Quantifying Misattribution Unfairness in Authorship Attribution

Pegah Alipoormolabashi
Stony Brook University
palipoormola@cs.stonybrook.edu

Ajay Patel
University of Pennsylvania
ajayp@seas.upenn.edu

Niranjan Balasubramanian
Stony Brook University
niranjan@cs.stonybrook.edu

Abstract

Authorship misattribution can have profound consequences in real life. In forensic settings simply being considered as one of the potential authors of an evidential piece of text or communication can result in undesirable scrutiny. This raises a fairness question: Is every author in the candidate pool at equal risk of misattribution? Standard evaluation measures for authorship attribution systems do not explicitly account for this notion of fairness. We introduce a simple measure, Misattribution Unfairness Index (MAUI_k), which is based on how often authors are ranked in the top k for texts they did *not* write. Using this measure we quantify the unfairness of five models on two different datasets. All models exhibit high levels of unfairness with increased risks for some authors. Furthermore, we find that this unfairness relates to how the models embed the authors as vectors in the latent search space. In particular, we observe that the risk of misattribution is higher for authors closer to the centroid (or center) of the embedded authors in the haystack. These results indicate the potential for harm and the need for communicating with and calibrating end users on misattribution risk when building and providing such models for downstream use.

1 Introduction

Authorship attribution has various sensitive uses in multiple domains such as literature (Zhao and Zobel, 2007; Stańczyk and Cyran, 2007) and forensics (M. et al., 2016). In some uses, authorship misattribution or even suspected authorship can have dire consequences for the individuals involved. For example, consider a scenario where a forensic investigation hinges on finding who wrote a piece of text from a large pool of potential authors. Automatic authorship attribution is often seen as a way to reduce the number of candidates for such settings (Tschuggnall et al., 2019). However, even being suspected of authorship in this setting can result in further scrutiny. Given such undesirable consequences, we ask a question of fairness: *Are some authors more likely to be misattributed than others?*

To answer this question, we study the commonly used *needle-in-the-haystack* formulation (Rivera-Soto et al., 2021; Wang et al., 2023; Wen et al., 2024). In this setting, the *haystack* is a collection of known authors and documents they have authored D_h. Given some documents D_q from a query author i.e., ones whose authorship is unknown as of yet, the attribution task is to find a candidate author (*needle*) from the haystack who most likely authored the query documents. The task is then framed as a ranking problem, where systems rank haystack authors by comparing an embedding of their documents $e_i = \mathrm{enc}(D_h(a_i))$ to an embedding of the query documents $e_q = \mathrm{enc}(D_q)$. Such *embed-and-rank* solutions are appealing since they allow efficient scaling to large haystacks with many authors (Douze et al., 2024).

Existing evaluation measures for these solutions do not tackle fairness notions. Prior work mostly use effectiveness metrics, such as Mean Reciprocal Rank (MRR), or Recall at various ranks (R@k), or Recall at various ranks (R@k) (Andrews and Bishop, 2019; Khan et al., 2021; Rivera-Soto et al., 2021; Burrows et al., 2009). These metrics are designed to reduce the amount of manual scrutiny by assessing whether the correct author is ranked as highly as possible. They do not capture or measure the risks of getting ranked highly for other authors.

We make two contributions to remedy this gap: **1)** We introduce a way to measure the unfairness in rankings induced by models and empirically assess unfairness across five models over two datasets. **2)** We perform an analysis and provide a potential explanation for unfairness in how embedding distributions relate to misattribution risks of authors. Our results show that model rankings exhibit high levels of unfairness and authors closer to the center of the embedding space are at a higher risk for

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 1030–1041
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

misattribution. These call for further research both in evaluation and in modeling of authorship attribution systems to reduce the potential for unfairness-related harms.

2 Misattribution Unfairness in Author Ranking

We introduce a notion of fairness where all users carry equal risks of being misattributed i.e., are equally likely to be ranked high for documents that they have not authored. This is similar to the common definition of fairness in retrieval settings (Biega et al., 2020). In retrieval, the focus is that all authors receive relevance-proportional attention in the rankings. Here we focus on reducing undue or disproportional presence in ranking.

Suppose there are N_h authors in the haystack and N_q query authors selected at random from the haystack. For any single query, if we consider an unbiased ranking i.e. a random permutation of the authors, the probability of any specific author being ranked higher than k is $\frac{k}{N_h}$. When we query for N_q times, an author is expected to get ranked higher than k for $E_k = \lceil \frac{k}{N_h} \times N_q \rceil$ times.

The unfairness of model-induced rankings can thus be characterized in terms of how the actual counts of authors being ranked at top k exceeds the expected count[1]. Let c_j^k denote the number of times author a_j is ranked in the top k. Then, we can quantify the unfairness as follows:

$$\mathrm{MAUI_k} = \frac{\sum_{j=1}^{N_h} \max(0, (c_j^k - E_k))}{k \times (N_q - E_k)} \quad (1)$$

This metric normalizes the sum of the differences by its highest possible value which happens in the worst case: when the same k authors are ranked in top k for all queries. This scales the values between 0 and 1, 0 being most fair and 1 being the least.

3 Evaluation

We experiment with multiple authorship models and datasets and use cosine similarity for ranking.

3.1 Experimental Setup

Datasets (i) Reddit: We use the evaluation partition of the dataset by Andrews and Bishop (2019) with 111,396 candidate authors and 25,000 query authors. (ii) Bloggers: We select 9000 bloggers from the authorship corpus of Blogger posts (Schler et al., 2006) and use 2500 of them as queries. (iii) Fanfiction: We randomly select 20,000 fanfiction authors, and use 7500 of them as queries.

Models We use five text embedding models: 1. SBERT (Reimers and Gurevych, 2019): A sentence transformer model based on DistilRoBERTa (Sanh et al., 2019) 2. LUAR (Rivera-Soto et al., 2021): A universal authorship embedding model trained on the Reddit Million User Dataset (Khan et al., 2021). 3. Style Embedding (Wegmann et al., 2022): A sentence transformer built on RoBERTa-base (Liu et al., 2019b) and trained for style representation. We call this model Wegmann. 4. StyleDist. (Patel et al., 2024): Another style embedding model trained on a combination of Reddit comments and synthetic data. 5. MPNet$_{\mathrm{AR}}$: Microsoft's sentence transformer (MPNet$_{\mathrm{st}}$) by Song et al. (2020) that we train for authorship representation. Section A.1 of the appendix provides more details on models and datasets.

We show the performance of the five embedding models on the two datasets in Table 1. In the rest of this section we look at the models' embeddings and rankings from other angles.

3.2 Misattribution Unfairness in Model Rankings

We measure MAUI$_{\mathrm{k}}$ (eq. 1) for different values of k and show the results in Table 2. Higher values show higher unfairness. Putting Table 2 and 1 together, we can see that there is no clear relationship between how good a model is in correctly attributing authors and how fair it is in misattributing authors. For example, Wegmann is the worst performing model in terms of R@8 and MRR, but it consistently shows the least unfairness. MPNet$_{\mathrm{AR}}$ and StyleDist. have very close MAUI scores, yet their ranking performance varies greatly. Even LUAR with very high ranking scores does not guarantee unfairness in misattribution. In fact, after SBERT, LUAR is the most unfair model to the Bloggers.

To present the unfairness from another viewpoint, we count the number of authors who carry higher risks of being ranked in top $k = 10$. Tables 3a and 3b show how these counts compare to

[1] We acknowledge that expecting a random permutation does not factor in the demographic or stylistic "relevance" of non-query authors to the query author. For example, an author who shares a regional dialect with the query author is more relevant to the query in this regard. While this notion of relevance is useful for attribution, it can be unfair in forensic settings as if a member of Demographic X is a suspect of a crime, undue scrutiny should not be brought to all members of Demographic X. While imperfect, a random permutation is a reasonable calibration to benchmark systems against and measure the magnitude of bias towards certain authors.

	Blogs		Reddit		Fanfiction	
Model	R@8	MRR	R@8	MRR	R@8	MRR
SBERT	0.61	0.48	0.15	0.10	0.28	0.22
LUAR	0.97	0.90	0.82	0.71	0.53	0.44
MPNet$_{\text{AR}}$	0.96	0.88	0.40	0.30	0.30	0.25
Wegmann	0.45	0.32	0.08	0.05	0.09	0.06
StyleDist.	0.68	0.55	0.09	0.06	0.16	0.12

Table 1: Recall-at-8 and Mean Reciprocal Rank scores of embedding models on three datasets.

the expected count E_{10}. The number of unfairly misattributed authors and the severity of this issue vary across models. Similar to the trends in Table 2 Wegmann is the most fair of the models on both Reddit and Blogs datasets. Again, for Blogs, LUAR is highly unfair despite its superior performance. Looking at the last column of Table 3b, LUAR's number of misattributed authors is more than MPNet$_{\text{AR}}$, StyleDist., and Wegmann together. Table 3 shows the counts of the number of authors subject to unfair misattribution. The differences between columns show that the extent of this misattribution risk varies for different authors. In section D of the appendix we show how the risk is distributed among authors. Looking at the authors with the most misattribution we see that for instance, with SBERT there is -at least- one author for whom the misattribution risk is almost 40x the random case.

3.3 Relation to Embedding Distribution

How the author embeddings are distributed is central to the effectiveness of attribution and fairness of misattribution in rankings. One way to analyze this relationship is via the distance of author embeddings to the centroid (center) of the embeddings.

To this end, we average all author embeddings to compute the centroid and measure distance of each author to the centroid as $1 - cosine$. We plot authors' mean rank (average of their ranking across all queries) against their distance to centroid in figure 1. For better visualization the distance values are scaled to the [0,1] range. We see strong correlations between authors' distance to the centroid and their risk of misattribution, as measured by their average rank over all queries. Across models and datasets the closer an author is to the centroid, the lower their average rank is; i.e. authors closer to the centroid are ranked higher on average. Each author is ranked for queries that are chosen at random, therefore an author's average ranking should not correlate with their distance from the centroid. We expect the plots to be close to a horizontal line. The expected average rank for every author is the middle of the ranked list i.e., half the size of haystack. Among the models we compare, Wegmann curves are closest to the ideal. Note that these correlations depend largely on how the embeddings are distributed. Indeed, we find that distance distributions vary considerably across different models. See Figure 2 in Appendix Section B.

3.4 Unfairness of missed attribution

The centroid analyses can also help us understand which authors are harder to find i.e., ones who are not recognized as the author of their own documents. To assess how often an author is ranked high for their own documents, we average their reciprocal ranks over multiple query subsets drawn from their documents[2]. Authors that are ranked close to the top when queried have higher MRRs than those who are not.

We then test three hypotheses relating MRR to authors' distance from centroid: (i) Authors with higher MRR are further away from the centroid compared to authors with lower MRR, (ii) Authors with higher MRR are further away from the centroid compared to a random subset of authors. (iii) Authors with lower MRR are closer to the center than a random subset of authors. We use the Mann-Whitney U test [3] (Mann and Whitney, 1947) to accept or reject our hypotheses. Table 7 in Appendix shows statistics for all tests. The numbers show that for Reddit the first and the second hypotheses are statistically supported for all models and the third only for LUAR. For blogs, only the third hypothesis for SBERT is rejected. See figures 3 and 4 for visualizations of this phenomenon. To summarize, while authors closer to centroid are more likely to be ranked higher when queried for

[2] Specifically, we perform four queries per each query author, each comprising four of their documents.

[3] This test does not assume normality.

	k			
Model	5	10	15	20
SBERT	0.20	0.31	0.36	0.39
LUAR	0.06	0.12	0.16	0.18
MPNet$_{\text{AR}}$	0.09	0.17	0.22	0.25
Wegmann	0.03	0.09	0.13	0.15
StyleDist.	0.07	0.15	0.19	0.22

(a) Reddit

	k			
Model	5	10	15	20
SBERT	0.24	0.36	0.37	0.40
LUAR	0.15	0.26	0.27	0.31
MPNet$_{\text{AR}}$	0.12	0.23	0.23	0.27
Wegmann	0.06	0.14	0.13	0.17
StyleDist.	0.11	0.22	0.21	0.25

(b) Blogs

	k			
Model	5	10	15	20
SBERT	0.21	0.30	0.34	0.36
LUAR	0.17	0.25	0.28	0.30
MPNet$_{\text{AR}}$	0.17	0.25	0.28	0.30
Wegmann	0.08	0.12	0.14	0.15
StyleDist.	0.12	0.18	0.21	0.22

(c) Fanfiction

Table 2: (MAUI$_{\text{k}}$) for different k values across models and datasets. MAUI$_{\text{k}}$ is defined in section 2 as a measure of an authorship attribution system's unfairness in misattributing authors. The scores range between 0 (most fair) and 1 (most unfair).

other authors' documents, they are not necessarily ranked high for their own. These results also demonstrate the importance of embedding distributions to the fairness of authorship rankings.

4 Related Work

Automatic authorship analysis has a rich body of work (see Huang et al., 2025; Tyo et al., 2022), with a heavy emphasis on usefulness and reliability of features used in authorship analysis (e.g. Chaski (2001); Baayen et al. (2002), or the effects of the attribution setup (e.g. Stamatatos (2013); Sari et al. (2018)). While these provide general insights into authorship attribution, to the best of our knowledge, there have been no specific studies that focus on notions of fairness in authorship attribution. Fairness of NLP models has been studied in many application domains including dialogue (Liu et al., 2019a), language modeling (Cao et al., 2022; Qian et al., 2022), text generation (Fleisig et al., 2023), classification (Pruksachatkun et al., 2021) and clinical NLP (Meng et al., 2022). Gallegos et al. (2024) survey research on bias and fairness in large language models focusing on metrics, probing datasets, and bias mitigation techniques. There have also been extensive research on fairness and bias related metrics. Czarnowska et al. (2021) survey, categorize, and compare some of the fairness and bias related metrics.

	$> 2 \times E_{10}$	$> 4 \times E_{10}$	$> 5 \times E_{10}$
SBERT	8487	2582	1599
LUAR	4290	242	54
MPNet$_{\text{AR}}$	6054	701	299
Wegmann	2967	18	3
StyleDist.	5411	382	106

(a) Reddit

	$> 2 \times E_{10}$	$> 4 \times E_{10}$	$> 5 \times E_{10}$
SBERT	858	296	214
LUAR	821	189	96
MPNet$_{\text{AR}}$	816	122	56
Wegmann	496	9	1
StyleDist.	789	104	33

(b) Blogs

	$> 2 \times E_{10}$	$> 4 \times E_{10}$	$> 5 \times E_{10}$
SBERT	1319	328	104
LUAR	1161	170	38
MPNet$_{\text{AR}}$	1155	181	43
Wegmann	316	0	0
StyleDist.	780	32	0

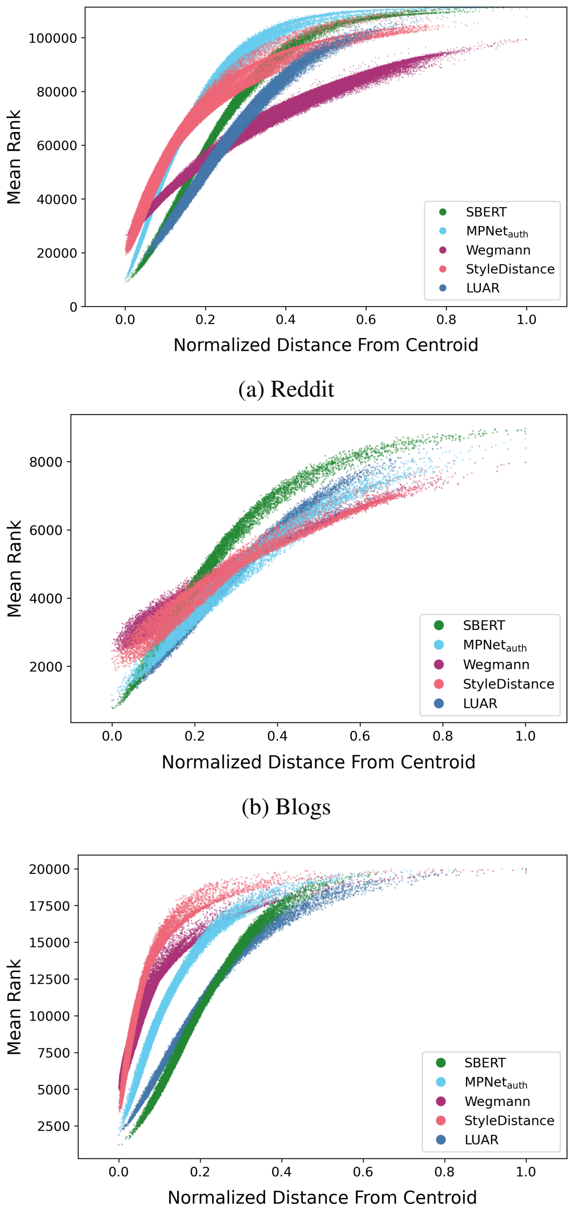
(c) Fanfiction

Table 3: Number of authors who are ranked in top 10 more than the expected number i.e. E_{10}. Note that the size of the haystack and the number of the queries are different across datasets, hence the huge difference in the numbers between the three tables.

Model	**Reddit**	**Blogs**	**Fanfic**
LUAR	9.75	10.0	12.0
SBERT	39.00	21.75	5.8
Wegmann	4.50	4.25	3.6
MPNet	12.25	10.25	12.0
StyleDist	8.25	7.0	19.6

Table 4: Extreme cases of unfair misattribution risk. Numbers show the ratio of the number of times an author is ranked in top 10 to E_{10} for the author bearing the highest risk of misattribution.

The closest to our work is research on fairness of embedding-based document retrieval systems. Their main concerns are: group fairness (Yang



(a) Reddit

(b) Blogs

(c) Fanfiction

Figure 1: Relationship between authors' average rank and their distance from the centroid. Distances are min-max normalized. Authors' mean rank is highly correlated with their distance from centroid.

and Stoyanovich, 2016), individual fairness (Biega et al., 2018; García-Soriano and Bonchi, 2021), their trade-off with each other and with relevance (Gao and Shah, 2019), and recently unfairness in ranking with LLMs (Wang et al., 2024). These works aim to ensure relevant documents from certain individuals or groups are fairly represented in the top ranks. In contrast, in our authorship setting, the focus is on the risks of being included in the top ranks in forensic or law enforcement settings.

5 Conclusion

Authorship attribution carries profound risks in settings such as forensic and law enforcement uses. In these settings, measures of effectiveness alone are not adequate for evaluating and understanding the impact of misattribution. To this end, we argued that we also need to understand if the rankings induced by attribution models distribute misattribution risks equally. The unfairness measure we introduced helps quantify these risks. Our empirical measurements of models used to produce authorship embeddings shows that in an embed-and-rank approach there are many authors who are at a substantially higher risk of being ranked in the top k. We further showed that this risk correlates with how the embeddings are distributed. Our findings call for careful consideration of such notions of misattribution fairness in evaluation, development, and deployments of authorship analysis systems.

Limitations

In practice, the anticipated "fair" ranking of the haystack authors is not entirely random, but rather a function of authors' relatedness to the query. The distribution of queries therefore impacts the "most fair" baseline, and consequently the unfairness measurements. This impact is not accounted for in this paper. Our results hold for the set of queries selected for in our experiments, which were chosen at random. It is possible different selection methods of queries may impact unfairness measurements. Additionally, this work's focus is limited to cases where over-attribution is undesired. A broader study would also cover cases where under-attribution is problematic.

Ethical Considerations

Potential Misuse Results and analyses presented in this paper are meant to prompt researchers in authorship attribution to focus on unfairness, assessing it, and preventing its potential harms. People and institutions using authorship attribution may misuse these results to justify automatically pre-empting groups of candidate authors. We do not investigate characteristics of authors who are easy or difficult to track down. Nevertheless, a malicious agent may reproduce our experiments, analyze hard-to-find authors (near-centroid authors), and use characteristics of their writing as an authorship obfuscation method.

Our Work We report licenses and sources of every artifact we used in Appendix A. We did not investigate individuals or use demographic information (e.g. reported age and gender published in the Blogs corpus) about authors in any of our experiments and operate under the assumption that no author should have significantly greater than random expectation of appearing in the top k results when they are not the query author of interest regardless of demographics.

Acknowledgments

This research is supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200005. 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

Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. *Preprint*, arXiv:1910.04979.

Harald Baayen, Hans Van Halteren, Anneke Neijt, and Fiona Tweedie. 2002. An experiment in authorship attribution. In *6th JADT*, volume 1, pages 69–75. Citeseer.

Asia J. Biega, Fernando Diaz, Michael D. Ekstrand, and Sebastian Kohlmeier. 2020. Overview of the trec 2019 fair ranking track. *Preprint*, arXiv:2003.11650.

Asia J. Biega, Krishna P. Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.

Steven Burrows, Alexandra L. Uitdenbogerd, and Andrew Turpin. 2009. Application of information retrieval techniques for source code authorship attribution. In *Database Systems for Advanced Applications*, pages 699–713, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, J. Dhamala, and A. G. Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. *ArXiv*, abs/2203.13928.

Carole E Chaski. 2001. Empirical evaluations of language-based author identification techniques. *Forensic linguistics*, 8:1–65.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. The faiss library.

Eve Fleisig, Aubrie Amstutz, Chad Atalla, Su Lin Blodgett, Hal Daumé, Alexandra Olteanu, Emily Sheng, Dan Vann, and Hanna M. Wallach. 2023. Fairprism: Evaluating fairness-related harms in text generation. In *Annual Meeting of the Association for Computational Linguistics*.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Ruoyuan Gao and Chirag Shah. 2019. How fair can we go: Detecting the boundaries of fairness optimization in information retrieval. *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*.

David García-Soriano and Francesco Bonchi. 2021. Maxmin-fair ranking: Individual fairness under group-fairness constraints. *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*.

Baixiang Huang, Canyu Chen, and Kai Shu. 2025. Authorship attribution in the era of llms: Problems, methodologies, and challenges. *Preprint*, arXiv:2408.08946.

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the cross-domain authorship verification task at pan 2020. In *Conference and Labs of the Evaluation Forum*.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Walter Daelemans, Martin Potthast, and Benno Stein. 2019. Overview of the cross-domain authorship attribution task at pan 2019. In *Conference and Labs of the Evaluation Forum*.

Aleem Khan, Elizabeth Fleming, Noah Schofield, Marcus Bishop, and Nicholas Andrews. 2021. A deep metric learning approach to account linking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5275–5287, Online. Association for Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019a. Does gender matter? towards fairness in dialogue systems. *ArXiv*, abs/1910.10486.

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

Coulthard M., Johnson A., and Wright D. 2016. *An Introduction to Forensic Linguistics: Language in Evidence (2nd ed.)*. Routledge.

Henry B. Mann and Douglas R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18:50–60.

Chuizheng Meng, Loc Trinh, Nan Xu, James Enouen, and Yan Liu. 2022. Interpretability and fairness evaluation of deep learning models on mimic-iv dataset. *Scientific Reports*, 12.

Ajay Patel, Jiacheng Zhu, Justin Qiu, Zachary Horvitz, Marianna Apidianaki, Kathleen McKeown, and Chris Callison-Burch. 2024. Styledistance: Stronger content-independent style embeddings with synthetic parallel examples. *Preprint*, arXiv:2410.12757.

Yada Pruksachatkun, Satyapriya Krishna, J. Dhamala, Rahul Gupta, and Kai Wei Chang. 2021. Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification. *ArXiv*, abs/2106.10826.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer nlp. In *Conference on Empirical Methods in Natural Language Processing*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *Preprint*, arXiv:1908.10084.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Yunita Sari, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? exploring the most useful features for authorship attribution. In *International Conference on Computational Linguistics*.

Jonathan Schler, Moshe Koppel, Shlomo Engelson Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Preprint*, arXiv:2004.09297.

Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n -gram features. *Journal of law and policy*, 21:7.

Urszula Stańczyk and Krzysztof A Cyran. 2007. Machine learning approach to authorship attribution of literary texts. *International journal of applied mathematics and informatics*, 1(4):151–158.

Michael Tschuggnall, Benjamin Murauer, and Günther Specht. 2019. Reduce & attribute: Two-step authorship attribution for large-scale problems. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 951–960, Hong Kong, China. Association for Computational Linguistics.

Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton. 2022. On the state of the art in authorship attribution and authorship verification. *Preprint*, arXiv:2209.06869.

Andrew Wang, Cristina Aggazzotti, Rebecca Kotula, Rafael A. Rivera Soto, Marcus Bishop, and Nicholas Andrews. 2023. Can authorship representation learning capture stylistic features? *Transactions of the Association for Computational Linguistics*, 11:1416–1431.

Yuan Wang, Xuyang Wu, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2024. Do large language models rank fairly? an empirical study on the fairness of LLMs as rankers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5712–5724, Mexico City, Mexico. Association for Computational Linguistics.

Anna Wegmann, Marijn Schraagen, and Dong Nguyen. 2022. Same author or just same topic? towards content-independent style representations. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 249–268, Dublin, Ireland. Association for Computational Linguistics.

Zichen Wen, Dadi Guo, and Huishuai Zhang. 2024. Aidbench: A benchmark for evaluating the authorship identification capability of large language models. *ArXiv*, abs/2411.13226.

Ke Yang and Julia Stoyanovich. 2016. Measuring fairness in ranked outputs. *Preprint*, arXiv:1610.08559.

Ying Zhao and Justin Zobel. 2007. Searching with style: Authorship attribution in classic literature. In *ACM international conference proceeding series*, volume 244, pages 59–68. Citeseer.

A Technical Details of Experiments

A.1 Models and Datasets

Blogs Data

We use the corpus of Blogger posts collected by Schler et al. (2006) as it is commonly used in authorship analysis research. This dataset "may be freely used for non-commercial research purposes."[4] We filter for authors with more than 10 and fewer than 200 posts. From those we pick 9000 candidate authors, 2500 of which are used as queries and needles. Each author in the haystack has 16 blog posts (randomly selected) and each author in the queries has 10.

Reddit Data

For evaluation we use the evaluation partition of the Reddit dataset created by Andrews and Bishop (2019). It has 111,396 candidate authors and 25,000 query authors. Each author in the haystack has 16 Reddit comments, and each author in the queries has 16 Reddit comments as well. For finetuning MPNet$_{\text{AR}}$ we use a part of the Million User Dataset by Khan et al. (2021).

Fanfiction Data

We use a subset of the fanfiction dataset from PAN19 (Kestemont et al., 2019) and PAN20 (Kestemont et al., 2020) for evaluation. This subset consists of 20000 haystack authors, 7500 of whom are used as queries.

A.2 Finetuning MPNet$_{\text{st}}$

We choose to finetune MPNet$_{\text{st}}$ (Song et al., 2020) because it has a different structure than the BERT-based models. Also, all-mpnet-base-v2 [5] is a top-performing sentence transformer. Out of 12 layers we keep 8 frozen for training. This model is licensed under the apache-2.0 license. We use cached multiple-negative ranking loss implemented in the sentence-transformers library[6]. Training with multiple negatives has shown to be very effective in the case of LUAR (Rivera-Soto et al., 2021). We set the learning rate to $5e^{-5}$ with a linear scheduler. Maximum sequence length is set to 512, and we train with a batch size of 200. We train for 5000 steps on the training subset of Reddit data by Khan et al. (2021). Training computation was done on an NVIDIA RTX A6000 GPU.

[4] See https://u.cs.biu.ac.il/ koppel/BlogCorpus.htm

[5] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

[6] https://www.sbert.net/index.html

Model	Base Sentence-Transformer	Training Objective	Training Data	# of Parameters	License
SBERT	DistilRoBERTa	-	-	82M	apache-2.0
LUAR	paraphrase-distilroberta-base-v1	Batch Contrastive Loss*	Reddit (MUD)	82M	apache-2.0
MPNet$_{\text{AR}}$	all-mpnet-base-v2	Multiple-Negative Ranking Loss	Reddit(MUD)	109M	apache-2.0
Wegmann	RoBERTa-base	Triplet Loss	Reddit	124M	mit
StyleDist.	RoBERTa-base	Contrastive Loss	Reddit + Synthetic Data	124M	mit

Table 5: Overview of embedding models we use in experiments. We perform inference on these models using an NVIDIA TITAN Xp GPU.

A.3 Indexing and Retrieving Embeddings

We use Faiss (Douze et al., 2024)[7] for efficient indexing of author embeddings and performing fast searches. All the indexing and search computations are done on CPU. Faiss library is under an MIT-license.

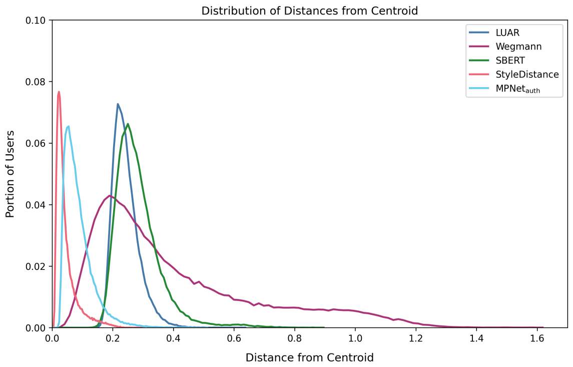
A.4 Libraries and Packages

Here is a list of major python libraries we use (with no particular order):

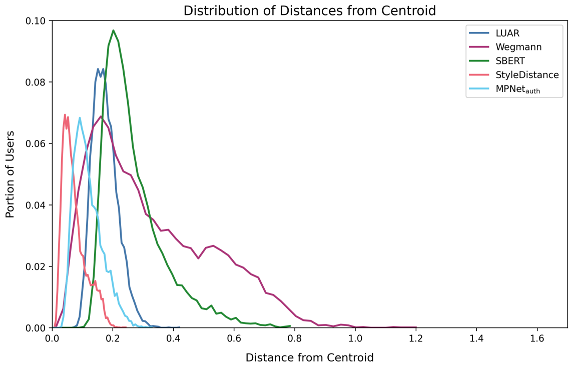
- torch==2.3.1
- sentence-transformers==3.3.1
- transformers==4.45.2
- tensorflow-datasets==4.9.6
- scikit-learn==1.5.0
- scipy==1.14.0
- numpy==1.26.4
- nltk==3.8.1
- matplotlib==3.9.2
- seaborn==0.13.2
- faiss-cpu==1.8.0.post1

B Embedding Distributions

Since authorship attribution in this setting is based entirely upon embeddings and their distances, we use a simple distance-to-centroid measure to characterize the distribution of author embeddings under different models. Intuitively, if all author embeddings are clustered close to each other then the distances to the centroid will be small. If they are spread out then the distances to the centroid will be large. Per each dataset and model we find the centroid of all haystack author embeddings, then calculate each author's distance from this centroid. The distance measure is $1 - cosine$, so it can take values from 0 to 2.



(a) Reddit



(b) Blogs

Figure 2: Distribution of authors' distinctness (i.e. their distance from the centroid computed as: $1 - cosine$)

Figure 2 shows the distribution of the distance value per each model for Reddit and Blogs datasets. The diversity of the distributions in terms of domain and curve shape inspired us to investigate how it affects the effectiveness and fairness of authorship attribution.

[7] https://github.com/facebookresearch/faiss

Model	Max	Mean	Std
LUAR	9.75	1.62	0.50
SBERT	39.00	2.37	1.67
Wegmann	4.50	1.50	0.32
MPNet	12.25	1.79	1.74
StyleDist	8.25	1.69	0.56

(a) Reddit

Model	Max	Mean	Std
LUAR	10.0	2.07	1.10
SBERT	21.75	2.61	2.12
Wegmann	4.25	1.58	0.40
MPNet	10.25	1.92	0.85
StyleDist	7.00	1.84	0.70

(b) Blogs

Model	Max	Mean	Std
LUAR	12.0	2.03	1.05
SBERT	19.6	2.32	1.58
Wegmann	3.6	1.54	0.39
MPNet	12.0	2.04	1.09
StyleDist	5.80	1.75	0.64

(c) Fanfiction

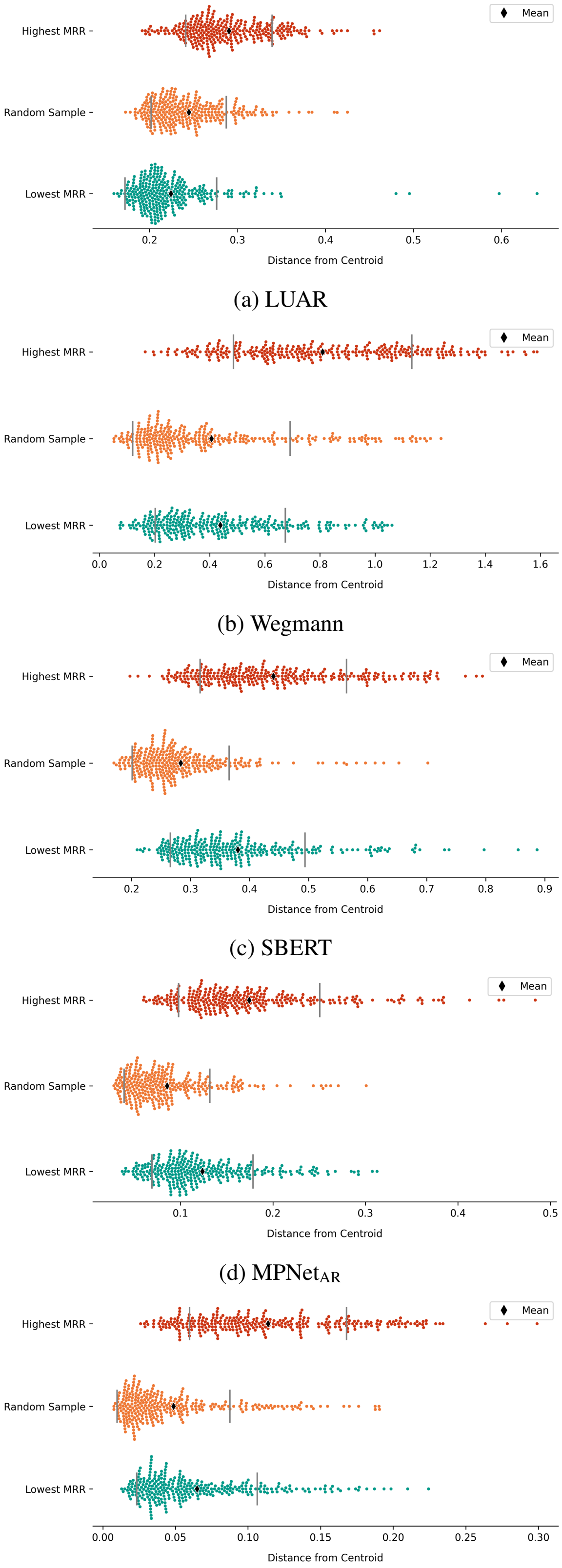
Table 6: Variation of the risk of unfair misattribution across different authors measured by the ratio of the number of times an author is ranked in top 10 to E_{10}.

C Analysis of Needle Authors' Distance to Centroid

Test statistics corresponding to section 3.4 are presented in 7. Additionally, we sample 300 Reddit authors with highest MRRs, 300 authors with lowest MRRs, and 300 random authors to visualize the relationship between needle authors' MRR and their distance from the centroid. Figure 3 shows the distribution of the sampled authors' distances from the centroid. Similar plots for Blogs experiments are in figure 4.

D Risk of Misattribution for Different Authors

Not all authors who are subject to over-misattribution carry the same risk. Per dataset and model, we obtain the ratio of the number of times an author is ranked in top 10 to E_{10}: $u_j^{10} = \frac{c_j^{10}}{E_{10}}$. For all authors contributing to the MAUI$_{10}$ score (eq. 1) this ratio is more than 1. Table 3 shows for how many authors this ratio is higher than 2, 4, and 5. Table 6 shows the maximum, mean, and standard deviation of u^{10} across authors.



(a) LUAR

(b) Wegmann

(c) SBERT

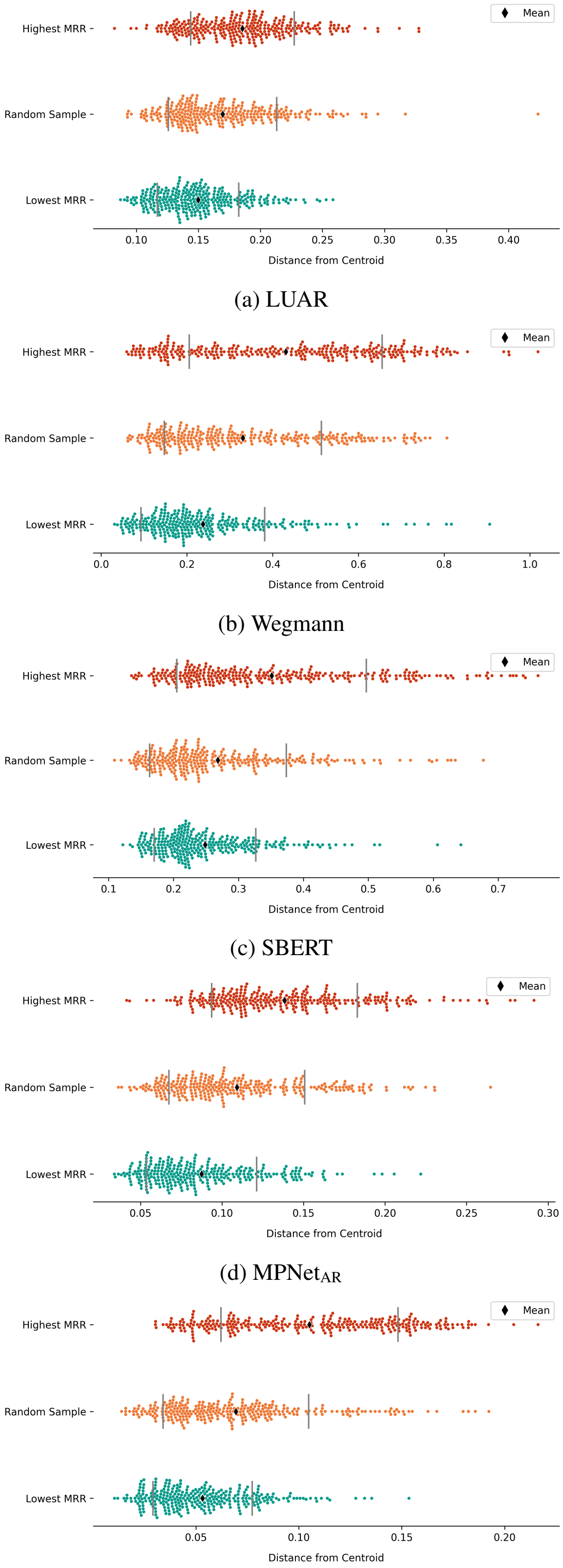
(d) MPNet$_{\text{AR}}$

(e) StyleDist.

Figure 3: Reddit - Distribution of Needle Authors' Distances from the Centroid. A comparison between random samples of authors, a group of easily found authors and a group of hard to find authors.

		Reddit			Blogs		
		Hyp. (i)	Hyp. (ii)	Hyp. (iii)	Hyp. (i)	Hyp. (ii)	Hyp. (iii)
LUAR	U	79199.0	71937.5	30635.0	67607.0	56918.5	34128.0
	p	$1.12e^{-58}$	$3.46e^{-37}$	$6.63e^{-12}$	$8.91e^{-27}$	$9.91e^{-9}$	$1.52e^{-7}$
Wegmann	U	74052.0	75528.0	50676.0	67255.0	55346.5	31362.0
	p-value	$6.36e^{-43}$	$3.51e^{-47}$	1.0	$5.22e^{-6}$	$5.49e^{-7}$	$6.66e^{-11}$
SBERT	U	59269.0	78719.5	70489.0	65060.0	60090.5	41619.0
	p-value	$9.05e^{-12}$	$4.22e^{-57}$	1.0	$1.72e^{-21}$	$5.9e^{-13}$	0.055
MPNet$_{\text{AR}}$	U	64956.0	79814.5	66651.5	74715.0	61172.5	29477.0
	p-value	$2.75e^{-21}$	$9.91e^{-61}$	1.0	$8.26e^{-45}$	$1.30e^{-14}$	$1.33e^{-13}$
StyleDist.	U	70389.0	79633.0	62192.5	76030.0	63759.5	30604.5
	p-value	$2.94e^{-33}$	$4.03e^{-60}$	1.0	$1.12e^{-48}$	$4.97e^{-19}$	$6e^{-12}$

Table 7: Mann-Whitney U-statistics and p-values of testing three hypotheses about our authorship attribution experiments: (i) Authors with higher MRR are further away from the centroid compared to authors with lower MRR, (ii) Authors with higher MRR are further away from the centroid compared to a random subset of authors. (iii) Authors with lower MRR are closer to the center than a random subset of authors.



(a) LUAR

(b) Wegmann

(c) SBERT

(d) MPNet_{AR}

(e) MPNet_{AR}

Figure 4: Blogs - Distribution of Needle Authors' Distances from the Centroid. A comparison between random samples of authors, a group of easily found authors and a group of hard to find authors.