



Poster: Style Pooling: Automatic Text Style Obfuscation for Hiding Sensitive Attributes

Fatemehsadat Mireshghallah, Taylor Berg-Kirkpatrick
{fatemeh, tberg}@ucsd.edu

Abstract—Text style can reveal sensitive attributes of the author (e.g. race or age) to the reader, which can, in turn, lead to privacy violations and bias in both human and algorithmic decisions based on text. For example, the style of writing in job applications might reveal protected attributes of the candidate which could lead to bias in hiring decisions, regardless of whether hiring decisions are made algorithmically or by humans. We propose a VAE-based framework that obfuscates stylistic features of human-generated text through style transfer *by automatically re-writing the text itself*. Our framework operationalizes the notion of obfuscated style in a flexible way that enables two distinct notions of obfuscated style: (1) a minimal notion that effectively *intersects* the various styles seen in training, and (2) a maximal notion that seeks to obfuscate by adding stylistic features of *all sensitive attributes* to text, in effect, computing a *union* of styles. Our style-obfuscation framework can be used for multiple purposes, however, we demonstrate its effectiveness in improving the fairness of downstream classifiers. We also conduct a comprehensive study on style pooling’s effect on fluency, semantic consistency, and attribute removal from text, in two and three domain style obfuscation. 

This paper appeared in the proceedings of EMNLP 2021 .


Index Terms—Fairness, Privacy, Natural Language Processing, Language Modeling

I. INTRODUCTION


Machine learning (ML) algorithms are used in a wide range of tasks, including high-stakes applications like determining credit ratings, setting insurance policy rates, making hiring decisions, and performing facial recognition. It has been shown that such algorithms can produce outcomes that are biased towards a certain gender or race. Ideally, high-stakes decisions made by either humans or ML algorithms, should not be influenced by irrelevant, protected attributes like nationality, age, or gender. In many instances, the input data used for making high-stakes decisions is text that is authored by a human candidate – for example, hiring decisions are often based on bios and personal statements. Recent work shows that automatic hiring-decision models trained on bios are less likely to select female candidates for certain roles (e.g. architect, software engineer, and surgeon) even when the gender of the author is not explicitly provided to the system. Bias is, of course, not limited to algorithmic decisions, humans make biased decisions based on text, even when the protected attributes of the author are not explicitly revealed. Together, these results indicate that both algorithms and humans can (1) decipher protected attributes of authors based on stylistic features of text, and (2) whether consciously or not, be biased by this information. A large body of prior work has

attempted to address *algorithmic bias* by modifying different stages of the natural language processing (NLP) pipeline. While effective in many cases, such approaches do nothing to mitigate bias in decisions made by humans based on text. We propose a fundamentally different approach. Rather than mitigating bias in learning algorithms that make decisions based on text, we propose a framework that obfuscates stylistic features of human-generated text *by automatically re-writing the text itself*. By obfuscating stylistic features, readers (human or algorithms) will be less able to infer protected attributes that enable bias.

II. APPROACH

We introduce a novel framework that enables ‘style pooling’: the automatic transduction of user-generated text to a central, obfuscated style. Notions of ‘centrality’ can themselves introduce bias – for example, a system might learn to obfuscate by mapping all text to the dominant style seen in its training corpus. This might ‘white-wash’ text, ignoring stylistic features of underrepresented groups in the learned notion of central style. Our framework operationalizes the notion of centrality in a more flexible way: our probabilistic approach allows us to choose between two distinct notions of centrality. First, we define a variant of our model which is incentivized to learn a minimal notion of central style that effectively *intersects* the various styles seen in training. This is achieved through the design of this variant’s probabilistic prior. We further equip this variant with a novel “de-boosting” mechanism, which amplifies the use of words that are less likely to leak sensitive attributes, and de-incentivizes the use of words whose presence might hint at a particular sensitive attribute. Second, we propose an alternative prior that instead incentivizes a maximal notion of style that seeks to obfuscate by adding stylistic features of all protected attributes to text – in effect, computing a *union* of styles. Table  shows our intersection and union obfuscation applied to sentences from the Blogs dataset, and highlights the differences between them.

While we propose both these obfuscations in our framework and leave it to the users to choose, it is worth noting that the cognitive process literature shows that when humans are confronted with conflicting biasing information, they tend to form an opinion about the conflicting text, based on their own implicit biases. Therefore, removing sensitive stylistic features may be more effective than combining them. This is also commensurate with our findings, where we observed that intersection more successfully improves the fairness metric.

Figure  shows an overview of our framework, where we depict a grouping of authors by age into three domains.

¹Code, models, and data is available at <https://github.com/mireshghallah/style-pooling>

TABLE I: Example Blog sentences transformed with A4NT [2] and our proposed Intersection and Union obfuscations. Our Intersection obfuscation aims at changing the style such that it does not reflect either teen or adult style. However, the union, tries to reflect both by making changes like adding “...” to the beginning of the sentence (adult style) while keeping the “grr” (teen style). Or by adding exclamation marks at the end of the sentence.

Age	Input Sentence (Original Data)	A4NT (Baseline)	Intersection	Union
Teen	grr ... now i get cold quicker .	grr now i get cold lol .	hmmm ... now i get cold grr ... now i get cold quicker .
Teen	it was so fricken hilarious .	it was so boring hilarious .	it was so utterly hilarious .	it was so totally hilarious
Adult	well i 've just been too busy .	well i 've just been kinda fun .	well i 've just been too busy .	well i 've just been too busy .
Adult	these were common phrases .	these were common teacher .	these were common .	these were common ! !

Our generative process assumes each sentence x_i , with corresponding domain $d(i)$, is generated as follows: First, a latent sentence y_i is sampled from a central prior, $p_{prior}(y_i)$, which is domain agnostic. Then, x_i is sampled conditioned on y_i from a transduction model, $p(x_i|y_i; \theta_{y \rightarrow x}^{d(i)})$. We let $\theta_{y \rightarrow x}^{D_j}$ represent the parameters of the transduction model for the j th domain. We extensively discuss p_{prior} in the full paper. For now, we assume the prior distributions are pretrained on the observed data and therefore omit their parameters for simplicity of notation. The log marginal likelihood of the observed data, which we approximate during training, can be written as:

$$\begin{aligned} \log p(X^{D_1}, \dots, X^{D_M}; \theta_{y \rightarrow x}^{D_1}, \dots, \theta_{y \rightarrow x}^{D_M}) \\ = \log \sum_y p(X^{D_1}, \dots, X^{D_M}; \theta_{y \rightarrow x}^{D_1}, \dots, \theta_{y \rightarrow x}^{D_M}) \end{aligned} \quad (1)$$

III. EVALUATION

We extensively evaluate our proposed framework on a wide range of tasks. First, we compare and contrast our “intersection” and “union” obfuscations on a modified version of the Yelp dataset and show that our intersection obfuscation successfully removes these misspellings and replaces them by the dominant spelling of the word 99.20% of the time. Then, we evaluate our framework on the Blogs data, where the sensitive attribute is age, and we measure the impact our obfuscations have on the fairness of a job classifier, using the the TPR-gap measure. We also evaluate the removal of sensitive attributes, fluency of the generated text, and the uncertainty of a sensitive attribute classifier for our framework, in both two and three domain setups.

For the sake of space, we only show one of our results here. The top section of Table II shows the linguistic and sensitive-attribute classification metrics for two domain obfuscations. We can see that de-boosting (denoted as DB) offers a trade-off between the linguistic quality of the generated text and the obfuscation of sensitive attributes. The *Intersection* obfuscation with de-boosting multiplier of 25 outperforms A4NT, with lower classifier accuracy, higher entropy and much lower Confident Response (CR) rate from the classifier. In general, the *Intersection* obfuscation, even without de-boosting does well on *Entropy* and *CR*, which shows that our method is doing well at creating doubt in terms of what the age of the author is. Our *Union* obfuscation is behaving differently from the *Intersection*, and is inferior in terms of obfuscating the text, with higher classifier accuracy and lower entropy. However, it has higher lexical diversity, which could hint at it trying to keep sentences diverse and “adding styles”, whereas the *Intersection* is only keeping the common words and is therefore decreasing the lexical diversity.

TABLE II: Linguistic and sensitive-attribute classifier results for Blogs data, considering *two* sensitive age domains of teens and adults. For BT accuracy and entropy higher is better, for PPL and Confident Response (CR) lower is better.

Metric	Original	A4NT	Intersection			Union
			No DB	DB = 25	DB = 40	
Ling.	BT Accuracy (%)	100.00	66.49	95.41	87.39	88.63
	GPT-2 PPL	41.71	44.85	41.6	42.80	58.15
	Lex. Div. (%)	3.22	2.28	2.50	1.47	0.97
Clf.	Clf. Accuracy (%)	64.73	61.31	64.23	60.90	59.81
	Entropy	0.87	0.86	0.87	0.93	0.95
	CR (%)	14.21	15.72	13.95	4.78	2.47

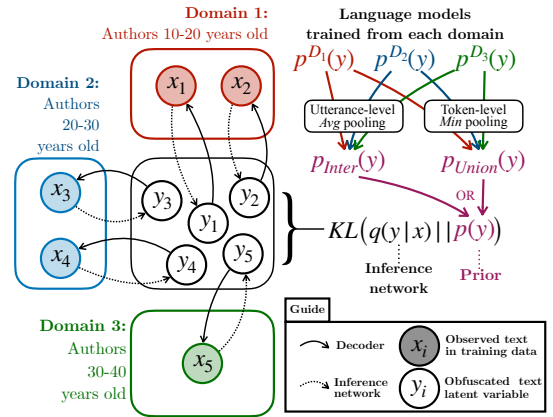


Fig. 1: Proposed unsupervised framework for *style pooling*: inducing a centralized obfuscated style. x_i represent observed text which are clustered by their sensitive attribute (age). y_i are corresponding latent variables representing the induced obfuscated text. Training leverages an amortized inference setup similar to a VAE-style training, but, critically the prior is produced by pooling language models from each domain using two different strategies targeting (1) intersected style and (2) the union of all styles in the corpus.

REFERENCES

- [1] Fatemehsadat Mireshghallah and Taylor Berg-Kirkpatrick, “Style pooling: Automatic text style obfuscation for improved classification fairness,” in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), Selected for Oral Presentation*, November 2021.
- [2] Rakshith Shetty, Bernt Schiele, and Mario Fritz, “A4nt: Author attribute anonymity by adversarial training of neural machine translation,” in *27th USENIX Security Symposium (USENIX Security 18)*, Baltimore, MD, Aug. 2018, pp. 1633–1650, USENIX Association.