contrast to commercial translation products that are costly and can be replaced at any time.

Secondly, this model translates between languages directly without using English as a reference/pivot language. Many of the existing neural machine translation models use English as a *pivot language* where translation is done either *from* English or *to* English. For example, if the task is the translate from French to Chinese, one has to translate from French to English, then from English to Chinese. This approach defeats the whole point of multi-hop translation where the goal is to use the differences between languages in phrasing the same idea to change the writing style of a sentence.

Lexical Substitution Using BERT (LSB) (Mansoorizadeh et al., 2016) masks random words in a sentence, then use BERT language model to replace these words with ones that fit the context.

Mutant-X (Mahmood et al., 2019) replaces words based on their GloVE word embeddings given that the candidate replacement has the same sentiment. This technique requires knowledge of the authorship attribution classifier, specifically, the probability of each author, to do the obfuscation.

Heuristic Obfuscation Search (A*) (Bevendorff et al., 2019) was originally developed as an imitation approach to obfuscation. The algorithm requires a target author profile which is the tri-grams frequency. This rule-based approach changes the text while incurring costs, and the goal is to generate a document with a high similarity to a target profile with minimum cost.

4.3 Authorship Identification

For authorship identification, we use the state-of-the-art (Altakrori et al., 2021) cross-topic, authorship identification technique to evaluate evasion of obfuscation techniques namely, Masking (Stamatatos, 2018). The main idea of this approach is to mask words in a document, where masking is done by replacing the characters in the word with asterisks, then use word- or character-level n-grams to represent as features. The choice of which words are masked is based on the hyperparameter k[3]. In a document, any word that is not in the k-most frequent words in the British National Corpus (BNC) must be masked. After masking and extracting the n-gram features, a Support Vector Machines (SVM) with linear kernel is used as a classifier.

[2] https://huggingface.co/facebook/m2m100_418M

[3] All the hyperparameters are provided in Appendix A.2.

Anon. Tech.	EBG dataset				C50 dataset			
	5 Authors		10 Authors		5 Authors		10 Authors	
	Acc. (%)	Diff.	Acc.	Diff.	Acc.	Diff.	Acc.	Diff.
None (Original text)	96.4	-	77.6	-	76.0	-	67.3	-
A*	93.5	2.9	71.1	6.5	**72.0**	**4.0**	**64.0**	**3.3**
Back Translation	**84.0**	**12.4**	**64.2**	**13.4**	73.3	2.7	65.3	2.0
Lexical Sub (BERT)	91.5	4.9	78.4	-0.8	76.0	0	67.3	0
Mutant-X	86.4	10.0	73.6	4.0	74.7	1.3	66.7	0.6

Table 3: Obfuscation performance characterized by the change in the identification accuracy (Acc. %) using word masking and character n-grams as features, and a linearSVM classifier. (**Diff.** is the difference between the identification accuracy for the original text and the accuracy after obfuscation. (Lower identification accuracy (higher difference) is better. A negative sign means the accuracy increased instead of decreasing. **Bold**: best result per column.)

4.4 Content Preservation

To evaluate the content preservation, we chose the EBG dataset with the ten authors configuration. From all the original test documents and the four obfuscated versions, we randomly selected 10% of the documents. These documents were split into sentences, and a sentence was included or excluded from the evaluation samples based on a coin flip. This resulted in 212 sampled sentences, an average of 42 sentences per obfuscation techniques. To avoid cherry-picking samples that favor one metric vs. another, we did not exclude any of the sampled sentences. However, we discuss the consequence of this in the results section below.

To evaluate content preservation of these samples, we used HuggingFace implementation for both token-based and model-based evaluation tools. For the question answering approach, we used (Scialom et al., 2021) that generated the questions from the original document instead of needing a reference[4].

In brief, we used BLEU (Papineni et al., 2002), ROUGE-1, 2, and L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), BERTScore (Zhang et al., 2020), and QuestEval (Scialom et al., 2021) to evaluate content preservation. For example, for BLEU, we consider the obfuscation text as the translation of the original text and we report the average BLEU score over the 212 sampled sentences.

Appendix A.6 provides various obfuscated examples for each obfuscation technique. With these examples, we inspect what was each technique good at, and what did it fail to do. A more detailed study on the types of errors made by obfuscation techniques can be found in (Gröndahl and Asokan, 2020).

4.5 Characterizing Misattribution

As described in Sec. 3.4, we calculate the change in entropy before and after training. We follow the same training procedure that was used for identification. However, instead of using the authors' probabilities to find the most likely author we calculate the entropy for that output distribution.

Finally, we normalize the entropy scores to make them comparable with other content-preservation scores that are bounded between zero and 1. To do that, we divide the entropy scores by the entropy of the uniform distribution with K authors, where K is the number of authors in each dataset.

5 Experimental Results

5.1 Evaluating Evasion

As mentioned earlier in Sec. 3.2, the successful evasion of an obfuscation technique is measured by the drop in authorship identification accuracy after obfuscation. In Table 3, the first row shows the identification accuracy on the original test documents, i.e., before obfuscation. The rows below it show the identification accuracy after obfuscating the test documents. Here, the lower the attribution accuracy after obfuscation the better is an obfuscation algorithm at evading detection.

We make the following observations from Table 3. First, despite being a baseline, back translation outperforms both obfuscation techniques on the EBG dataset, and comes as a close second on the C50 dataset after A* of Bevendorff et al.. Contrast to the literature, back translation is not a weak baseline anymore.

The other general observation that we make is that identifying the original author –even without obfuscation– becomes much harder as the number of candidate author increases. Specifically, as the number of authors increased from five authors

[4] The authors made their code available online.

Anon. Tech.	Rouge-1	Rouge-2	Rouge-L	BLEU	METEOR	BERTSc.	QuestE
None (original text[5])	1.000	0.981	1.000	0.981	1.000	1.000	0.678
A*	**0.906**	**0.858**	**0.906**	**0.766**	0.867	0.966	0.582
Back Translation	0.704	0.471	0.681	0.312	0.722	0.958	**0.620**
Lexical Sub (BERT)	0.848	0.696	0.845	0.593	0.844	0.965	0.599
Mutant-X	0.902	0.814	0.902	0.746	**0.915**	**0.976**	0.555

Table 4: Content Preservation scores on 212 sampled sentences from the EBG-10 dataset. (The first row is the score for the original text. Higher is better. **Bold** is for the maximum value per column)

	EBG				C50			
Anon. Tech.	5 Authors		10 Authors		5 Authors		10 Authors	
	Ent.	Diff	Ent.	Diff	Ent.	Diff	Ent.	Diff
None (Original text)	73.9	-	83.3	-	79.4	-	84.3	-
A*	78.8	+4.9	86.8	-3.5	79.0	+0.4	84.4	-0.1
Back Translation	**82.3**	**-8.4**	**88.8**	**-5.5**	**83.1**	**-3.7**	**87.3**	**-3.0**
Lexical Sub (BERT)	80.5	-6.6	84.6	-1.3	80.2	-0.8	85.3	-1.0
Mutant-X	72.6	+1.3	85.2	-1.9	82.0	-2.6	83.2	+1.1

Table 5: Characterizing the misattribution using the normalized entropy score (%) (Ent. is the normalized entropy score. **Diff.** is the difference between the entropy score for the original text and the entropy after obfuscation. Higher entropy (Lower diff) is better. **Bold** is the best value per column.)

to ten authors, the authorship attribution accuracy dropped by around %20 and %10 on the EBG and the C50 dataset, respectively. We conduct more analysis on the robustness of the evaluated obfuscation techniques in Sec. 6. Specifically, we use different identification techniques, with various writing style features, and report the results of this analysis in Table 6.

5.2 Content Preservation

Table 4 shows the result of content preservation using various evaluation metrics. Naturally, A* has the best performance on the token-based metrics given that most of the modifications are done at the character level, i.e., has lower tendency to change words. Similarly, Mutant-X has the highest model-best scores because words are replaced based on their embeddings.

Conversely, back translation has the worst scores in both token-based an model-based measures. In contrast, it has the closest score to the original text using the QA-based approach which, as mentioned earlier, has better correlation with human scores than token-based and model-based metrics.

We manually investigated the quality of sentences that were obfuscated using the back translation technique (See Appendix A.6). In these sentences, one can see that back translation rephrases the sentence and maintains the original content despite using different tokens. This is a clear indication that the QA-based approach is more trustworthy to measure the content preservation than the commonly used token-based approaches.

5.3 Characterizing Unfair Misattribution Using Entropy

Table 5 shows the normalized entropy scores[6] that are used to characterize unfair misattribution. The higher the normalized entropy, the closer the probability distribution of the predicted authors is to the uniform distribution. In that case, the model has no preference for one particular author, or has low confidence in its outcome. Back translation has the best performance on this evaluation metric, measured by the increase in normalized entropy from that before obfuscation (the first row). Our interpretation is that by translating to different languages, back translation is generating text in a generic style that is hard to attribute to one particular author. In contrast, A* tries to imitate a specific author's writing style to avoid detection, while Mutant-X requires a set of candidate authors and a classifier to do the obfuscation, and only stops when the obfuscated text is attributed to a different author.

6 Analysis of Robustness

In this section, we conduct a battery of tests on different attribution features. The goal of this study

[5] We investigated the sentences sampled for evaluation to see why original text does not have a perfect score (1.0) on all token-based metrics. We noticed that one sentence had only one word, that is "originally,". As a result, Rouge-2 could not find any bi-grams in this sentence and output a score of 0.

[6] Table 9 in Appendix A shows the normalized entropy scores with standard deviation.

		Identification technique						
		Stylo.		N-grams		Masking		**Average**
	Anon. tech.	-	POS	Ch.	W.	Ch.	W.	
EBG 5	No Anon.	81.2	90.0	91.5	96.4	88.8	96.4	90.7 ± 5.2
	A*	51.3	78.0	91.5	94.6	76.4	93.5	80.9 ± 15.1
	Back Translation	59.7	67.3	89.7	96.4	83.1	84.0	80.0 ± 12.7
	Lexical Sub (BERT)	68.2	80.2	89.7	96.4	95.3	91.5	86.9 ± 9.9
	Mutant-X	81.2	84.7	91.5	96.4	95.5	86.4	89.3 ± 5.6
EBG 10	No Anon.	58.8	53.9	73.3	75.1	53.1	77.6	65.3 ± 10.3
	A*	47.8	37.2	74.2	71.8	51.7	71.1	59.0 ± 14.1
	Back Translation	46.2	43.5	66.3	73.2	59.5	64.2	58.8 ± 10.7
	Lexical Sub (BERT)	55.0	52.0	71.5	73.7	58.7	78.4	64.9 ± 10.1
	Mutant-X	55.0	52.2	74.1	77.0	52.3	73.6	64.0 ± 11.0
C50 5	No Anon.	65.3	68.0	84.0	84.0	61.3	76.0	73.1 ± 8.9
	A*	65.3	66.7	81.3	85.3	58.7	72.0	71.6 ± 9.2
	Back Translation	64.0	65.3	82.7	81.3	58.7	73.3	70.9 ± 9.0
	Lexical Sub (BERT)	65.3	68.0	85.3	88.0	62.7	76.0	74.2 ± 9.7
	Mutant-X	60.0	62.7	84.0	84.0	54.7	74.7	70.0 ± 11.6
C50 10	No Anon.	58.7	69.3	69.3	64.0	53.3	67.3	63.6 ± 5.9
	A*	56.7	69.3	68.0	61.3	54.7	64.0	62.3 ± 5.4
	Back Translation	55.3	64.7	65.3	62.0	52.0	65.3	60.8 ± 5.2
	Lexical Sub (BERT)	56.7	70.7	68.7	62.0	54.7	67.3	63.4 ± 6.0
	Mutant-X	56.0	67.3	69.3	62.7	52.7	66.7	62.4 ± 6.1

Table 6: Obfuscation performance using different sets of features with a Support Vector Machines classifier. The colored row represents the identification accuracy on the original text.

is to characterize the obfuscation performance under different types of writing style features that vary between stylometric features and content features. The results are shown in Table 6. As can be shown, the performance of obfuscation techniques varies drastically based on the choice of obfuscation technique. Because of that, it is important to evaluate a proposed technique against authorship identification techniques with different feature representations.

7 Conclusion

In this work, we demonstrated the importance of using state-of-the-art evaluation tools to measure the performances of authorship obfuscation techniques. In addition, our experiments revealed that current obfuscation techniques have key weaknesses and have been outperformed by a baseline, namely back translation in multiple evaluation aspects. Furthermore, we identified a critical issue with respect to the fairness of obfuscation techniques. Our proposed misattribution measure investigates the side-effect of a successful detection evasion by identifying another author as the most plausible author of the obfuscated text. As a result, we argue that an attack on the confidence of the identification model, by generating text in a generic style would confuse the identification model and make it unusable in real-life applications. Finally, we argue that evaluation of authorship obfuscation tools should follow the rapidly evolving domain of evaluation tools while keep the potential users and real-life applications when developing and evaluating novel obfuscation techniques.

Acknowledgments

We would like to thank the reviewers for their valuable discussion during the rebuttal period.

The first author is supported by the Doctoral Scholarship from Fonds de Recherche du Quebec Nature et Technologies (FRQNT-275545). In addition, this research is supported in part by the Discovery Grants (RGPIN-2018-03872) and CREATE Grants (CREATE-554764-2021) from the Natural Sciences and Engineering Research Council of Canada, and Canada Research Chairs Program (950-230623). The fourth author is supported by a Canada CIFAR AI Chair.

8 Limitations

One potential limitation of this work is that obfuscation can be misused in a similar way that authorship identification can be misused. However, it is important that the public be aware of the existence of such tools, and for researchers to have better obfuscation techniques to raise the bar for identification techniques. Another limitation is that we could have used more datasets in our analysis. We note that, our results —particularly, where a baseline outperforms state-of-the-art obfuscation techniques— would still be interesting regardless off the number of datasets. In addition, all the datasets for authorship obfuscations similar characteristics in terms of size.

Another potential limitation of this work is the lack of human evaluation for content preservation. While question answering approaches have been shown to correlate well with human evaluation scores for factual consistency, it would have been interesting to analyze the cases when such techniques fail. In particular, what type of errors do such techniques make? For example, do these techniques produce ungrammatical sentences or generate grammatical but nonsensical sentences.

References

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.

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/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Georgios Barlas and Efstathios Stamatatos. 2020. Cross-domain authorship attribution using pre-trained language models. In *Artificial Intelligence Applications and Innovations*, pages 255–266, Cham. Springer International Publishing.

Georgios Barlas and Efstathios Stamatatos. 2021. A transfer learning approach to cross-domain authorship attribution. *Evolving Systems*, pages 1–19.

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, Florence, Italy. Association for Computational Linguistics.

Janek Bevendorff, Tobias Wenzel, Martin Potthast, Matthias Hagen, and Benno Stein. 2020. On divergence-based author obfuscation: An attack on the state of the art in statistical authorship verification. *it-Information Technology*, 62(2):99–115.

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.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. USE: Universal sentence encoder for english. In *Proc. of the 2018 conference on empirical methods in natural language processing: system demonstrations (EMNLP)*, pages 169–174.

José Eleandro Custódio and Ivandré Paraboni. 2019. An ensemble approach to cross-domain authorship attribution. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 201–212. Springer.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzmán, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5960–5969, Online. Association for Computational Linguistics.

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.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep

Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, et al. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Jade Goldstein-Stewart, Ransom Winder, and Roberta Sabin. 2009. Person identification from text and speech genre samples. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 336–344, Athens, Greece. Association for Computational Linguistics.

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

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. q^2: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gary Kacmarcik and Michael Gamon. 2006. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 444–451, Sydney, Australia. Association for Computational Linguistics.

Yashwant Keswani, Harsh Trivedi, Parth Mehta, and Prasenjit Majumder. 2016. Author masking through translation. In *CLEF (Working Notes)*, pages 890–894.

Dmitry V Khmelev. 2000. Disputed authorship resolution through using relative empirical entropy for markov chains of letters in human language texts. *Journal of quantitative linguistics*, 7(3):201–207.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Olga V Kukushkina, Anatoly A Polikarpov, and Dmitry V Khmelev. 2001. Using literal and grammatical statistics for authorship attribution. *Problems of Information Transmission*, 37(2):172–184.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Asad Mahmood, Faizan Ahmad, Zubair Shafiq, Padmini Srinivasan, and Fareed Zaffar. 2019. A girl has no name: Automated authorship obfuscation using mutant-x. *Proceedings on Privacy Enhancing Technologies*, 2019(4):54–71.

Muharram Mansoorizadeh, Taher Rahgooy, Mohammad Aminiyan, and Mahdy Eskandari. 2016. Author obfuscation using wordnet and language models—notebook for pan at clef 2016. In *CLEF 2016 Evaluation Labs and Workshop–Working Notes Papers*, pages 5–8.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark. Association for Computational Linguistics.

James O'Shea. 2013. Alphabetical order 277 word new function word list. `https://semanticsimilarity.files.wordpress.com/2013/08/jim-oshea-fwlist-277.pdf`. [Retrieved Oct. 2019].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Martin Potthast, Matthias Hagen, and Benno Stein. 2016. Author obfuscation: Attacking the state of the art in authorship verification. In *CLEF (Working Notes)*, pages 716–749.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6490–6500, Online. Association for Computational Linguistics.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, Jacopo Staiano, Alex Wang, and Patrick Gallinari. 2021. QuestEval: Summarization asks for fact-based evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6594–6604, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Efstathios Stamatatos. 2017. Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1138–1149, Valencia, Spain. Association for Computational Linguistics.

Efstathios Stamatatos. 2018. Masking topic-related information to enhance authorship attribution. *Journal of the Association for Information Science and Technology*, 69(3):461–473.

Kalaivani Sundararajan and Damon Woodard. 2018. What represents "style" in authorship attribution? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2814–2822, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

W. J. Teahan. 2000. Text classification and segmentation using minimum cross-entropy. In *Content-Based Multimedia Information Access - Volume 2*, RIAO '00, page 943–961. Le Centre de Hautes Etudes Internationales D'Informatique Documentaire, Paris, FRA.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Benjamin Weggenmann and Florian Kerschbaum. 2018. Syntf: Synthetic and differentially private term frequency vectors for privacy-preserving text mining. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 305–314. ACM.

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.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *ArXiv preprint*, abs/1609.08144.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Appendices

A.1 Hardware and Runtime

The experiments for this paper where run on a workstation with one GPU type Quadro RTX 8000, with four CPUs and 32GB of RAM. Run time (estimated by wandb.com) is as follows.

1. Obfuscation run-time: ∼10 days, that is ∼256 Hrs total.

2. Authorship identification run time: ∼0.6 day, that is ∼14.5 Hrs total.

A.2 Hyperparameters

Table 7 shows the ranges of hyperparameters that were used for Masking (the main identification technique) and the other writing style features that were used in the ablation study in Table 6.

Hyperparameter	Range
k	100, 200, 300, 400, 500, and 1000, 2000, 3000, 4000, 5000
f_t	5, 10, 15, 20, 25, 30, 35, 40, 45, 50
n_{ch}	3, 4, 5, 6, 7, 8
n_w	1, 2, 3
$epochs$	2, 5
$vocab\_size$	2000, 5000

Table 7: Hyperparameters for masking and n-gram based feature representations. k is the threshold for masking, n_w is the word-level and POS n-grams, n_{ch} is the character-level n-gram, and f_t is the minimum frequency threshold in the whole dataset.

A.3 Corpus Statistics, with Mean and SD

See Table 8.

A.4 Misattribution Harm

Table 9 shows the normalized entropy scores (with SD) while Table 10 shows the identification accuracy on the left of the table and unnormalized entropy scores to characterize the misattribution behavior. The goal of this table is to show that raw entropy scores are less intuitive than the normalized values bound between zero and one.

A.5 Stylometric Features

Table 11 shows details of the static, stylometric features that were used in the ablation study.

	C50				EBG			
Authors	**5**		**10**		**5**		**10**	
Training set								
Docs	75		150		55		110	
Docs / authors:	15	(0.0)	15	(0.0)	11	(0.0)	11	(0.0)
Avg. doc Len (W)	478	(46.4)	452	(60.8)	496	(6.1)	494	(4.8)
Avg. doc Len (C)	3007	(273.1)	2861	(366.9)	3157	(24.0)	3120	(41.8)
Testing set								
Docs	75		150		55		110	
Docs / authors:	15	(0.0)	15	(0.0)	7	(4.0)	6	(3.2)
Avg. doc Len (W)	480	(86.2)	479	(77.6)	496	(14.1)	497	(12.5)
Avg. doc Len (C)	3032	(567.2)	3036	(473.9)	3068	(102.7)	3046	(130.8)
Total docs	150		300		90		169	

Table 8: Corpora statistics. (Mean and SD)

Anon. Tech.	EBG		C50	
	5 Authors	10 Authors	5 Authors	10 Authors
None (Original text)	73.9 ± 4.4	83.3 ± 3.8	79.4 ± 6.8	84.3 ± 2.9
A*	78.8 ± 4.9	86.8 ± 4.0	79.0 ± 6.5	84.4 ± 2.1
Back Translation	**82.3 ± 4.3**	**88.8 ± 2.1**	**83.1 ± 4.8**	**87.3 ± 3.1**
Lexical Sub (BERT)	80.5 ± 4.5	84.6 ± 2.5	80.2 ± 6.6	85.3 ± 2.4
Mutant-X	72.6 ± 4.7	85.2 ± 2.8	82.0 ± 5.4	83.2 ± 2.6

Table 9: Characterizing the misattribution using the normalized entropy score (%).

	Identification accuracy (%)				Entropy			
	EBG		C50		EBG		C50	
Anon. T.	5 Au.	10 Au.	5 Au.	10 Au.	5 Au.	10 Au.	5 Au.	10 Au.
None	96.4	77.6	76.0	67.3	1.72 ± 0.1	2.77 ± 0.1	1.84 ± 0.2	2.80 ± 0.1
A*	93.5	71.1	72.0	64.0	1.83 ± 0.1	2.88 ± 0.1	1.83 ± 0.1	2.81 ± 0.1
Back T.	84.0	64.2	73.3	65.3	1.91 ± 0.1	2.95 ± 0.1	1.93 ± 0.1	2.90 ± 0.1
LS BERT	91.5	78.4	76.0	67.3	1.87 ± 0.1	2.81 ± 0.1	1.86 ± 0.2	2.84 ± 0.1
Mutant-X	86.4	73.6	74.7	66.7	1.69 ± 0.1	2.83 ± 0.1	1.90 ± 0.1	2.76 ± 0.1

Table 10: Identification accuracy (left) and Misattribution harm (right) characterized by raw entropy score.

<table>
<tr><th>Lexical Features - Character-Level</th><th>Lexical Features - Word-Level</th></tr>
<tr><td>1. Characters count (N)
2. Ratio of digits to N
3. Ratio of letters to N
4. Ratio of uppercase letters to N
5. Ratio of tabs to N
6. Frequency of each alphabet (A-Z), ignoring case (26 features)
7. Frequency of special characters: <>%|{ } []/\@#˜ +-*=$ˆ &_()’ (24 features).</td><td>1. Tokens count (T)
2. Average sentence length (in characters)
3. Average word length (in characters)
4. Ratio of alphabets to N
5. Ratio of short words to T (a short word has a length of 3 characters or less)
6. Ratio of words length to T. Example: 20% of the words are 7 characters long. (20 features)
7. Ratio of word types (the vocabulary set) to T</td></tr>
<tr><th colspan="2">Syntactic Features</th></tr>
<tr><td colspan="2">1. Frequency of Punctuation: , . ? ! : ; ’ " (8 features)
2. Frequency of each function words (O’Shea, 2013) (277 features)</td></tr>
</table>

Table 11: List of stylometric features.

A.6 Qualitative Analysis

In this section, we provide examples for each obfuscation system, and comment on what each one does. Note that the examples were cherry-picked in order to highlight the different issues in each approach.

Tables 12 to 15 provide examples for A*, Mutant-X, back translation, and lexical substitution with BERT, respectively. A more detailed study on categorizing the types of errors made by an obfuscation technique can be found in (Gröndahl and Asokan, 2020).

Original	The decline of the Kongo due to a series of wars with the Portuguese in the seventeenth century,
Modified	The decrease of the Kongo due to ab polynomial of wars with the Portuguese in the seventeenth week,
Original	The continued fragmentation of its nationalist movements set Angola apart from other Portuguese colonies.
Modified	The continued fragmentation of its nationalist movements set Angola apart from other Portuguese colonies!
Original	The oppressive tropical climate and hostile African neighbors made life difficult for settlers, many of whom lacked agricultural experience or expertise.
Modified	The opperss tropical control and troops African neighbors made life difficult fsettlers, many of whom lacked agricutlural bonanza ro xepertise.
Changes	Word replacement, punctuation replacement, flipping characters and introducing typos.
Observation	In some cases, the replaced words fit the context to some extent. In other cases, the new words were completely out of context. This is mainly because word replacement did not consider the whole sentence but rather the word to be replaced.

Table 12: Obfuscated examples generated using the A* obfuscation technique (Bevendorff et al., 2019)

Original	Protect personal information with the MyID identity theft monitoring solution.
Modified	Protect personal info with the MyID identity theft monitoring solution.
Original	The possibility that Internet users will be able to hide what they do from the ubiquitous ad tracking is a big win for consumers concerned with Internet privacy.
Modified	The prospect that Internet users will be able to hide what they do from the ubiquitous ad tracking is a big victorious for consumers concerned with Internet privacy.
Original	The other example was that of a woman who had fallen and broken her arm.
Modified	The other example was that of a schoolgirl who had fallen and broken her arm.
Changes	Controlled word replacement that is based on the sentiment of the word to be replaced.
Observation	Similar to other techniques that use word replacement, sometimes the replaced word either have the wrong part of speech, or changes the meaning. It does better than naive word replacement techniques because of the added rules on candidate words.

Table 13: Obfuscated examples generated using Mutant-X (Mahmood et al., 2019)

Original	Some of the relevant items with regard to maintaining and strengthening health systems include:Neither side purposely disrupted health systems during the conflict.
Modified	Some of the relevant points regarding the preservation and strengthening of health systems are: no side that targeted health systems during the conflict.
Original	Zimbabwe lost over two thirds of their physicians in the 1990s.
Modified	Zimbabwe lost more than two-thirds of its doctors in the 1990s.
Original	The initial reasons for United States intervention in Angola were primarily economic.
Modified	The initial reasons for the U.S. intervention in Angola were mainly economic.
Original	Two important US officials in Luanda, Robert W. Hultslander, the CIA station chief, and Tom Killoran, the American Consul General, agreed that ...
Modified	Two main U.S. officials in Luanda, Robert W. Hultslander, the CIA head of state, and Tom Killoran, the U.S. consulate, accepted that ...
Original	Over half of Cuba's doctors left during the revolution.
Modified	More than half of the Cuban doctors were abandoned during the revolution.
Original	Since 1961, the US had been supporting Holden Roberto with a modest stipend of $10,000 a year.
Modified	Since 1961, the United States has supported Holden Roberto with a modest stock exchange of $10,000 per year.
Changes	Word replacement, rephrasing sentences, contracting/expanding acronyms, and adding spaces.
Observation	Back translation is a powerful text generation tool. Rephrasing a sentence implicitly replaces some words with synonyms that fit the context, and in some cases changes the grammatical structure of the sentence as well. In addition, expanding an acronym, e.g., replacing US with United States, or vise versa or adding proper spacing might hide some writing habits of the author.

Table 14: Obfuscated examples generated using back translation (Fan et al., 2021)

Original	Du Mortier and Coninx describe their use of MHUs with the International Committee of the Red Cross during the conflict in Columbia in 2005.
Modified	Du Mortier and Coninx describe his use of MHUs with the International Committee of the white Cross during the conflict in maryland in 2005.
Original	They are generally expensive, however, and the fact that they only provide services intermittently tends to affect when they are appropriate for use.
Modified	They are not expensive, however, so the fact that they can provide services intermittently tends toward affect when they are appropriate for use.
Original	Sloppy dressers generally look as if they slept in the clothes they are wearing.
Modified	Sloppy dressers generally look as if they slept on the clothes they are wearing.
Changes	Word replacement.
Observation	As shown in the examples, word replacement, even when the context is considered, sometimes lead to choosing the wrong word. Here, the chosen word fits the context, i.e. sensible, but changes the meaning compared to the original sentence.

Table 15: Obfuscated examples generated using Lexical Substitution using BERT (Mansoorizadeh et al., 2016)