Improving Authorship Privacy: Adaptive Obfuscation with the Dynamic Selection of Techniques

Hemanth Kandula **Damianos Karakos** **Haoling Qiu** **Brian Ulicny**
RTX BBN Technologies
{hemanth.kandula, damianos.karakos, haoling.qiu, brian.ulicny}@rtx.com

Abstract

Authorship obfuscation, the task of rewriting text to protect the original author's identity, is becoming increasingly important due to the rise of advanced NLP tools for authorship attribution techniques. Traditional methods for authorship obfuscation face significant challenges in balancing content preservation, fluency, and style concealment. This paper introduces a novel approach, the Obfuscation Strategy Optimizer (OSO), which dynamically selects the optimal obfuscation technique based on a combination of metrics including embedding distance, meaning similarity, and fluency. By leveraging an ensemble of language models OSO achieves superior performance in preserving the original content's meaning and grammatical fluency while effectively concealing the author's unique writing style. Experimental results demonstrate that the OSO outperforms existing methods and approaches the performance of larger language models. Our evaluation framework incorporates adversarial testing against state-of-the-art attribution systems to validate the robustness of the obfuscation techniques. We release our code publicly at https://github.com/BBN-E/ObfuscationStrategyOptimizer

1 Introduction

The digital age has brought about profound changes in how information is created, shared, and analyzed. One critical aspect of this transformation is the increasing capability to attribute texts to their authors using powerful authorship attribution systems by analyzing text style alone (Abbasi and Chen, 2008; Narayanan et al., 2012; Rivera-Soto et al., 2021). These create both opportunities and challenges, particularly when they intersect with issues of privacy and anonymity. Authorship obfuscation seeks to address these challenges by modifying a text's stylistic features to prevent the identification of its author. The need for such measures spans various domains, from protecting journalists and political dissidents against persecution to preserving anonymity in peer review processes. The primary goal is to protect the public from potential abuses of authorship attribution techniques, which could stifle free speech or target whistleblowers. Authorship obfuscation involves strategically altering writing style to obscure stylistic signatures that might trace the text back to its author, thereby protecting their identity (Kacmarcik and Gamon, 2006). The challenge lies in concealing the author's style without compromising the text's content integrity.

Current approaches to authorship obfuscation vary widely, from using large language models (LLMs) like ChatGPT, which, while powerful, require substantial computational resources and potentially compromise privacy if proprietary data retention is involved. On the other end of the spectrum are more localized, machine translation system (Keswani et al., 2016), rule-based systems (Karadzhov et al., 2017) or iterative-change algorithms (Mahmood et al., 2019) that often struggle with the dual demands of effective obfuscation and content preservation. More recently (Fisher et al., 2024), on the other hand, proposed an inference-time algorithm that utilizes constrained decoding for author anonymity, providing flexibility and user-specified control. (Hallinan et al., 2023) proposed a style transfer method that effectively adjusts styles from arbitrary sources to target styles while preserving content. Each exhibits diverse strengths and weaknesses due to variations in data, architectures, and hyperparameters, making them complementary to each other. Therefore, it is important to dynamically ensemble these systems to generate consistently better obfuscation for each input. Considering the diverse strengths and weaknesses of these methods, it is crucial to develop an ensembling method that harnesses their complementary potentials.

We introduce the Obfuscation Strategy Optimizer (OSO), an ensemble-based approach de-

Proceedings of the Fifth Workshop on Privacy in Natural Language Processing, pages 137–142

signed to dynamically select the optimal obfuscation strategy that aligns with users' needs. Users will be able to leverage OSO over outputs from different kinds of obfuscation systems to optimize the trade-off between style concealment and content preservation while preserving the semantic integrity and readability of the original text. OSO can operate over many other obfuscation systems and align with new users' needs with very small configuration efforts.

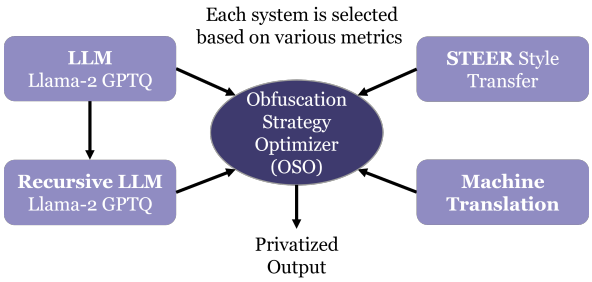

Figure 1: Overview of obfuscation strategy optimizer

2 Approach: Obfuscation Selection for Authorship Obfuscation

We propose a novel approach leveraging an Obfuscation Strategy Optimizer (OSO) to improve authorship obfuscation. The OSO dynamically selects the most effective obfuscation method from a set of available techniques based on specific metrics. This approach addresses the inherent challenges of authorship obfuscation, maintaining content integrity, ensuring fluency, and concealing the author's style. The OSO offers a flexible and adaptive solution that can be applied in real time, making it suitable for diverse applications where privacy and authorship concealment are paramount. The OSO operates by evaluating multiple candidate obfuscations for a given text and selecting the one that optimally balances privacy, sense preservation, and fluency. The candidate obfuscations are generated using various methods, including language models and style transfer techniques, as delineated in Figure 1. The selection process is guided by a combination of quantitative metrics that assess the quality of each obfuscation along the dimensions of author embedding distance, meaning similarity, and fluency.

The OSO evaluates each candidate obfuscation using the following metrics:

Privacy is measured using the cosine distance of LUAR Authorship Attribution model AA (Rivera-Soto et al., 2021) embeddings from the original y_{orig} and obfuscated y_{obf} texts. Higher values indicate greater stylistic divergence, which is desirable for privacy:

$$AADist_{system_i} = C_D(AA(y_{orig}), AA(y_{obf_i})) \quad (1)$$

Meaning Similarity between the y_{orig} and y_{obf} is measured using embedding distance generated with SentenceTransformers (Reimers and Gurevych, 2019). Higher similarity scores indicate better preservation of the original content's meaning. Document meaning similarity is determined by the average of sentence similarity,

$$MS_{system_i} = SBERT(y_{orig}, y_{obf_i}) \quad (2)$$

Fluency is calculated by two metrics first one evaluates the grammatical correctness $CoLA$ of the obfuscated text y_{obf} using a binary RoBERTa-large classifier trained on the CoLA dataset (Warstadt et al., 2019) Eq. 3. The second one was measured using the perplexity PPL Eq. 4 of the text, computed with GPT-2 large [1]. Texts with higher grammatical scores and lower perplexity are more fluent and natural-sounding.

$$CoLA_{system_i} = CoLA(y_{obf_i}) \quad (3)$$

$$PPL_{system_i} = Perplexity(y_{obf_i}) \quad (4)$$

The OSO combines the above metrics into a single objective function to select the best obfuscation candidate per each author. The selection metric for n docs of author a is given by:

$$\text{OSO}_a = \arg\max_{\text{system}_i} \left(\frac{1}{n} \sum_{\text{doc}}^{n} \begin{pmatrix} \log(AADist_i) \\ +\log(MS_i) \\ +\log(CoLA_i) \\ -\log(PPL_i) \end{pmatrix} \right) \quad (5)$$

3 Experiments & System Evaluation

We conducted experiments to evaluate the performance of the OSO compared to individual obfuscation methods. These experiments were designed to measure the effectiveness of OSO in preserving content, ensuring fluency, and achieving style concealment. We used a diverse dataset comprising texts from multiple authors to assess the OSO's generalizability. We also compared it with existing baseline authorship obfuscation methods such as Mutant-X (Mahmood et al., 2019) and JamDec (Fisher et al., 2024).

[1] https://huggingface.co/openai-community/gpt2-large

Datasets	Methods	Privacy		Meaning		Fluency
		AADist	Δ Acc.	SBERT	METEOR	COLA
AMT	Original	0.0	0.0	1.0	1.0	0.88
	Mutant-X	-	0.39	-	**0.84**	0.53
	JamDec	-	0.41	-	0.61	0.79
	Machine Translation	0.2133	0.29	0.64	0.75	0.72
	STEER Style Transfer	0.1976	0.30	0.64	0.50	0.76
	Llama-2 7B	0.1955	0.31	**0.87**	0.36	0.91
	Recursive Llama-2 7B	0.2087	0.42	0.85	0.35	0.92
	OSO (proposed)	**0.2441**	**0.43**	0.86	0.42	**0.94**
BLOG	Original	**0**	0.0	1.0	1.0	0.78
	Mutant-X	-	0.44	-	**0.55**	0.47
	JamDec	-	0.32	-	0.53	0.74
	Machine Translation	0.3184	0.25	0.58	0.48	0.70
	STEER Style Transfer	0.4202	0.32	0.57	0.45	0.90
	Llama-2 7B	0.3726	0.49	**0.81**	0.35	0.88
	Recursive Llama-2 7B	0.4335	0.33	0.78	0.31	0.89
	OSO (proposed)	**0.4416**	**0.51**	0.78	0.32	**0.90**

Table 1: Performance comparison of various obfuscation methods on AMT and Blog datasets.

Methods	Privacy		Meaning		Fluency
	AADist	EER	SBERT	METEOR	COLA
Original	0.0	0.0340	1.0	1.0	0.82
Machine Translation	0.2462	0.0817	0.68	**0.48**	0.72
STEER Style Transfer	0.2075	0.0885	0.63	0.47	0.78
Llama-2 7B	0.3242	0.1742	0.65	0.37	0.90
Recursive Llama-2 7B	**0.3427**	0.1857	**0.77**	0.36	0.91
OSO (proposed)	0.3347	**0.2058**	**0.77**	0.37	**0.93**

Table 2: Performance comparison of various obfuscation methods on the HRS dataset.

3.1 Obfuscation Candidates

The Obfuscation Strategy Optimizer dynamically selects the optimal obfuscation method from multiple candidates based on preserving the meaning, and maintaining the fluency while picking the output to maximize the preservation of anonymity. The candidates generated include:

Machine Translation: We adapted sequence-to-sequence models, initially developed for machine translation, by training them on parallel data generated prompting Llama-2 to restyle original texts. We utilized the Fairseq toolkit (Ott et al., 2019) to train transformer-based models.

STEER Style Transfer: The second candidate uses STEER (Hallinan et al., 2023) to rewrite the text in the style of a specific domain, such as Twitter. This approach leverages style transfer to embed the text within a different stylistic context, thereby obfuscating the original author's style.

LLM Rewriting: We paraphrase the original text using an LLM, specifically the Llama-2 7B model (Touvron et al., 2023), optimized through GPTQ quantization (Frantar et al., 2022). This quantization process reduces the model size dramatically from 38GB to 3.4GB, while the runtime on the entire document is decreased from approximately 4 minutes to just about 30 seconds on an Nvidia V100 GPU. This quantization not only increases the processing speed but also reduces the resource consumption significantly, making it far more efficient compared to larger models like those in the ChatGPT.

Recursive LLM Rewriting: The final candidate involves a recursive approach where the output of the initial LLM rewrite is further rewritten by LLM. This double-layer obfuscation aims further to distance the text from the original author's style.

3.2 Datasets

We conducted our experiments using three datasets to evaluate the performance of the Obfuscation Strategy Optimizer (OSO) in various contexts. The datasets include the Extended Brennan–Greenstadt Corpus (EBG) (Brennan et al., 2012), the Blog Authorship Corpus (Schler et al., 2006), and the HRS-HIATUS research datasets.

The **Extended Brennan–Greenstadt Corpus (EBG)** (Brennan et al., 2012) is a collection of short paragraphs gathered from Amazon Mechanical Turk (AMT), used for tasks involving authorship attribution and obfuscation due to its diverse range of writing styles and topics. We used the 10-author version of the EBG dataset.

The **Blog Authorship Corpus** (Schler et al., 2006) consists of diary-style entries from blog.com, featuring a broad array of personal writing styles. We used 10-author versions of the dataset, respectively. This corpus is valuable for evaluating obfuscation techniques in more informal and varied writing styles.

The **HRS-HIATUS Research datasets**, derived from the IARPA HIATUS program[2], are specifically designed to address the dual challenges of authorship attribution and privacy preservation across various genres. These datasets encompass a wide range of sources, including BoardGameGeek, Instructables, GlobalVoices, and StackExchange (both liberal arts and STEM). They feature writings from a diverse group of 114 authors and include 885 query documents, which are texts whose authorship is to be determined, alongside 16k authors and 37k candidate authors, who are potential matches in the attribution process, across all collections. This variety in content sources, along with the inclusion of both genre-specific and cross-genre datasets, provides comprehensive coverage for evaluating authorship obfuscation strategies. Further details are discussed in Appendix A. For summary statistics, see Table 3.

3.3 Evaluation Metrics

We evaluate all methods using automated metrics to assess privacy preservation, content preservation, and fluency. For privacy, we use LUAR embedding distance from 2 and the drop rate in accuracy Δ Acc., which measures the average percentage drop in correctly attributing obfuscated text to the true author compared to the original text using the AAmodel from (Mahmood et al., 2019). Additionally, for the HRS dataset, we use the equal error rate (EER), which is the point where the false acceptance rate (FAR) equals the false rejection rate (FRR), providing a single measure of the system's overall accuracy in distinguishing between authors.

For content preservation, we use the METEOR score (Banerjee and Lavie, 2005) between the original and obfuscated text, which evaluates token overlap. As a second metric, we use the SentenceBERT-based cosine similarity (Eq. 2). While this metric evaluates the semantic closeness and token overlap between the original and obfuscated texts, they do not inherently guarantee the preservation of factual accuracy.

For fluency, we evaluate using the CoLA model on grammatical correctness as described in 2 Eq. 3 Some of these metrics are used in OSO for privacy preservation, content preservation, and fluency. As shown in Figure 1, the metrics are used to select the best system for each author.

3.4 Results

Table 1 and Table 2 summarize the performance of various obfuscation methods, including OSO, across key metrics such as author embedding distance (AADist), meaning similarity, and fluency. The results highlight OSO's superior ability to effectively balance these metrics. Unlike individual methods that may excel in one aspect but falter in others, OSO consistently ensures high levels of style concealment, content preservation, and text fluency by dynamically selecting the most suitable obfuscation method for each text instance. For instance, while the LLM approach in the AMT and Blog datasets achieves a high meaning similarity score, it does so at the expense of privacy, evidenced by lower AADist and Δ Acc. scores compared to OSO. Similarly in the HRS dataset, OSO surpasses other methods by achieving the highest EER for privacy, the highest meaning similarity according to SBERT, and the highest fluency with the Cola score. This not only demonstrates the best balance of privacy and content preservation but also the highest fluency scores. This shows OSO's effectiveness in providing a balanced approach to text obfuscation across different datasets, leveraging the strengths of various techniques while mitigating their limitations. Additionally, it is worth noting that content semantics can be preserved without direct token overlap through the use of synonyms,

[2]https://www.iarpa.gov/research-programs/hiatus

and SBERT effectively captures such content similarities compared to METEOR

4 Conclusion

In this work, we proposed a novel Obfuscation Strategy Optimizer (OSO) to improve authorship obfuscation. By leveraging multiple obfuscation techniques and dynamically selecting the most effective one based on a set of well-defined metrics, the OSO offers a robust and flexible solution to protect authorship privacy. Our experimental results highlight the efficacy of the OSO in maintaining content integrity and fluency while effectively obfuscating the author's style. Future work will involve expanding the OSO with additional obfuscation techniques and further refining the algorithm. We aim to explore more scalable optimization methods, such as heuristic searches and reinforcement learning-based strategies, to improve the OSO's performance and efficiency.

Limitations

While OSO demonstrates promising results, there are a few limitations. Firstly, OSO's performance is influenced by the effectiveness of the attribution system used to evaluate privacy preservation. If the AA system fails to perform well for certain genres or domains, the privacy metrics may become unreliable, undermining the overall obfuscation effectiveness. Secondly, the specific metrics used, such as CoLA, may carry inherent biases. For instance, CoLA often performs better with standard English, as the typical definition of fluency tends to favor text written in this form and for that reason, it may not be appropriate in some settings (e.g., the generated text will not have the same appeal if it sounds too "formal"). Additionally, the creation of obfuscation candidates relies on pre-trained language models, which are known to occasionally generate factually incorrect or hallucinatory information (Ji et al., 2023). While we use content-preserving metrics, these do not guarantee the factual integrity of obfuscated texts compared to original text. Both hallucinations (overgeneration) and omissions negatively impact these metrics, reflecting the discrepancies between the original and obfuscated texts. Ideally, we should employ methods from Information Extraction to ensure that the facts mentioned in the two documents are identical—neither more nor less. This approach would help maintain factual integrity, which is crucial, especially in sensitive domains. This underscores the need for further research in this area.

Acknowledgments

We would like to thank Skyler Hallinan, Jillian Fisher, and Yejin Choi from the University of Washington for fruitful discussions on the topic of authorship obfuscation in the IARPA HIATUS project. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2022-22072200003. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

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

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.

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.

Jillian Fisher, Ximing Lu, Jaehun Jung, Liwei Jiang, Zaid Harchaoui, and Yejin Choi. 2024. Jamdec: Unsupervised authorship obfuscation using constrained decoding over small language models. *arXiv preprint arXiv:2402.08761*.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2022. Gptq: Accurate post-training quantization for generative pre-trained transformers. *arXiv preprint arXiv:2210.17323*.

Skyler Hallinan, Faeze Brahman, Ximing Lu, Jaehun Jung, Sean Welleck, and Yejin Choi. 2023. STEER: Unified style transfer with expert reinforcement. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7546–7562, Singapore. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

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.

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: (best of the labs track at clef-2017). In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8*, pages 173–185. Springer.

Yashwant Keswani, Harsh Trivedi, Parth Mehta, and Prasenjit Majumder. 2016. Author masking through translation. *CLEF (Working Notes)*, 1609:890–894.

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*.

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Alex Warstadt, Amanpreet Singh, and Samuel R Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

A Detailed Description of HIATUS Research Datasets (HRS)

The HRS-HIATUS Research datasets from the IARPA HIATUS program[3] aim to bridge the gap between authorship attribution and privacy preservation. These datasets contain articles of different genres including tabletop game reviews from BoardGameGeek, instructions for making projects from Instructables, news articles from GlobalVoices, and user answers from StackExchange on liberal arts and STEM topics. Articles average 862 English words and have undergone Personally Identifiable Information (PII) removal using Microsoft's Presidio tool.

During testing, the corpus is split into a query set and a candidate set. The query set comprises approximately 0.5% of total authors and about 0.7% of total articles. The candidate set can come from the same or different genres. Performers are tasked with obfuscating the text from the query set such that it significantly differs from texts written by the same author in the candidate set, thereby testing the efficacy of obfuscation methods in disguising authorial style.

The datasets consist of 127,273 documents authored by 179 different authors. Below is a detailed table that outlines the structure of these datasets:

Source	Docs		Authors		Avg
	Query	Cand.	Query	Cand.	Words
BoardGameGeek	102	25,769	36	16,946	862
Instructables	46	25,722	19	16,997	865
GlobalVoices	65	25,617	26	16,962	862
StackExchange LA	87	25,526	30	16,950	863
StackExchange STEM	97	25,786	32	16,981	862
Mixed from HRS1.1-5	270	34,453	92	17,196	864

Table 3: Dataset statistics of HIATUS Research datasets (HRS)

[3] https://www.iarpa.gov/research-programs/hiatus