Fair NLP Models with Differentially Private Text Encoders

Gaurav Maheshwari, Pascal Denis, Mikaela Keller, Aurélien Bellet

Fair NLP Models with Differentially Private Text Encoders

Gaurav Maheshwari, Pascal Denis, Mikaela Keller, Aurélien Bellet
Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
first_name.last_name@inria.fr

Abstract

Encoded text representations often capture sensitive attributes about individuals (e.g., race or gender), which raise privacy concerns and can make downstream models unfair to certain groups. In this work, we propose FEDERATE, an approach that combines ideas from differential privacy and adversarial training to learn private text representations which also induces fairer models. We empirically evaluate the trade-off between the privacy of the representations and the fairness and accuracy of the downstream model on four NLP datasets. Our results show that FEDERATE consistently improves upon previous methods, and thus suggest that privacy and fairness can positively reinforce each other.

1 Introduction

Algorithmically-driven decision-making systems raise fairness concerns (Raghavan et al., 2020; van den Broek et al., 2019) as they can be discriminatory against specific groups of people. These systems have also been shown to leak sensitive information about the data of individuals used for training or inference, and thus pose privacy risks (Shokri et al., 2017). Societal pressure as well as recent regulations push for enforcing both privacy and fairness in real-world deployments, which is challenging as these notions are multi-faceted concepts that need to be tailored to the context. Moreover, privacy and fairness can be at odds with one another: recent studies have shown that preventing a model from leaking information about its training data negatively impacts the fairness of the model and vice versa (Bagdasaryan et al., 2019; Pujol et al., 2020; Cummings et al., 2019; Chang and Shokri, 2020).

This paper studies fairness and privacy and their interplay in the NLP context, where these two notions have often been considered independently from one another. Modern NLP heavily relies on learning or fine-tuning encoded representations of text. Unfortunately, such representations often leak sensitive attributes (e.g., gender, race, or age) present explicitly or implicitly in the input text, even when such attributes are known to be irrelevant to the task (Song and Raghunathan, 2020). Moreover, the presence of such information in the representations may lead to unfair downstream models, as has been shown on various NLP tasks such as occupation prediction from text bios (De-Arteaga et al., 2019), coreference resolution (Zhao et al., 2018), or sentiment analysis (Kiritchenko and Mohammad, 2018).

Privatizing encoded representations is thus an important, yet challenging problem for which existing approaches based on subspace projection (Bolukbasi et al., 2016; Wang et al., 2020; Karve et al., 2019; Ravfogel et al., 2020) or adversarial learning (Li et al., 2018; Coavoux et al., 2018; Han et al., 2021) do not provide a satisfactory solution. In particular, these methods lack any formal privacy guarantee, and it has been shown that an adversary can still recover sensitive attributes from the resulting representations with high accuracy (Elazar and Goldberg, 2018; Gonen and Goldberg, 2019).

Instead of relying on adversarial learning to prevent attribute leakage, Lyu et al. (2020); Plant et al. (2021) recently propose to add random noise to text representations so as to satisfy differential privacy (DP), a mathematical definition which comes with rigorous guarantees (Dwork et al., 2006). However, we uncover a critical error in their privacy analysis which drastically weakens their privacy claims. Moreover, their approach harms accuracy and fairness compared to adversarial learning.

In this work, we propose a novel approach (called FEDERATE) to learn private text representations and fair models by combining ideas from DP with an adversarial training mechanism. More specifically, we propose a flexible end-to-end architecture in which (i) the output of an arbitrary text

encoder is normalized and perturbed using random noise to make the resulting encoder differentially private, and (ii) on top of the encoder, we combine a classifier branch with an adversarial branch to actively induce fairness, improve accuracy and further hide specific sensitive attributes.

We empirically evaluate the privacy-fairness-accuracy trade-offs achieved by FEDERATE over four datasets and find that it simultaneously leads to more private representations and fairer models than state-of-the-art methods while maintaining comparable accuracy. Beyond the superiority of our approach, our results bring valuable insights on the complementarity of DP and adversarial learning and the compatibility of privacy and fairness. On the one hand, DP drastically reduces undesired leakage from adversarially trained representations, and has a stabilizing effect on the training dynamics of adversarial learning. On the other hand, adversarial learning improves the accuracy and fairness of models trained over DP text representations.

Our main contributions are as follows:

- We propose a new approach, FEDERATE, which combines a DP encoder with adversarial learning to learn fair and accurate models from private representations.
- We identify and fix (with a formal proof) a critical mistake in the privacy analysis of previous work on learning DP text representations.
- We empirically show that FEDERATE leads to more private representations and fairer models than state-of-the-art methods while maintaining comparable accuracy.
- Unlike previous studies, our empirical results suggest that privacy and fairness are compatible in our setting, and even mutually reinforce each other.

The paper is organized as follows. Section 2 provides background on differential privacy. Section 3 presents our approach. Section 4 reviews related work. Experimental results and conclusions are given in Sections 5 and 6.

2 Background: Differential Privacy

Differential Privacy (DP) (Dwork et al., 2006) provides a rigorous mathematical definition of the privacy leakage associated with an algorithm. It does not depend on assumptions about the attacker's capabilities and comes with a powerful algorithmic framework. For these reasons, it has become a de-facto standard in privacy currently used by the US Census Bureau (Abowd, 2018) and several big tech companies (Erlingsson et al., 2014; Fanti et al., 2016; Ding et al., 2017). This section gives a brief overview of DP, focusing on the aspects needed to understand our approach (see Dwork and Roth (2014) for an in-depth review of DP).

Over the last few years, two main models for DP have emerged: (i) Central DP (CDP) (Dwork et al., 2006), where raw user data is collected and processed by a trusted curator, which then releases the result of the computation to a third party or the public, and (ii) Local DP (LDP) (Kasiviswanathan et al., 2011) which removes the need for a trusted curator by having each user locally perturb their data before sharing it. Our work aims to create an encoder that leads to a private embedding of an input text, which can then be shared with an untrusted curator for learning or inference. We thus consider LDP, defined as follows.

Definition 1 (Local Differential Privacy)**.** A randomized algorithm $M : X \to O$ is ϵ-differentially private if for all pairs of inputs $x, x' \in X$ and all possible outputs $o \in O$:

$$\Pr[M(x) = o] \leq e^{\epsilon} \Pr[M(x') = o]. \quad (1)$$

LDP ensures that the probability of observing a particular output o of M should not depend too much on whether the input is x or x'. The strength of privacy is controlled by ϵ, which bounds the log-ratio of these probabilities for any x, x'. Setting $\epsilon = 0$ corresponds to perfect privacy, while $\epsilon \to \infty$ does not provide any privacy guarantees (as one may be able to uniquely associate an observed output to a particular input). In our approach described in Section 3, x will be an input text and M will be an encoding function which transforms x into a private vector representation that can be safely shared with untrusted parties.

Laplace mechanism. As clearly seen from Definition 1, an algorithm needs to be randomized to satisfy DP. A classical approach to achieve ϵ-DP for vector data is the Laplace mechanism (Dwork et al., 2006). Given the desired privacy guarantee ϵ and an input vector $\mathbf{x} \in \mathbb{R}^D$, this mechanism adds centered Laplace noise $\text{Lap}(\frac{\Delta}{\epsilon})$ independently to each dimension of $\mathbf{x}$. The noise scale $\frac{\Delta}{\epsilon}$ is calibrated to ϵ and the *L1-sensitivity* Δ of inputs:

$$\Delta = \max_{\mathbf{x},\mathbf{x}' \in X} \|\mathbf{x} - \mathbf{x}'\|_1. \quad (2)$$

In our work, we propose an architecture in which the Laplace mechanism is applied on top of a trainable encoder to get private representations of input texts, and is further combined with adversarial training to learn fair models.

3 Approach

We consider a scenario similar to Coavoux et al. (2018), where a user locally encodes its input data (text) x into an intermediate representation $E_{priv}(x)$ which is then shared with an untrusted curator to predict the label y associated with x using a classifier C. Additionally, an attacker (which may be the untrusted curator or an eavesdropper) may observe the intermediate representation $E_{priv}(x)$ and try to infer some sensitive (discrete) attribute z about x (e.g., gender, race etc.). Our goal is to learn an encoder E_{priv} and classifier C such that (i) the attacker performs poorly at inferring z from $E_{priv}(x)$, (ii) the classifier $C(E_{priv}(x))$ is fair with respect to z according to some fairness metric, and (iii) C accurately predicts the label y.

To achieve the above goals we introduce FEDERATE (for Fair modEls with DiffERentiAlly private Text Encoders), which combines two components: a differentially private encoder and an adversarial branch. Figure 1 shows an overview of our proposed architecture.

3.1 Differentially Private Encoder

We propose a generic private encoder construction $E_{priv} = priv \circ E$ composed of two main components. The first component E can be any encoder which maps the text input to some vector space of dimension D. It can be a pre-trained language model along with a few trainable layers, or it can be trained from scratch. The second component $priv$ is a randomized mapping which transforms the encoded input to a differentially private representation. Given the desired privacy guarantee $\epsilon > 0$, this mapping is obtained by applying the Laplace mechanism (see Section 2) to a normalized version of the encoded representation $E(x)$:

$$priv(E(x)) = E(x)/\|E(x)\|_1 + \boldsymbol{\ell}, \quad (3)$$

where each entry of $\boldsymbol{\ell} \in \mathbb{R}^D$ is sampled independently from Lap($\frac{2}{\epsilon}$). We will prove that $E_{priv} = priv \circ E$ satisfies ϵ-DP in Section 3.4.

3.2 Adversarial Component

To improve the fairness of the downstream classifier C, we model the adversary by another classifier A which aims to predict z from the privately encoded input $E_{priv}(x)$. The encoder E_{priv} is optimized to fool A while maximizing the accuracy of the downstream classifier C. Specifically, given $\lambda > 0$, we train E_{priv}, C and A (parameterized by θ_E, θ_C, and θ_A respectively) to optimize the following objective:

$$\min_{\theta_E,\theta_C} \max_{\theta_A} \mathcal{L}_{class}(\theta_E, \theta_C) - \lambda \mathcal{L}_{adv}(\theta_E, \theta_A), \quad (4)$$

where $\mathcal{L}_{class}(\theta_E, \theta_C)$ is the cross-entropy loss for the $C \circ E_{priv}$ branch and $\mathcal{L}_{adv}(\theta_E, \theta_A)$ is the cross-entropy loss for the $A \circ E_{priv}$ branch.

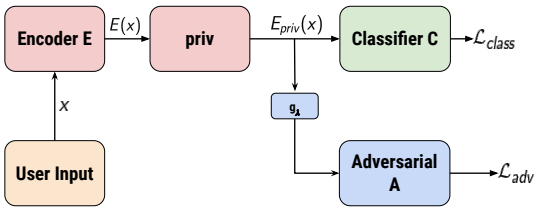

Figure 1: Overview of our FEDERATE approach. The text input x is transformed to $E(x) \in \mathbb{R}^D$ by the text encoder E. The encoded input is then made private by the privacy layer $priv$, which involves normalization and addition of Laplace noise. The resulting private representation $E_{priv}(x) \in \mathbb{R}^D$ is then used by the main task classifier C. It also serves as input to the adversarial layer A which is connected to the main branch via a radient reversal layer g_λ. The light red boxes represent the Differentially Private Encoder (Sec. 3.1), and the light blue boxes represent the Adversarial component (Sec. 3.2).

3.3 Training

We train the private encoder E_{priv} and the classifier C from a set of public tuples (x, y, z) by optimizing (4) with backpropagation using a gradient reversal layer g_λ (Ganin and Lempitsky, 2015). The latter acts like an identity function in the forward pass but scales the gradients passed through it by $-\lambda$ in the backward pass. This results in E_{priv} receiving opposite gradients to A. We give pseudo-code in Appendix A.

3.4 Privacy Analysis

We show the following privacy guarantee.

Theorem 1. *Our encoder E_{priv} and the downstream predictions $C \circ E_{priv}$ satisfy ϵ-DP.*

The proof is given in Appendix B. Theorem 1 shows that the encoded representations produced by E_{priv} have provable privacy guarantees: in particular, it bounds the risk that the sensitive attribute

z of a text x is leaked by $E_{priv}(x)$.[1] These privacy guarantees naturally extend to the downstream prediction $C(E_{priv}(x))$ due to the post-processing properties of DP (see Appendix B for details).

Error in previous work. We found a critical error in the privacy analysis of previous work on differential private text encoders (Lyu et al., 2020; Plant et al., 2021). In a nutshell, they incorrectly state that normalizing each entry of the encoded representation in $[0, 1]$ allows to bound the sensitivity of their representation by 1, while it can in fact be as large as D (the dimension of the representation). As a result, the privacy guarantees are dramatically weaker than what the authors claim: the ϵ values they report should be multiplied by D. In contrast, the L1 normalization we use in (3) ensures that the sensitivity of E is bounded by 2. We provide more details in Appendix C.

Interestingly, Habernal (2021) recently identified an error in ADePT (Krishna et al., 2021), a differentially private auto-encoder for text rewriting. However, the error in ADePT is different from the one in Lyu et al. (2020); Plant et al. (2021): the problem with ADePT is that it calibrates the noise to L2 sensitivity, while the Laplace mechanism requires L1 sensitivity. These errors call for greater scrutiny of differential privacy-based approaches in NLP—our work contributes to this goal.

4 Related Work

Adversarial learning. In order to improve model fairness or to prevent leaking sensitive attributes, several approaches employ adversarial-based training. For instance, Li et al. (2018) propose to use a different adversary for each protected attribute, while Coavoux et al. (2018) consider additional loss components to improve the privacy-accuracy trade-off of the learned representation. Han et al. (2021) introduce multiple adversaries focusing on different aspects of the representation by encouraging orthogonality between pairs of adversaries. Recently, Chowdhury et al. (2021) propose an adversarial scrubbing mechanism. However, they purely focus on information leakage, and not on fairness. Moreover, unlike our approach, these methods do not offer formal privacy guarantees. In fact, it has been observed that one can recover the sensitive attributes from the representations by training a post-hoc non linear classifier (Elazar and Goldberg, 2018). This is confirmed by our empirical results in Section 5.

Sub-space projection. A related line of work focuses on debiasing text representations using projection methods (Bolukbasi et al., 2016; Wang et al., 2020; Karve et al., 2019). The general approach involves identifying and removing a sub-space associated with sensitive attributes. However, they rely on a manual selection of words in the vocabulary which is difficult to generalize to new attributes. Furthermore, Gonen and Goldberg (2019) showed that sensitive attributes still remain present even after applying these approaches.

Recently, Ravfogel et al. (2020) propose Iterative Null space Projection (INLP). It involves iteratively training a linear classifier to predict sensitive attributes followed by projecting the representation on the classifier's null space. However, the method can only remove linear information from the representation. By leveraging DP, our approach provides robust guarantees that do not depend on the expressiveness of the adversary, thereby providing protection against a wider range of attacks.

DP and fairness. Recent work has studied the interplay between DP and (group) fairness in the setting where one seeks to prevent a model from leaking information about individual training points. Empirically, this is evaluated through membership inference attacks, where an attacker uses the model to determine whether a given data point was in the training set (Shokri et al., 2017). While Kulynych et al. (2022) observed that DP reduces disparate vulnerability to such attacks, it has also been shown that DP can exacerbate unfairness (Bagdasaryan et al., 2019; Pujol et al., 2020). Conversely, Chang and Shokri (2020) showed that enforcing a fair model leads to more privacy leakage for the unprivileged group. This tension between DP and fairness is further confirmed by a formal incompatibility result between ϵ-DP and fairness proved by Cummings et al. (2019), albeit in a restrictive setting. Some recent work attempts to train models under both DP and fairness constraints (Cummings et al., 2019; Xu et al., 2020; Liu et al., 2020), but this typically comes at the cost of enforcing weaker privacy guarantees for some groups. Finally, Jagielski et al. (2019) train a fair model under DP constraints only for the sensitive attribute.

A fundamental difference between this line of work and our approach lies in the kind of privacy

[1] More generally, the DP guarantee bounds the risk that *any* attribute of x is leaked through $E_{priv}(x)$.

we provide. While the above approaches study (central) DP as a way to design algorithms which protect training points from membership inference attacks on the model, we construct a private encoder such that the encoded representation does not leak sensitive attributes of the input. Thus, unlike previous work, we provide privacy guarantees with respect to the model's intermediate representation for data unseen at training time, and empirically observe that in this case privacy and fairness are compatible and even mutually reinforce each other.

DP representations for NLP. In a setting similar to ours, Lyu et al. (2020) propose to use DP to privatize model's intermediate representation. Unlike their method, we actively promote fairness by using an adversarial training mechanism, which leads to more private representations and fairer models in practice. Importantly, we also uncover a critical error in their privacy analysis (see Sec. 3.1).

Concurrent to and independently from our work, Plant et al. (2021) propose an adversarial-driven DP training mechanism. However, they do not consider fairness, whereas we focus on enforcing both fairness and privacy. Moreover, their method has the same incorrect analysis as Lyu et al. (2020).

5 Experiments

Recall that we are interested in approaches that are not only accurate but also fair and private at the same time. However, these three dimensions are not independent and are not straightforwardly amenable to a single evaluation metric. Thus, we present experiments aiming at (i) showcasing the privacy-fairness-accuracy tradeoffs of different approaches and then (ii) analyzing privacy-accuracy and fairness-accuracy tradeoffs separately. We begin by describing the datasets and the metrics.

Datasets. We consider 4 different datasets: (i) *Twitter Sentiment* (Blodgett et al., 2016) consists of 200k tweets annotated with a binary sentiment label and a binary "race" attribute corresponding to African American English (AAE) vs. Standard American English (SAE) speakers; (ii) *Bias in Bios* (De-Arteaga et al., 2019) consists of 393,423 textual biographies annotated with an occupation label (28 classes) and a binary gender attribute; (iii) *CelebA* (Liu et al., 2015) is a binary classification dataset with a binary sensitive attribute (gender); (iv) *Adult Income* (Kohavi, 1996) consists of 48,842 instances with binary sensitive attribute (gender). Our setup for the first two dataset is similar to Ravfogel et al. (2020) and Han et al. (2021). Appendix D.2 provides detailed description of these datasets, including sizes, pre-processing, and the challenges they pose to privacy and fairness tasks. Due to lack of space, results and analyses for *Adult Income* and *CelebA dataset* are given in Appendix D.5, but note that they exhibit similar trends. The preprocessed versions of the datasets can be downloaded from this anonymized URL.[2]

Fairness metrics. For Twitter Sentiment we report the True Positive Rate Gap (TPR-gap), which measures the true positive rate difference between the two sensitive groups (gender/race) and is closely related to the notion of equal opportunity. Formally, denoting by $y \in \{0, 1\}$ the ground truth binary label, $\hat{y}$ the predicted label and $z \in \{g, \neg g\}$ the sensitive attribute, TPR-gap is defined as:

$$\text{TPR-gap} = P_g(\hat{y} = 1|y = 1) - P_{\neg g}(\hat{y} = 1|y = 1).$$

For Bias in Bios, which has 28 classes, we follow Romanov et al. (2019) and report the root mean square of TPR-gaps (GRMS) over all occupations $y \in O$ to obtain a single number:

$$\text{GRMS} = \sqrt{(1/|O|)\sum_{y\in O}(\text{TPR-gap}_y)^2}. \quad (5)$$

Privacy metrics. We report two metrics for privacy: (i) **Leakage**: the accuracy of a two-layer classifier which predicts the sensitive attribute from the encoded representation, and (ii) **Minimum Description Length (MDL)** (Voita and Titov, 2020), which quantifies the amount of "effort" required by such a classifier to achieve a certain accuracy. A higher MDL means that it is more difficult to retrieve the sensitive attribute from the representation. The metric depends on the dataset and the representation dimension, and thus cannot be compared across different datasets. We provide more details about these metrics in Sec. D.1.

Methods and model architectures. We compare **`FEDERATE`** to the following methods: (i) **`Adversarial`** implements standard adversarial learning (Li et al., 2018), which is equivalent to our approach without the $priv$ layer, (ii) **`Adversarial + Multiple`** (Han et al., 2021) implements multiple adversaries, (iii) **`INLP`** (Ravfogel et al., 2020) is a subspace projection approach, and (iv) **`Noise`** learns DP text

[2] https://drive.google.com/uc?id=1ZmUE-g6FmzPPbZyw3EOki7z4bpzbKGWk

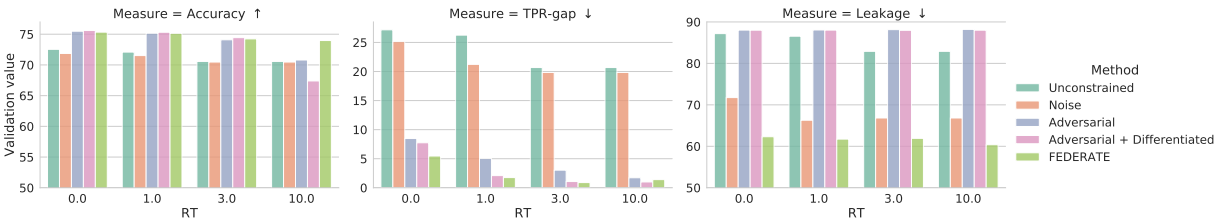

Figure 2: Validation accuracy, fairness and privacy of various approaches for different relaxation threshold (RT) (see Section 5.1) on Twitter Sentiment. When RT is increased, we select models with potentially lower accuracy on the validation set but are more fair (lower TPR-gap). Our approach FEDERATE consistently achieves better accuracy-fairness-privacy trade-offs than its competitors across all RTs.

representations as proposed by Lyu et al. (2020) but with corrected privacy analysis: this corresponds to our approach without the adversarial component. These methods have been described in details in Section 4 and their hyperparametrs in Appendix D.4. We also report the performance of two simple baselines: **Random** simply predicts a random label, and **Unconstrained** optimizes the classification performance without special consideration for privacy or fairness.

To provide a fair comparison, all methods use the same architecture for the encoder, the classifier and (when applicable) the adversarial branches. In order to evaluate across varying model complexities, we employ different architectures for the different datasets. For Twitter Sentiment, we follow the architecture employed by Han et al. (2021), while for Bias in Bios we use a deeper architecture. The exact architecture, hyperparameters, and their tuning details are provided in Appendix D.3-D.4. We implement FEDERATE in PyTorch (Paszke et al., 2019). Our implementation, training, and evaluation scripts are available here.[3]

5.1 Accuracy-Fairness-Privacy Trade-off

In this first set of experiments, we explore the tridimensional trade-off between accuracy, fairness, and privacy and the inherent tension between them. These metrics are potentially all equally important and represent different information on different scales. Thus, they cannot be trivially combined into a single metric. Moreover, this trade-off is influenced by the choice of method but also some of its hyperparameters (e.g., the value of ϵ and λ in our approach). Previous studies (Han et al., 2021; Lyu et al., 2020) essentially selected hyperparameter values that maximize validation accuracy, which may lead to undesirable or suboptimal trade-offs. For instance, we found that this strategy does not always induce a fairer model than the Unconstrained baseline, and that it is often possible to obtain significantly more fair models at a negligible cost in accuracy. Based on these observations, we propose to use a Relaxation Threshold (RT): instead of selecting the hyperparameters with highest validation accuracy α^*, we consider all models with accuracy in the range $[\alpha^* - \text{RT}, \alpha^*]$. We then select the hyperparameters with best fairness score within that range.[4]

Figure 2 presents the (validation) accuracy, fairness and privacy scores related to different RT for each method on Twitter Sentiment. The first thing to note is that FEDERATE achieves the best fairness and privacy results with accuracy higher or comparable to competing approaches. We also observe that setting RT$=0.0$ (i.e., choosing the model with highest validation accuracy) leads to a significantly more unfair model in all approaches, while fairness generally improves with increasing RT. This improvement comes at a negligible or small cost in accuracy. In terms of privacy, we find no significant differences across RTs.

We now showcase detailed results with RT fixed to 1.0 which is found to provide good trade-offs for all approaches in Figure 2, see Table 1a for Twitter Sentiment and Table 1b for Bias in Bios (and Appendix D.6 for additional results). For both datasets, we observe that all adversarial approaches induce a fairer model than Unconstrained or Noise, with FEDERATE performing best. In terms of accuracy, all adversarial approaches perform similarly on Twitter Sentiment. Interestingly, they achieve higher accu-

[3]The work-in-progress version of the codebase is currently available at https://github.com/saist1993/DPNLP.

[4]We can also incorporate privacy into our hyperparameter selection strategy but, for the datasets and methods in our study, we found no significant change in Leakage across different hyperparameters.

Method	Accuracy ↑	TPR-gap ↓	Leakage ↓	MDL ↑
Random	50.00 ± 0.00	0.00 ± 0.00	-	31.3 ± 0.10
Unconstrained	72.09 ± 0.73	26.26 ± 0.87	86.56 ± 0.83	15.21 ± 0.88
INLP	67.62 ± 0.57	9.19 ± 1.08	80.27 ± 2.50	24.82 ± 3.28
Noise	71.52 ± 0.51	21.23 ± 2.50	66.29 ± 3.55	21.10 ± 1.81
Adversarial	75.16 ± 0.65	5.03 ± 2.94	88.06 ± 0.20	16.16 ± 1.05
Adversarial + Multiple	75.32 ± 0.60	2.09 ± 1.18	88.03 ± 0.47	15.85 ± 1.46
FEDERATE	75.15 ± 0.59	1.75 ± 1.41	61.74 ± 5.05	22.94 ± 1.25

(a) Results on Twitter Sentiment dataset.

Method	Accuracy ↑	GRMS ↓	Leakage ↓	MDL ↑
Random	3.53 ± 0.01	0.00 ± 0.00	–	265.44 ± 0.13
Unconstrained	79.29 ± 0.32	15.88 ± 0.80	75.92 ± 2.73	173.99 ± 7.08
INLP	75.96 ± 0.47	12.81 ± 0.09	59.91 ± 0.08	253.36 ± 1.05
Noise	77.88 ± 0.32	13.89 ± 0.31	62.23 ± 0.99	241.22 ± 2.97
Adversarial	79.02 ± 0.20	13.06 ± 0.39	69.47 ± 1.64	206.78 ± 13.02
Adversarial + Multiple	79.30 ± 0.20	13.38 ± 0.63	68.24 ± 1.12	222.35 ± 10.04
FEDERATE	77.79 ± 0.11	11.02 ± 0.55	56.92 ± 0.98	257.94 ± 1.93

(b) Results on Bias in Bios dataset.

Table 1: Test results on (a) *Twitter Sentiment*, and (b) *Bias in Bios* with fixed Relaxation Threshold of 1.0. Fairness is measured with TPR-Gap or GRMS (lower is better), while privacy is measured by Leakage (lower is better) and MDL (higher is better). The MDL achieved by Random gives an upper bound for that particular dataset. Results have been averaged over 5 different seeds. Our proposed FEDERATE approach is the only method which achieves high levels of both fairness and privacy while maintaining competitive accuracy.

racy than Unconstrained. We attribute this to a significant mismatch in the train and test distribution due to class imbalance. On Bias in Bios, we observe a small drop in accuracy of our proposed approach in comparison to Adversarial, albeit with a corresponding gain in fairness. We hypothesize that this is due to the choice of possible hyperparameters for FEDERATE (we did not consider very large values of ϵ which would recover Adversarial), meaning that FEDERATE pushes for more fairness (and privacy) at a potential cost of some accuracy. We explore the pairwise trade-offs (fairness-accuracy and privacy-accuracy) in more details in Section 5.2.

In terms of both privacy metrics, FEDERATE significantly outperforms all adversarial methods on both datasets. In fact, in line with previous studies (Han et al., 2021), the leakage and MDL of purely adversarial methods are similar to that of Unconstrained. On both datasets, Noise achieves slightly weaker privacy than FEDERATE with much worse accuracy and fairness.

FEDERATE also consistently outperforms INLP in all dimensions.

In summary, the results show that FEDERATE stands out as the only approach that can simultaneously induce a fairer model *and* make its representation private while maintaining high accuracy. Furthermore, these results empirically demonstrate that our measures of privacy and fairness are indeed compatible with one another and can even reinforce each other.

5.2 Pairwise Trade-offs

In the previous experiments, we explored the tridimensional trade-off and found FEDERATE to attain better trade-offs than all other methods. Here, we take a closer look at the pairwise fairness-accuracy and privacy-accuracy trade-offs separately. We find that FEDERATE outperforms the Adversarial and Noise approach in their corresponding dimension, suggesting that FEDERATE is a better choice even for bidimensional trade-offs. This experiment also validates the superiority of

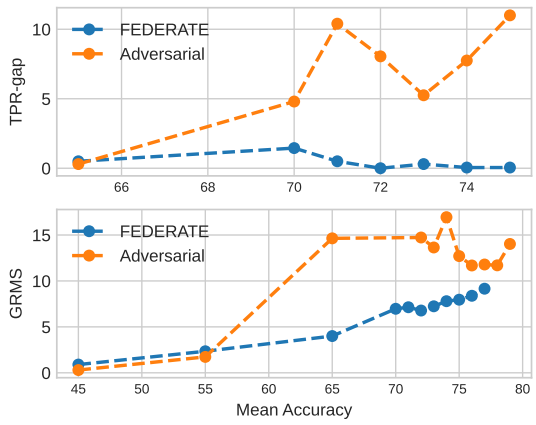

Figure 3: Fairness-accuracy trade-off on Twitter Sentiment (top) and Bias in Bios (bottom). A missing point means that the accuracy interval was not found within our hyperparameter search. FEDERATE provides better fairness across most accuracy intervals in comparison to Adversarial over both datasets.

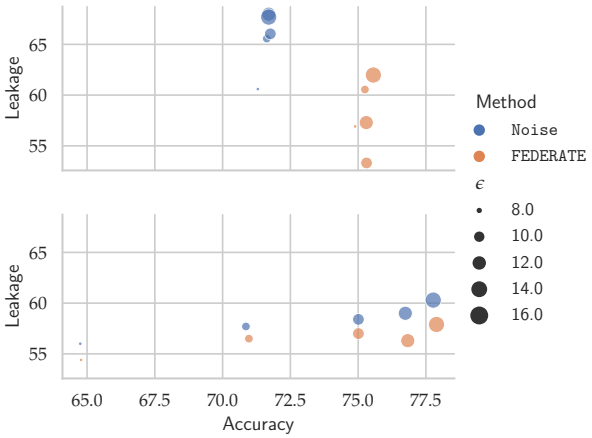

Figure 4: Privacy-accuracy trade-off on Twitter Sentiment (top) and Bias in Bios (bottom), with associated values of ϵ. FEDERATE gives lower leakage and better or comparable accuracy to Noise over both datasets.

combining adversarial learning and DP over using either approach alone.

Fairness-accuracy trade-off. We plot best validation fairness scores over different accuracy intervals for the two datasets in Figure 3. The interval is denoted by its mean accuracy (i.e., $[71.5, 72.5]$ is represented by 72). We then find the corresponding best fairness score for the interval. We observe:

- *Better fairness-accuracy trade-off:* FEDERATE provides better fairness than the Adversarial approach for almost all accuracy intervals. In the case of Bias in Bios, Adversarial is able to achieve higher accuracy (albeit with a loss in fairness). We note that this high accuracy regime can be matched by FEDERATE with a larger ϵ.

- *Smoother fairness-accuracy trade-off:* Interestingly, FEDERATE enables a smoother exploration of the accuracy-fairness trade-off space than Adversarial. As adversarial models are notoriously difficult to train, this suggests that the introduction of DP noise has a stabilizing effect on the training dynamics of the adversarial component.

Privacy-accuracy trade-off. We plot privacy and accuracy with respect to ϵ, the parameter controlling the theoretical privacy level in Figure 4. In general, the value of ϵ correlates well with the empirical leakage. On Bias in Bios, FEDERATE and Noise are comparable in both accuracy and privacy. However, for Twitter Sentiment, our approach outperforms Noise in both accuracy and privacy for every ϵ. We hypothesize this difference in the accuracy to be a case of mismatch between train-test split, suggesting FEDERATE to be more robust to these distributional shifts. These observations suggest that FEDERATE either improves upon Noise in privacy-accuracy tradeoff or remains comparable. For completeness, we also present the same results as a table in Appendix D.6.

6 Conclusion and Perspectives

We proposed a DP-driven adversarial learning approach for NLP. Through our experiments, we showed that our method simultaneously induces private representations and fair models, with a mutually reinforcing effect between privacy and fairness. We also find that our approach improves upon competitors on each dimension separately. While we focused on privatizing sensitive attributes like race or gender, our approach can be used to remove other types of unwanted information from text representations, such as tenses or POS tag information, which might not be relevant for certain NLP tasks.

A possible limitation of this work is that it not tailored to a specific definition of fairness like equal odds. Instead, it enforces fairness by removing certain protected information, which can correlate with specific fairness notions. Similarly, we do not provide any formal fairness guarantees for our method, as we do for privacy. In the future, we aim to investigate fairness methods that explicitly optimize for a specific fairness definition and explore other privacy threats (e.g., reconstruction attacks).

7 Acknowledgement

The authors would like to thank the Agence Nationale de la Recherche for funding this work under grant number ANR-19-CE23-0022, as well as the ARR reviewers for their feedback and suggestions.

References

John M Abowd. 2018. The us census bureau adopts differential privacy. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2867–2867.

Yossi Adi, Neil Zeghidour, Ronan Collobert, Nicolas Usunier, Vitaliy Liptchinsky, and Gabriel Synnaeve. 2019. To reverse the gradient or not: an empirical comparison of adversarial and multi-task learning in speech recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 3742–3746. IEEE.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 15453–15462.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, Texas. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Hongyan Chang and Reza Shokri. 2020. On the privacy risks of algorithmic fairness. *CoRR*, abs/2011.03731.

Somnath Basu Roy Chowdhury, Sayan Ghosh, Yiyuan Li, Junier Oliva, Shashank Srivastava, and Snigdha Chaturvedi. 2021. Adversarial scrubbing of demographic information for text classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 550–562. Association for Computational Linguistics.

Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018. Privacy-preserving neural representations of text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 1–10. Association for Computational Linguistics.

Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, UMAP 2019, Larnaca, Cyprus, June 09-12, 2019*, pages 309–315. ACM.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Bolin Ding, Janardhan Kulkarni, and Sergey Yekhanin. 2017. Collecting telemetry data privately. In *NIPS*.

Dheeru Dua and Casey Graff. 2017. UCI machine learning repository.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer.

Cynthia Dwork and Aaron Roth. 2014. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3-4):211–407.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 11–21. Association for Computational Linguistics.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. 2014. Rappor: Randomized aggregatable privacy-preserving ordinal response. In *CCS*.

Giulia Fanti, Vasyl Pihur, and Úlfar Erlingsson. 2016. Building a RAPPOR with the unknown: Privacy-preserving learning of associations and data dictionaries. In *PoPETs*.

Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. 2017. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 1615–1625. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. PMLR.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Ivan Habernal. 2021. When differential privacy meets NLP: the devil is in the detail. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1522–1528. Association for Computational Linguistics.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2760–2765. Association for Computational Linguistics.

Matthew Jagielski, Michael J. Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan R. Ullman. 2019. Differentially private fair learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 3000–3008. PMLR.

Saket Karve, Lyle Ungar, and João Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 40–48, Florence, Italy. Association for Computational Linguistics.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam D. Smith. 2011. What can we learn privately? *SIAM J. Comput.*, 40(3):793–826.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Ron Kohavi. 1996. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*, pages 202–207. AAAI Press.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. ADePT: Auto-encoder based differentially private text transformation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2435–2439, Online. Association for Computational Linguistics.

Bogdan Kulynych, Mohammad Yaghini, Giovanni Cherubin, Michael Veale, and Carmela Troncoso. 2022. Disparate vulnerability to membership inference attacks. In *PETS*.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5967–5976.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30, Melbourne, Australia. Association for Computational Linguistics.

Wenyan Liu, Xiangfeng Wang, Xingjian Lu, Junhong Cheng, Bo Jin, Xiaoling Wang, and Hongyuan Zha. 2020. Fair differential privacy can mitigate the disparate impact on model accuracy.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society.

Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. 2020. Too relaxed to be fair. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6360–6369. PMLR.

Lingjuan Lyu, Xuanli He, and Yitong Li. 2020. Differentially private representation for NLP: formal guar-

antee and an empirical study on privacy and fairness. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2355–2365. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Richard Plant, Dimitra Gkatzia, and Valerio Giuffrida. 2021. Cape: Context-aware private embeddings for private language learning. *arXiv preprint arXiv:2108.12318*.

David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. 2020. Fair decision making using privacy-protected data. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 189–199. ACM.

Manish Raghavan, Solon Barocas, Jon M. Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 469–481. ACM.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7237–7256. Association for Computational Linguistics.

Alexey Romanov, Maria De-Arteaga, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, Anna Rumshisky, and Adam Kalai. 2019. What's in a name? Reducing bias in bios without access to protected attributes. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4187–4195, Minneapolis, Minnesota. Association for Computational Linguistics.

Reza Shokri and Vitaly Shmatikov. 2015. Privacy-preserving deep learning. In *CCS*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 3–18. IEEE Computer Society.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *CCS '20: 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, USA, November 9-13, 2020*, pages 377–390. ACM.

Elmira van den Broek, Anastasia Sergeeva, and Marleen Huysman. 2019. Hiring algorithms: An ethnography of fairness in practice. In *Proceedings of the 40th International Conference on Information Systems, ICIS 2019, Munich, Germany, December 15-18, 2019*. Association for Information Systems.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 183–196. Association for Computational Linguistics.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5443–5453. Association for Computational Linguistics.

Yongkai Wu, Lu Zhang, and Xintao Wu. 2019. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3356–3362. ACM.

Depeng Xu, Wei Du, and Xintao Wu. 2020. Removing disparate impact of differentially private stochastic gradient descent on model accuracy. *CoRR*, abs/2003.03699.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers)*, pages 15–20. Association for Computational Linguistics.

APPENDIX

A Training Algorithm

We provide the pseudo-code of the training procedure of FEDERATE in Algorithm 1. Note that the combination of Steps 2-3-4 corresponds to E_{priv} in Sec. 3.

B Proof of Theorem 1

Proof. We start by proving that our noisy encoder $E_{priv} : X \to \mathbb{R}^D$ satisfies ϵ-DP. Recall that for any input text $x \in X$

$$E_{priv}(x) = priv \circ E(x) = E(x)/\|E(x)\|_1 + \boldsymbol{\ell},$$

where each entry of $\boldsymbol{\ell} \in \mathbb{R}^D$ is sampled independently from Lap($\frac{2}{\epsilon}$), the centered Laplace distribution with scale $2/\epsilon$. Let $\tilde{E}(x) = E(x)/\|E(x)\|_1$. The L1 sensitivity of $\tilde{E}$ is

$$\Delta_{\tilde{E}} = \max_{x,x' \in X} \|\tilde{E}(x) - \tilde{E}(x)'\|_1.$$

Since for any $x \in X$ we have $\|\tilde{E}(x)\|_1 = 1$, the triangle inequality gives $\Delta_{\tilde{E}} \leq 2$. The ϵ-DP guarantee then follows from the application of the Laplace mechanism (Dwork et al., 2006). Formally, let

$$p(y) = \frac{\epsilon}{4} e^{-\frac{|y|\epsilon}{2}}$$

denote the p.d.f. of Lap($2/\epsilon$). Consider two arbitrary input texts $x, x' \in X$ and let $\tilde{\mathbf{x}} = \tilde{E}(x) \in \mathbb{R}^D$ and $\tilde{\mathbf{x}}' = \tilde{E}(x') \in \mathbb{R}^D$ be their normalized encoded representations. Then, for any possible encoded output $\mathbf{e} = (e_1, \ldots, e_D) \in \mathbb{R}^D$, we have:

$$\begin{aligned}
\frac{\Pr[E_{priv}(x) = \mathbf{e}]}{\Pr[E_{priv}(x') = \mathbf{e}]} &= \prod_{d=1}^{D} \frac{p(e_d - \tilde{x_d})}{p(e_d - \tilde{x_d}')} && (6)\\
&= \prod_{d=1}^{D} \frac{e^{-\frac{\epsilon}{2}|e_d - \tilde{x}_d|}}{e^{-\frac{\epsilon}{2}|e_d - \tilde{x}'_d|}} \\
&= e^{\frac{\epsilon}{2}\sum_{d=1}^{D} |e_d - \tilde{x}'_d| - |e_d - \tilde{x}_d|} \\
&\leq e^{\frac{\epsilon}{2}\sum_{d=1}^{D} |\tilde{x}_d - \tilde{x}'_d|} && (7)\\
&= e^{\frac{\epsilon}{2}\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}'\|_1} \\
&\leq e^{\frac{\epsilon}{2}\Delta_{\tilde{E}}} = e^{\epsilon}, && (8)
\end{aligned}$$

where (6) follows from the independence of the noise across dimensions, (7) uses the triangle inequality, and (8) from the definition of $\Delta_{\tilde{E}}$ and the fact that $\Delta_{\tilde{E}} \leq 2$ as shown above.

The above inequality shows that E_{priv} satisfies ϵ-DP as per Definition 1. The fact that $C \circ E_{priv}$ also satisfies ϵ-DP follows from the post-processing property of DP, which ensures that the composition of any function with an ϵ-DP algorithm also satisfies ϵ-DP (Dwork and Roth, 2014).

□

C Error in Privacy Analysis of Previous Work

As briefly mentioned in Section 4, we found a critical error in the differential privacy analysis made in previous work by Lyu et al. (2020). This error is then reproduced in subsequent work by Plant et al. (2021). In this section, we explain this error and its consequences for the formal privacy guarantees of these methods, and provide a correction.

Recall from Section 2 that to achieve ϵ-DP with the Laplace mechanism, one must calibrate the scale of the Laplace noise needed to the L1 sensitivity of the encoded representation (see Eq. 2). This sensitivity bounds the worst-case change in L1 norm for any two arbitrary encoded user inputs $\mathbf{x}$ and $\mathbf{x}'$ of dimension D.

In order to bound the L1 sensitivity, Lyu et al. (2020) and Plant et al. (2021) propose to bound each entry of the encoded input $\mathbf{x} \in \mathbb{R}^D$ in the $[0, 1]$ range. Specifically, they normalize as follows:

$$\mathbf{x} \leftarrow \mathbf{x} - \min(\mathbf{x})/(\max(\mathbf{x}) - \min(\mathbf{x})), \qquad (9)$$

where $\min(\mathbf{x})$ and $\max(\mathbf{x})$ are respectively the minimum and maximum values in the vector $\mathbf{x}$. Lyu et al. (2020) and Plant et al. (2021) incorrectly claim that this allows to bound the L1 sensitivity by 1 and thus add Laplace noise of scale $\frac{1}{\epsilon}$. In fact, the sensitivity can be as large as D, as can be seen by considering the two inputs $\mathbf{x} = [0, 1, \ldots, 1]_D$ and $\mathbf{x}' = [1, 0, \ldots, 0]$ for which $\|\mathbf{x} - \mathbf{x}'\|_1 = D$. Therefore, to achieve ϵ-DP, the scale of the Laplace noise should be $\frac{D}{\epsilon}$ (i.e., D times larger than what the authors use). As a consequence, the differential privacy provided by their method are D times worse than claimed by Lyu et al. (2020) and Plant et al. (2021): the ϵ values they report should be multiplied by D, which leads to essentially void privacy guarantees.

While Lyu et al. (2020) claim to follow the approach of Shokri and Shmatikov (2015), they missed the fact that Shokri and Shmatikov (2015) do account for multiple dimensions by scaling the noise to the number of entries (denoted by c in their paper) that are submitted to the server, see

Algorithm 1: Training procedure of `FEDERATE` (one epoch).

Input: Model architecture composed of encoder E (parameterized by θ_E), classifier C (parameterized by θ_C), adversary A (parameterized by θ_A), loss function L

Output: Trained model

Data: Samples S=$\{x^i, y^i, z^i\}_{i=1}^m$ where x^i is the input text, y^i is the task label, and z^i is the sensitive attribute.

for $i \leftarrow 0$ **to** m **do**

`// For each sample in the dataset. This can be batch too.`

Encode: $\mathbf{x}^i \leftarrow E(x^i)$

Normalize: $\mathbf{x}^i \leftarrow \frac{\mathbf{x}^i}{\|\mathbf{x}^i\|_1}$

Privatize: $\mathbf{x}^i_{priv} \leftarrow \mathbf{x}^i + \boldsymbol{\ell}$, where each entry of the vector $\boldsymbol{\ell} \in \mathbb{R}^D$ is sampled independently from a centered Laplace distribution with scale $\frac{2}{\epsilon}$

Adversarial prediction: $\hat{z}^i \leftarrow A(\mathbf{x}^i_{priv})$

Update θ_A by backpropagating the loss $L(z^i, \hat{z}^i)$

Task classification: $\hat{y}^i \leftarrow C(\mathbf{x}^i_{priv})$

Update θ_E and θ_C by backpropagating the loss $L(y^i, \hat{y}^i) - \lambda \cdot L(z^i, \hat{z}^i)$

pseudo-code in Figure 12 of Shokri and Shmatikov (2015).

In contrast to Lyu et al. (2020) and Plant et al. (2021), our normalization in Eq. 3 guarantees by design that the L1 sensitivity is bounded by 2. We provide a complete and self-contained proof of our privacy guarantees in Section B.

D Experiments

This section gives more information on the experimental setup and also provides additional results.

D.1 Privacy metric

Leakage: We compute the leakage using a sklearn's MLPClassifier. We use the validation set of the original dataset as the train and the test set of the original dataset as the test.

Minimum Description Length (MDL) is a information-theoretic probing measure which captures the strength of regularity in the data. In this work, we employ the online coding approach (Voita and Titov, 2020) to calculate MDL. Online coding captures the regularity by characterizing the effort required to achieve a certain level of accuracy. Here, a portion of data is transmitted to the receiver at each step, which then uses all the data in the previous steps to understand the regularity in the current step. The regularity is obtained by training the model on the previously received data and then evaluating it on the current portion of the data.

Borrowing, the terminology from Voita and Titov (2020), consider a dataset D consisting of $\{(x_1, y_1), \cdots, (x_n, y_n)\}$ pairs, where the x_i's are the data representation, and the y_i's are the task label. In our case, x_i is the output of the encoder, and y_i is the sensitive attribute associated with the underlying text. Following the standard information theory setting, consider a sender Alice who wants to transmit labels $y_{1:n} = \{y_1 \cdots, y_n\}$ to a receiver Bob, and both of them have access to the data representation $x_{1:n} = \{x_1 \cdots, x_n\}$. In order to transmit labels $y_{1:n}$ efficiently (as few bits possible), Alice encodes $y_{1:n}$ using a model $p(y|x)$. According to Shannon-Huffman code, the minimum bits required to transmit these labels losslessly is:

$$L_p(y_{1:n}|x_{1:n}) = -\sum_{i=1}^{n} \log_2 p(y_i|x_i).$$

In the online coding setting of MDL, the labels are transmitted in blocks of n timesteps $t_0 < t_1 < \cdots t_n$. Alice starts by encoding $y_{1:t_1}$ with a uniform code, then both Alice and Bob learn a model $p_{\theta_1}(y|x)$ that predicts y from x using data $\{(x_i, y_i)\}_{i=1}^{t1}$. Alice then uses this model to communicate the next data block $y_{t_1:t_2}$, and both learns a new model using larger chunk of data $\{(x_i, y_i)\}_{i=1}^{t2}$. This continues till the whole set of labels $y_{1:n}$ is transmitted. The total code length required for transmission using this setting is given

as:

$$L_{online}(y_{1:n}|x_{1:n}) = t_1 \log_2 C - \sum_{i=1}^{n-1} \log_2 p_{\theta_i}(y_{t_i+1:t_i}|x_{t_i+1:t_i}). \quad (10)$$

where $y_i \in \{1, 2, \cdots, C\}$. In our case, the online code length $L_{online}(y_{1:n}|x_{1:n})$ is shorter, if it is easier for probing model to perform well with fewer training instances. This implies that the sensitive information is more easily available in the encoder's representation.

We compute MDL using sklearn's MLPClassifier at timesteps corresponding to 0.1%, 0.2%, 0.4%, 0.8%, 1.6%, 3.2%, 6.25%, 12.5%, 25%, 50% and 100% of each dataset as suggested by Voita and Titov (2020).

D.2 Datasets

Twitter Sentiment (Blodgett et al., 2016) consists of 200k tweets annotated with a binary sentiment label and a binary "race" attribute corresponding to African American English (AAE) vs. Standard American English (SAE) speakers. The initial representation of tweets are obtained from a Deepmoji encoder (Felbo et al., 2017). The dataset is evenly balanced with respect to the four sentiment-race subgroup combinations. To create bias in the training data, we follow Elazar and Goldberg (2018) and change the race proportion in each sentiment class to have 40% AAE-happy, 10% AAE-sad, 10% SAE-happy, and 40% SAE-sad. Test data remains balanced. This setup is particularly challenging regarding privacy and fairness, as the model may exploit the correlation between the protected attribute and the main class label, which is reinforced due to skewing. The mismatch between the train-test distribution is also relevant for our setup, where the system may be trained on publicly available datasets or collected via an opt-in policy and may therefore not closely resemble the test distribution. This dataset is made available for research purposes only.[5]

Bias in Bios (De-Arteaga et al., 2019) consists of 393,423 textual biographies annotated with an occupation label (28 classes) and a binary gender attribute. Similar to Ravfogel et al. (2020), we encode each biography with BERT (Devlin et al., 2019), using the last hidden state over the CLS token. We use the same train-valid-test split as De-Arteaga et al. (2019). As the dataset was collected by scrapping the web, it tends to reflect common gender stereotypes and contains explicit gender indicators (e.g., pronouns), making it more challenging to prevent models from relying on these gendered words. It is also more complex than Twitter Sentiment in terms of the number of classes. Dataset is released under MIT License.[6]

CelebA (Liu et al., 2015) consists of over 200,000 images of the human face, alongside with 40 binary attributes labels describing the content of the images. Following the standard setting as described in (Lohaus et al., 2020), we use 38 of these attributes as features, "Smiling" as the class label, and "Sex" as the sensitive attribute. We use 60% of the data as train, 20% as validation, and the remaining as the test split. The CelebA dataset is available for non-commercial research purposes.[7]

Adult Income (Kohavi, 1996) consists of a U.S. 1994 Census database segment and has 48842 instances with 14 features each. We apply the preprocessing as proposed by (Wu et al., 2019) resulting in a total of 9 features for each instance. The objective is to predict whether a given data point earns more than fifty thousand U.S. dollars or less. We consider sex (binary) as the sensitive attribute. Like CelebA, We use 60% of the data as train, 20% as validation, and the remaining as the test split. The license of the dataset is unknown, however it is commonly used in several fairness papers and is avaialbe at (Dua and Graff, 2017).

D.3 Model Architecture

Twitter Sentiment. The encoder consists of two layers with ReLU activation and a fixed dropout of 0.1. The classifier is linear, and the adversarial branch consists of three layers. We use a fixed dropout of 0.1 in all the layers with ReLU activation, apart from the last layer.

Bias in Bios. The encoder consists of three layers and a fixed dropout of 0.1. The classifier also consists of three layers, and the adversarial branch consists of two layers. We use a fixed dropout of 0.1 in all the layers with ReLU activation, apart from the last layer.

[5] http://slanglab.cs.umass.edu/TwitterAAE/

[6] https://github.com/Microsoft/biosbias

[7] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

In case of *Adult Income* and *CelebA* dataset we use the same model as for *Twitter Sentiment*.

D.4 Hyperparameters

For all our experiments, we use Adam optimizer with a learning rate of 0.001 and batch size of 2000. We give additional tuning details of the different methods below. A single experiment takes about 30 minutes to run on Intel Xenon CPU. We will also provide the PyTorch model description in the README of the source code for easier reproduction.

- `Adversarial`: We perform a grid search over λ varying it between 0.1 to 3.0 with an interval of 0.2. Moreover, following previous work (Lample et al., 2017; Adi et al., 2019), instead of a constant λ, we increase it over the epochs using the update scheme $\lambda_i = 2/(1 + e^{-p_i}) - 1$, where p_i is the scaled version of the epoch number. We also experimented with increasing the λ linearly, as well as keeping it constant, but found the above update scheme to perform the best in various settings. We also use this scheme in all other adversarial approaches.

- `Adversarial + Multiple`: Similar to `Adversarial`, we vary λ between 0.1 to 3.0 with an interval of 0.2. Apart from λ, `Adversarial + Multiple` has an additional hyperparameter λ_{ort} which corrresponds to the weight given to the orthogonality loss component. We vary λ_{ort} between 0.1 and 1.0. Here, we do a simultaneous grid search over λ and λ_{ort} resulting in 150 runs for each seed. We fix the number of the adversary to three which is the same as the original implementation by (Han et al., 2021).

- `FEDERATE`: In order to have comparable number of runs to `Adversarial + Multiple`, we experiments with following ϵ values: $8.0, 9.0, 10.0, 11.0, 12.0, 13.0, 14.0, 15.0, 16.0, 20.0$. Similar to above approach, we do a simultaneous grid search over λ and ϵ resulting in 150 runs for each seed.

- `INLP`: In the case of `INLP`, we always debias the representation after the penultimate classifier layer and before the final layer, which is consistent with the setting considered by the authors (Ravfogel et al., 2020). We also observe that this choice empirically led to the best results. We vary the number of iterations as a part of hyperparameter tuning. For Bias in Bios we vary the iterations between 15 and 45, while for Twitter Sentiment we vary between 2 to 7. We found that in case of Bias in Bios, performing less than 15 iterations resulted in the same behaviour as `Unconstrained` model over validation set while more than 45 iterations resulted in a random classifier. We observed the same in the Twitter Sentiment before 2 and after 7 iterations, respectively.

D.5 Extended Evaluation

Tables 2–3 present detailed results on CelebA and Adult Income dataset respectively. In terms of fairness over both the datasets, we observe that adversarial-based approaches induce a more fair model than Unconstrained or Noise, with `FEDERATE` outperforming all other methods. Interestingly, unlike Twitter Sentiment and Bias in Bios, all approaches have comparable accuracy, including `Noise` and `INLP`. We believe this to be the case due to these datasets being relatively more challenging than CelebA and Adult Income. As observed previously, purely adversarial-based approaches leak significantly more information than the DP-based approaches in terms of privacy. We observe that `Noise` and `INLP` performs marginally better in privacy than `FEDERATE`; however, they suffer significantly in the fairness metric. In fact, they induce fairness levels which are similar to `Unconstrained`.

Overall, the results show `FEDERATE` as the only viable choice to induce a fairer model and make its representation private while maintaining comparable accuracy. These observations are in line with previous experiments described in Sec. 5.1

D.6 Additional Results

Tables 4–6 present detailed results on Twitter Sentiment with different relaxation thresholds, which were summarized in Figure 2.

Table 7 provides the detailed privacy-fairness results which were summarized in Figure 4.

Method	Accuracy ↑	TPR-gap ↓	Leakage ↓	MDL ↑
Random	50.00 ± 0.00	0.00 ± 0.00	–	104.64 ± 0.11
Unconstrained	85.70 ± 0.21	12.25 ± 2.07	81.3 ± 0.89	67.82 ± 1.46
INLP	84.81 ± 0.47	12.69 ± 4.66	66.00 ± 1.32	100.17 ± 1.65
Noise	85.12 ± 0.47	12.49 ± 0.58	59.01 ± 0.65	103.93 ± 0.24
Adversarial	85.34 ± 0.22	7.83 ± 0.97	87.00 ± 2.22	46.61 ± 5.52
Adversarial + Multiple	84.92 ± 0.12	5.79 ± 1.44	84.38 ± 2.07	51.11 ± 4.06
FEDERATE	84.81 ± 0.34	2.68 ± 0.60	65.49 ± 3.48	98.53 ± 4.51

Table 2: Test results on CelebA dataset with fixed Relaxation Threshold of 1.0. Fairness is measured by TPR-Gap (lower is better), while privacy is measured by Leakage (lower is better) and MDL (higher is better). The MDL achieved by Random gives an upper bound for that particular dataset. The results have been averaged over 5 different seeds.

Method	Accuracy ↑	TPR-gap ↓	Leakage ↓	MDL ↑
Random	50.00 ± 0.00	0.00 ± 0.00	-	20.15 ± 0.083
Unconstrained	83.41 ± 0.32	12.73 ± 7.17	78.19 ± 1.0	16.38 ± 0.46
INLP	83.11 ± 0.51	3.91 ± 2.43	74.54 ± 0.67	19.93 ± 0.35
Noise	82.87 ± 0.37	8.01 ± 1.18	68.12 ± 0.94	19.38 ± 0.33
Adversarial	83.14 ± 0.53	7.02 ± 3.31	78.2 ± 0.18	16.1 ± 0.36
Adversarial + Multiple	83.14 ± 0.25	3.55 ± 2.16	81.37 ± 0.98	13.5 ± 1.09
FEDERATE	82.29 ± 0.9	2.73 ± 2.18	70.25 ± 4.81	18.1 ± 2.79

Table 3: Test results on Adult Income dataset with fixed Relaxation Threshold of 1.0. Fairness is measured by TPR-Gap (lower is better), while privacy is measured by Leakage (lower is better) and MDL (higher is better). The MDL achieved by Random gives an upper bound for that particular dataset. The results have been averaged over 5 different seeds.

Method	Accuracy ↑	TPR-gap ↓	Leakage ↓
Unconstrained	72.54 ± 0.57	27.17 ± 1.76	87.18 ± 0.32
Noise	71.87 ± 0.56	25.14 ± 3.47	71.75 ± 2.99
Adversarial	75.49 ± 0.71	8.47 ± 3.5	88.03 ± 0.24
Adversarial + Multiple	75.6 ± 0.53	7.74 ± 4.17	88.01 ± 0.28
FEDERATE	75.34 ± 0.56	5.46 ± 3.59	62.31 ± 5.69

Table 4: Test set results on Twitter Sentiment dataset (scores averaged over 5 different seeds, RT=0.0).

Method	Accuracy ↑	TPR-gap ↓	Leakage ↓
Unconstrained	70.57 ± 0.98	20.68 ± 0.99	82.91 ± 1.65
Noise	70.47 ± 0.43	19.84 ± 0.91	66.83 ± 3.32
Adversarial	74.09 ± 1.56	3.03 ± 2.65	88.14 ± 0.18
Adversarial + Multiple	74.44 ± 0.62	1.07 ± 0.74	87.98 ± 0.36
FEDERATE	74.24 ± 1.25	0.89 ± 0.46	61.92 ± 5.04

Table 5: Test set results on Twitter Sentiment dataset (scores averaged over 5 different seeds, RT=3.0).

Method	Accuracy ↑	TPR-gap ↓	Leakage ↓
Unconstrained	70.57 ± 0.98	20.68 ± 0.99	82.91 ± 1.65
Noise	70.47 ± 0.43	19.84 ± 0.91	66.83 ± 3.32
Adversarial	70.8 ± 2.77	1.72 ± 1.5	88.2 ± 0.24
Adversarial + Multiple	67.39 ± 1.16	1.0 ± 0.8	88.01 ± 0.12
FEDERATE	73.97 ± 1.6	1.4 ± 1.22	60.38 ± 5.46

Table 6: Test set results on Twitter Sentiment dataset (scores averaged over 5 different seeds, RT=10.0).

Method	ϵ	Twitter Sentiment		Bias in Bios	
		Accuracy ↑	Leakage ↓	Accuracy ↑	Leakage ↓
Noise	8.0	71.3	60.59	64.75	56
FEDERATE	8.0	74.89	56.91	64.78	54.4
Noise	10.0	71.63	65.57	70.86	57.7
FEDERATE	10.0	75.25	60.55	70.97	56.5
Noise	12.0	71.76	66.04	75.01	58.4
FEDERATE	12.0	75.31	53.31	75.01	57
Noise	14.0	71.7	67.98	76.74	59
FEDERATE	14.0	75.3	57.29	76.83	56.3
Noise	16.0	71.7	67.69	77.77	60.3
FEDERATE	16.0	75.56	61.98	77.89	57.9

Table 7: Accuracy-privacy trade-off for different noise level (as captured by ϵ).