Fair NLP Models with Differentially Private Text Encoders

Anonymous ACL submission

Abstract

Encoded text representations often capture sensitive attributes about individuals (e.g., gender, race or age), which can raise privacy concerns and contribute to making downstream models unfair to certain groups. In this work, we propose FEDERATE, an approach that combines ideas from differential privacy and adversarial learning 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 two challenging NLP tasks. Our results show that FEDERATE consistently improves upon previous methods.

1 Introduction

Algorithmically-driven decision-making systems have raised several fairness concerns (Raghavan et al., 2020; van den Broek et al., 2019) as they can be discriminatory against specific groups of people. On the other hand, these systems can 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 like GDPR 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. Furthermore, privacy and fairness can be at odds with one another. For instance, recent empirical and theoretical studies have shown that actively 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 these two notions and their interplay in the context of NLP, where fairness and privacy have often been considered independently from one another. Modern NLP heavily relies on learning or fine-tuning encoded representations of text, typically obtained as intermediate representations of a machine learning model. 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. Moreover, the presence of such information in the representations may lead to more unfair models downstream. For instance, even after scrubbing explicit gender indicators from text such as names and pronouns, De-Arteaga et al. (2019) found that occupation prediction models still show a large correlation between accuracy and gender, indicating the use of implicit gender information. Zhao et al. (2018) and Kiritchenko and Mohammad (2018) observed a similar phenomenon in coreference resolution and sentiment analysis. Privatizing encoded representations is thus an important, yet challenging problem for which existing approaches based on adversarial learning (Li et al., 2018; Coavoux et al., 2018; Han et al., 2021) or subspace projection (Bolukbasi et al., 2016; Wang et al., 2020; Karve et al., 2019; Ravfogel et al., 2020) do not provide a satisfactory solution. In particular, these methods lack any formal privacy guarantee, and it has been shown that an adversary can recover sensitive attributes from the resulting representations with high accuracy (Elazar and Goldberg, 2018).

In this work, we propose a novel approach (called FEDERATE) to learn private text representations by combining ideas from differential privacy (DP), a mathematical definition of privacy which comes with rigorous guarantees (Dwork and Roth, 2014), with an adversarial training mechanism. More specifically, we pro-

pose a flexible architecture in which (i) the output of an arbitrary text encoder is normalized and perturbed using random noise to make the resulting *private* 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. This architecture is trained end-to-end and can accommodate any type of text encoder while ensuring formal DP guarantees for the resulting text representations. This is in contrast to recent attempts at using DP in NLP (Lyu et al., 2020; Plant et al., 2021), for which we uncover a critical error in the privacy analysis.

We empirically evaluate the privacy-fairness-accuracy trade-off provided by `FEDERATE` on two NLP tasks: occupation prediction from bios (De-Arteaga et al., 2019) and sentiment analysis from tweets (Blodgett et al., 2016), where the sensitive attributes we consider are gender and race respectively. In contrast to previous studies which performed hyperparameter selection based only on validation accuracy, we propose a new criterion (applicable to all methods) which allows to obtain significant improvements in fairness or privacy for a small cost in accuracy. Our results show that `FEDERATE` simultaneously leads to more private representations and fairer models compared to state-of-the-art methods while maintaining comparable accuracy, and demonstrate that privacy and fairness are compatible in our setting and even mutually reinforce each other. Additionally, we find that `FEDERATE` provides better and smoother fairness-accuracy (resp. privacy-accuracy) trade-offs than purely adversarial (resp. purely noise-based) approaches on the large spectrum of possible trade-offs.

The paper is organized as follows. Section 2 provides some useful background on differential privacy. In Section 3, we present our approach. Section 4 reviews some related work. We describe our experimental results in Section 5, and conclude with final remarks in Section 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 of the attacker's capabilities and comes with a powerful algorithmic framework. For these reasons, it has become a de-facto standard in privacy and has been deployed in various settings, notably 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. We refer to 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 its data before sharing it. Our work aims to create an encoder that leads to a private representation of user text, which can then be shared with an untrusted curator for learning or inference. We thus consider LDP, defined as follows.

Definition 2.1 (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 user text and M will be an encoding function that the user applies to transform their text into a private representation before sharing it with untrusted parties. Among other desirable properties, DP is robust to post-processing: any function F applied over M is still ϵ-differential private.

Laplace mechanism. As clearly seen from Definition 2.1, 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 Laplace noise independently to each dimension in the input:

$$\mathbf{x}_{priv} \leftarrow \mathbf{x} + \boldsymbol{\ell}, \tag{2}$$

where each entry of the vector $\boldsymbol{\ell} \in \mathbb{R}^D$ is sampled independently from a centered Laplace distribution with scale $\frac{\Delta}{\epsilon}$, denoted by $\text{Lap}(\frac{\Delta}{\epsilon})$. The noise scale is calibrated to ϵ and the L1-sensitivity Δ of the inputs defined as:

$$\Delta = \max_{\mathbf{x},\mathbf{x}' \in X} \|\mathbf{x} - \mathbf{x}'\|_1. \tag{3}$$

In our work, we will apply the Laplace mechanism on top of a learned encoder to get private representations of input texts.

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 or race). 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 label y.

To achieve the above goals we introduce `FEDERATE` (for Fair modEls with DiffERentiAlly private Text Encoders), which combines ideas from DP and adversarial learning by integrating a randomized mapping into the encoder and modeling the adversary in the training phase to improve the fairness of the classifier.

Encoder architecture. We propose a generic encoder construction $E_{priv} = priv \circ E$ composed of two main components. The first component E can be any deterministic encoder which maps the user 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 $E(x)$:

$$priv(E(x)) = E(x)/\|E(x)\|_1 + \boldsymbol{\ell}, \tag{4}$$

where each entry of $\boldsymbol{\ell} \in \mathbb{R}^D$ is sampled independently from $\text{Lap}(\frac{2}{\epsilon})$. As the L1 sensitivity of the *normalized* representation is bounded by 2 for any E, $E_{priv} = priv \circ E$ is ϵ-DP.

Training phase. The objective of the training phase is to learn the parameters of the encoder E_{priv} and the classifier C from a set of tuples (x, y, z). During training, we model the adversary by a classifier A which aims to predict z, while the encoder E_{priv} is optimized to fool A while maximizing the accuracy of the downstream classifier C. Given $\lambda \geq 0$, we train E_{priv}, C and A (parameterized by θ_E, θ_C, and θ_A respectively) to optimize the objective:

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

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. We solve the problem with backpropagation using a gradient reversal layer (Ganin and Lempitsky, 2015), which acts like an identity function in the forward pass and 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.

Inference phase. Once trained, E_{priv} can be used to privately encode new data points which can then be fed into the classifier C for inference. Note that by the post-processing property of DP, applying C or any other function on top to E_{priv} preserves the ϵ-DP guarantee of E_{priv}. In our experiments, we will empirically evaluate the privacy of $E_{priv}(\cdot)$ and the fairness of $C(E_{priv}(\cdot))$ and show that our approach consistently provides better privacy-fairness-accuracy trade-offs than previous methods.

4 Related Work

This section reviews related lines of work, highlighting the main differences with our approach.

Adversarial learning. In order to improve the fairness of a model or to prevent its intermediate representations from leaking sensitive attributes of the input, several approaches in NLP also employ an adversarial-based training mechanism. 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) improve upon these approaches by introducing multiple adversaries focusing on different aspects of the representation. Their loss function encourages orthogonality between pairs of adversaries and leads to some improvements in the fairness of downstream models at the cost of higher training complexity. Unlike our approach, these methods do not offer formal privacy guarantees. Elazar and Goldberg (2018) have shown that it is often possible to recover the sensitive attributes from the representations by training a post-hoc classifier. This is also what we observe in our experiments (see 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. A key advantage over adversarial learning is that these methods do not use a task-specific loss. They instead rely on a manual selection of words in the vocabulary to estimate the sensitive sub-space, making them difficult to generalize to new attributes. Furthermore, Gonen and Goldberg (2019) found bias to be deeply ingrained in these representations, and showed that sensitive attributes remain present even after applying these approaches.

In order to circumvent these issues, 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. Although they show significant improvements in privacy and fairness over other projection approaches, the method is designed to remove linear information from the representation. As a result, a nonlinear adversary can still retrieve a significant amount of sensitive information. By leveraging DP, our approach provides robust guarantees that do not depend on the expressiveness of the adversary, thereby providing effective 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 the individual points used to train it. Empirically, this can be evaluated through membership inference attacks, where an attacker with access to the model seeks to determine whether a given data point was part of the training set (Shokri et al., 2017). While Kulynych et al. (2022) observed that DP helps to reduce disparate vulnerability to such attacks, several empirical studies have shown that DP can disproportionately impact the accuracy of the model for some groups and thus 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 members of 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) considered the problem of training a fair model under DP constraints only for the sensitive attribute.

A fundamental difference between the above line of work and our approach lies in the kind of privacy we provide. While the above approaches study DP as a way to design privacy-preserving learning algorithms which protect training points from membership inference attacks on the model, our goal is to construct a private encoder such that the encoded representations of two different points do not differ too much. 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 our setting privacy and fairness are compatible and even tend to 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 fair models by using an adversarial training mechanism. Our experiments show that our approach leads to more private representations and fairer models in practice. We also found a critical error in their privacy analysis, where they incorrectly bound the sensitivity of their representation by 1 while it can in fact be as large as D (the dimension of the representation, which is typically in the hundreds). As a result, the privacy guarantees are significantly weaker than the authors claim: the ϵ values they report should be multiplied by D. We provide more details in Appendix B.

Concurrent to and independently from our work, Plant et al. (2021) proposed an adversarial-driven DP training mechanism. However, they do not consider fairness, whereas our focus is on the combination of fairness and privacy. Moreover, their method reproduces the same incorrect analysis as Lyu et al. (2020) and provide similarly inflated privacy claims.

5 Experiments

In this section, we present experiments aiming to (i) assess the privacy-fairness-accuracy trade-offs of different approaches and (ii) analyze privacy-accuracy and fairness-accuracy tradeoffs separately. We begin by describing the setup, datasets, and metrics.

Datasets. We consider two 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), 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.

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.

Fairness metrics. For Twitter Sentiment we report 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. We also report False Positive Rate Gap (FPR-gap), which, coupled with TPR-gap, corresponds to equalized odds (Hardt et al., 2016). Formally, for a dataset with binary label $y \in \{0, 1\}$ and sensitive attribute $z \in \{g, \neg g\}$, TPR-gap is defined as:

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

where $\hat{y}$ is the predicted class. FPR-gap is defined similarly (see Appendix C.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 metric. To measure the privacy of a text encoder, we use the accuracy of a two-layer adversarial network which predicts the sensitive attribute from the representation (Leakage). This classifier is trained on the validation set and evaluated on the test set.

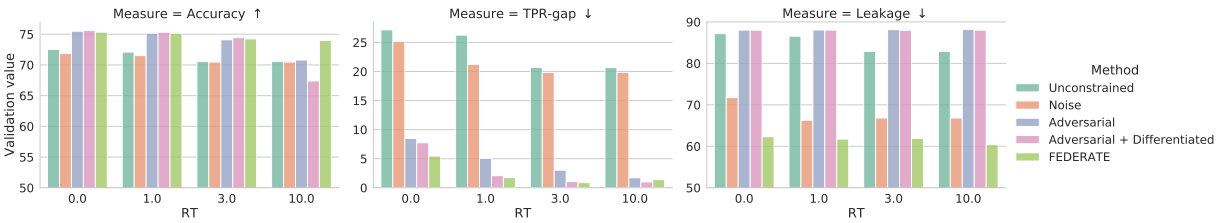

Figure 1: Fairness, accuracy, and privacy of various approaches for different RT on the validation set of Twitter Sentiment. For increasing RT, fairness improves for all approaches with little change in accuracy.

Method	Accuracy ↑	TPR-gap ↓	TNR-gap ↓	Leakage ↓
`Random`	50.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	-
`Unconstrained`	72.09 ± 0.73	26.26 ± 0.87	44.98 ± 1.82	86.56 ± 0.83
`INLP`	67.62 ± 0.57	9.19 ± 1.08	35.14 ± 1.26	80.27 ± 2.5
`Noise`	71.52 ± 0.51	21.23 ± 2.5	41.97 ± 2.02	66.29 ± 3.55
`Adversarial`	75.16 ± 0.65	5.03 ± 2.94	22.1 ± 4.23	88.06 ± 0.2
`Adversarial + Differentiated`	75.32 ± 0.6	2.09 ± 1.18	18.58 ± 1.25	88.03 ± 0.47
`FEDERATE`	75.15 ± 0.59	1.75 ± 1.41	16.48 ± 0.38	61.74 ± 5.05

Table 1: Test set results on Twitter Sentiment dataset (scores averaged over 5 different seeds, RT=1.0).

Methods and model architectures. We compare **`FEDERATE`** to the following competing methods: (i) **`Adversarial`** implements standard adversarial learning, which is equivalent to our approach without the *priv* layer, (ii) **`Adversarial + Differentiated`** (Han et al., 2021) implements multiple adversaries,[1] (iii) **`INLP`** (Ravfogel et al., 2020) is a subspace projection approach, and (iv) **`Noise`** learns DP text representations as proposed by Lyu et al. (2020) but with corrected privacy analysis: this corresponds to our approach without the adversarial component. 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 two datasets. In case of Twitter Sentiment dataset, 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 C.2-C.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. 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.[2]

[1] We do not evaluate **`Adversarial + Differentiated`** on Bias in Bios as it is expensive to train due to the multiple adversaries.

[2] We can also incorporate privacy into our hyperparameter selection strategy but, for the datasets and

Method	Accuracy ↑	GRMS ↓	Leakage ↓
`Random`	3.53 ± 0.01	0.00 ± 0.00	-
`Unconstrained`	79.29 ± 0.32	15.88 ± 0.80	75.92 ± 2.73
`INLP`	75.96 ± 0.47	12.81 ± 0.09	59.91 ± 0.08
`Adversarial`	79.02 ± 0.20	13.06 ± 0.39	69.47 ± 1.64
`Noise`	77.88 ± 0.32	13.89 ± 0.31	62.23 ± 0.99
`FEDERATE`	77.79 ± 0.11	11.02 ± 0.55	56.92 ± 0.98

Table 2: Test set results on Bias in Bios dataset (scores averaged over 5 different seeds, RT= 1.0).

Figure 1 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 (found to provide good trade-offs for all approaches in Figure 1), see Table 1 for Twitter Sentiment and Table 2 for Bias in Bios (see also Appendix C.4 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 over the Twitter Sentiment. Interestingly, these accuracies are higher than that of `Unconstrained` and `Noise`. We attribute this to a significant mismatch in the train and test distribution due to skewing. Over 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 it to be 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 detail in Section 5.2.

In terms of privacy, `FEDERATE` significantly outperforms all other adversarial approaches on both datasets. In fact, the leakage of purely adversarial approaches similar to that of `Unconstrained`, which is in line with previous studies (Han et al., 2021). Over Bias in Bios, `INLP` provides a similar level of privacy to `FEDERATE`, albeit with a worse accuracy. However, in the case of Twitter Sentiment, our proposed approach leaks significantly less information while also having higher accuracy. Finally, on both datasets, `Noise` achieves slightly weaker privacy than `FEDERATE` with much worse accuracy and fairness.

Overall, the results suggest that, although there are methods which can provide either privacy or fairness, `FEDERATE` stands out as the only approach that can simultaneously induce a fairer model *and* make its representation private. Furthermore, these results empirically demonstrate that our notions 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 considered accuracy, privacy, and fairness at the same time and found our approach to attain better trade-offs than all other approaches. Here, we take a closer look at the pairwise fairness-accuracy and privacy-accuracy trade-offs separately and show that `FEDERATE` outperforms the purely `Adversarial` or `Noise` approach in the corresponding dimension. This section can be seen as an ablation study which validates the superiority of combining adversarial learning and

methods in our study, we found no significant change in Leakage across the considered RT values, see Figure 1.

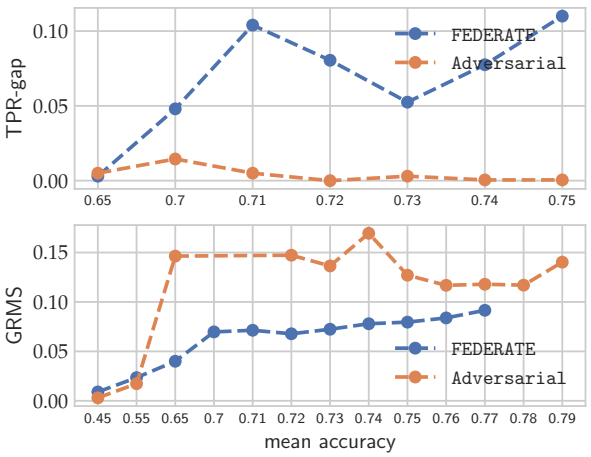

Figure 2: Fairness-accuracy trade-off on Twitter Sentiment (top) and Bias in Bios (bottom). A missing point means that the accuracy interval was not found within our hyperparameter search.

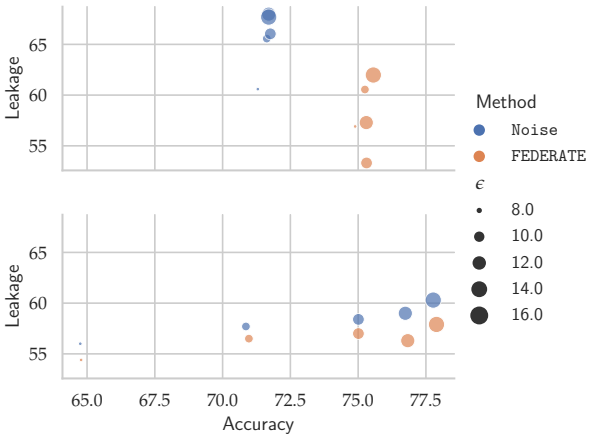

Figure 3: Privacy-accuracy trade-off on Twitter Sentiment (top) and Bias in Bios (bottom), with associated values of ϵ.

noise addition 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 2. The interval is denoted by mean accuracy, for instance, accuracy interval between 71.5 and 72.5 is represented with 72 and then we find the corresponding best validation fairness score. We find that our proposed approach provides better fairness than the `Adversarial` approach for almost all the 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. Interestingly, we find that `FEDERATE` enables a smoother exploration of the accuracy-fairness trade-off space than `Adversarial`, which shows much more erratic trajectories. Adversarial models are notoriously difficult to train, and 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 3. 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 C.4.

6 Conclusion and Perspectives

This work proposed a DP-driven adversarial learning mechanism for NLP. Through our experiments, we showed that our approach can simultaneously induce private representations and fair models, with a mutually reinforcing effect between privacy and fairness. We also find that our method improves upon competitors on each dimension separately. While we focused on privatizing certain sensitive attributes like race or gender, our approach could easily 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).

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.

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*.

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.

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.

Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323.

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.

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.

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

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.

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.

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.

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 Section 3.

B 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. 3). 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:

$$\mathbf{x} \leftarrow \mathbf{x} - \min(\mathbf{x})/(\max(\mathbf{x}) - \min(\mathbf{x})), \quad (6)$$

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 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. 4 guarantees by design that the L1 sensitivity is bounded by 2.

C Experiments

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

C.1 Fairness Measure

FPR-gap: Formally, for a classifier C, with binary labels $y \in \{0, 1\}$ and protected attribute $z \in \{g, \neg g\}$, FPR-gap is defined as:

$$\text{FPR-gap} = P_g(\hat{Y} = 0|Y = 0) - P_{\neg g}(\hat{Y} = 0|Y = 0) \quad (7)$$

where $\hat{Y}$ is the predicted class.

C.2 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.

C.3 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. 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.,

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 $L(y^i, \hat{y}^i) - \lambda \cdot L(z^i, \hat{z}^i)$

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 + Differentiated`: Similar to `Adversarial`, we vary λ between 0.1 to 3.0 with an interval of 0.2. Apart from λ, `Adversarial + Differentiated` 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 + Differentiated`, 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.

C.4 Additional Results

Tables 3–5 present detailed results on Twitter Sentiment with different relaxation thresholds, which were summarized in Figure 1.

Table 6 provides the detailed privacy-fairness results which were summarized in Figure 3.

Method	Accuracy ↑	TPR-gap ↓	TNR-gap ↓	Leakage ↓
`Unconstrained`	72.54 ± 0.57	27.17 ± 1.76	46.32 ± 1.01	87.18 ± 0.32
`Noise`	71.87 ± 0.56	25.14 ± 3.47	43.99 ± 1.55	71.75 ± 2.99
`Adversarial`	75.49 ± 0.71	8.47 ± 3.5	25.43 ± 4.27	88.03 ± 0.24
`Adversarial + Differentiated`	75.6 ± 0.53	7.74 ± 4.17	25.09 ± 4.19	88.01 ± 0.28
`FEDERATE`	75.34 ± 0.56	5.46 ± 3.59	20.44 ± 5.2	62.31 ± 5.69

Table 3: Test set results on Twitter Sentiment dataset (scores averaged over 5 different seeds, RT=0.0).

Method	Accuracy ↑	TPR-gap ↓	TNR-gap ↓	Leakage ↓
`Unconstrained`	70.57 ± 0.98	20.68 ± 0.99	42.4 ± 2.2	82.91 ± 1.65
`Noise`	70.47 ± 0.43	19.84 ± 0.91	44.25 ± 2.38	66.83 ± 3.32
`Adversarial`	74.09 ± 1.56	3.03 ± 2.65	18.69 ± 4.56	88.14 ± 0.18
`Adversarial + Differentiated`	74.44 ± 0.62	1.07 ± 0.74	16.43 ± 2.1	87.98 ± 0.36
`FEDERATE`	74.24 ± 1.25	0.89 ± 0.46	16.69 ± 0.98	61.92 ± 5.04

Table 4: Test set results on Twitter Sentiment dataset (scores averaged over 5 different seeds, RT=3.0).

Method	Accuracy ↑	TPR-gap ↓	TNR-gap ↓	Leakage ↓
`Unconstrained`	70.57 ± 0.98	20.68 ± 0.99	42.4 ± 2.2	82.91 ± 1.65
`Noise`	70.47 ± 0.43	19.84 ± 0.91	44.25 ± 2.38	66.83 ± 3.32
`Adversarial`	70.8 ± 2.77	1.72 ± 1.5	11.6 ± 4.86	88.2 ± 0.24
`Adversarial + Differentiated`	67.39 ± 1.16	1.0 ± 0.8	8.6 ± 3.47	88.01 ± 0.12
`FEDERATE`	73.97 ± 1.6	1.4 ± 1.22	14.69 ± 2.33	60.38 ± 5.46

Table 5: 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 6: Accuracy-privacy trade-off for different noise level (as captured by ϵ).