The Impact of Differential Privacy on Group Disparity Mitigation

Anonymous ACL submission

Abstract

The performance cost of differential privacy has, for some applications, been shown to be higher for minority groups; fairness, conversely, has been shown to disproportionally compromise the privacy of members of such groups. Most work in this area has been restricted to computer vision and risk assessment. In this paper, we evaluate the impact of differential privacy on fairness across four tasks, focusing on how attempts to mitigate privacy violations and between-group performance differences interact: Does privacy inhibit attempts to ensure fairness? To this end, we train (ε, δ)-differentially private models with empirical risk minimization and group distributionally robust training objectives. Consistent with previous findings, we find that differential privacy increases between-group performance differences in the baseline setting; but more interestingly, differential privacy *reduces* between-group performance differences in the robust setting. We explain this by reinterpreting differential privacy as regularization.

1 Introduction

Classification tasks in computer vision and natural language processing face the challenge of balancing performance with the need to prevent discrimination against protected demographic subgroups, satisfying fairness principles. In some tasks, we train our classifiers on private data and therefore also need our models to satisfy privacy guarantees.

Privacy-preserving algorithms, however, tend to disproportionally affect members of minority classes (Farrand et al., 2020). E.g., Bagdasaryan, Poursaeed, and Shmatikov (2019), show the performance cost of differential privacy (Dwork et al., 2006) in face recognition is higher for minority groups, suggesting that privacy and fairness are fundamentally at odds (Chang and Shokri, 2021; Agarwal, 2021).

In this paper, we evaluate two hypotheses at scale: (a) that the performance cost of differential privacy is unevenly distributed across demographic groups (Ekstrand, Joshaghani, and Mehrpouyan, 2018; Cummings et al., 2019; Bagdasaryan, Poursaeed, and Shmatikov, 2019; Farrand et al., 2020), and (b) that such effects can be mitigated by more robust learning objectives (Sagawa et al., 2020a; Pezeshki et al., 2020).

Contributions We build upon previous work suggesting that differential privacy and fairness are at odds: Differential privacy hurts minority groups the most, and reducing the fairness gap by focusing on minority groups during training typically puts their privacy at risk. We evaluate this hypothesis at scale by measuring the impact of differential privacy in terms of fairness across (1) a baseline empirical risk minimization and (2) under a group distributionally robust optimization. We conduct our experiments across four tasks of different modalities, assuming the group membership information is available at training time, but not at test time: face recognition (CelebA), topic classification, volatility forecasting based on earning calls, and sentiment analysis of product reviews. Our results confirm that differential privacy compromises fairness in the baseline setting; however, we demonstrate that differential privacy not only mitigates the decrease but also *improves* fairness compared to non-private experiments for 4/5 tasks in the distributionally robust setting. We explain this by reinterpreting differential privacy as an approximation of Gaussian noise injection, which is equivalent to strategies previously shown to determine the efficacy of group-robust learning.

2 Fairness and Privacy

Fair machine learning aims to ensure that induced models do not discriminate against individuals with specific values in their protected attributes (e.g.,

Findings of the Association for Computational Linguistics: NAACL 2024, pages 3952–3965

race, gender). We represent each data point as $z = (x, g, y) \in \mathcal{X} \times \mathcal{G} \times \mathcal{Y}$, with $g \in \mathcal{G}$ encoding its protected attribute(s).[1] Let $\mathcal{D}_y^g$ denote the distribution of data with protected attribute g and label y.

Several definitions of group fairness exist in the literature (Williamson and Menon, 2019), but here we focus on a generalization of approximately constant conditional (equalized) risk (Donini et al., 2018):[2]

Definition 2.1 (Δ-Fairness)**.** Let $\ell^{g_i}(\theta) = \mathbb{E}[\ell(\theta(x), y)|g = g_i]$ be the risk of the samples in the group defined by g_i, and $\Delta \in [0, 1]$. We say that a model θ is Δ-fair if for any two values of g, say g_i and g_j, $|\ell^{g_i}(\theta) - \ell^{g_j}(\theta)| < \Delta$.

Note that if ℓ coincides with the performance metric of a task, and $\delta = 0$, this is identical to performance or classification parity (Yuan et al., 2021).[3] Such a notion of fairness can be derived from John Rawls' theory on distributive justice and stability, treating model performance as a resource to be allocated. Rawls' *difference principle*, maximizing the welfare of the worst-off group, is argued to lead to stability and mobility in society at large (Rawls, 1971). Δ directly measures what is sometimes called Rawlsian *min-max fairness* (Bertsimas, Farias, and Trichakis, 2011). In our experiments, we measure Δ-fairness as the absolute difference between performance of the worst-off and best-off subgroups.

Recall the standard definition of (ε, δ)-privacy:

Definition 2.2. θ is (ε, δ)-private iff $\Pr[\theta(\mathcal{X})] \leq \exp(\varepsilon) \times \Pr[\theta(\mathcal{X}')] + \delta$ for any two distributions, $\mathcal{X}$ and $\mathcal{X}'$, different at most in one row.

Differential privacy thereby ensures that an algorithm will generate similar outputs on similar data sets. Note the multiplicative bound $\exp(\varepsilon)$ and the additive bound δ serve different roles: The δ term represents the possibility that a few data points are not governed by the multiplicative bound, which controls the level of privacy (rather than its scope). Note that it also follows directly that if $\varepsilon = 0$ and $\delta = 0$, absolute privacy is required, leading θ to be independent of the data.

Several authors have shown that differential privacy comes at different costs for minority subgroups (Ekstrand, Joshaghani, and Mehrpouyan, 2018; Cummings et al., 2019; Bagdasaryan, Poursaeed, and Shmatikov, 2019; Farrand et al., 2020). The more private the model is required to be, the larger group disparities it will exhibit.[4] This happens because differential privacy distributes noise where it is needed to reduce the influence of individual examples. Since outlier examples are likely to have disproportional influence on output distributions (Campbell, 1978; Chernick and Murthy, 1983), they are also disproportionally affected by noise injection in differential privacy.

Agarwal (2021) show that, in fact, a $(\varepsilon, 0)$-private and fully fair model – using equalized odds as the definition of fairness – will be unable to learn anything. To see this, remember that a fully private model is independent of the data and unable to learn from correlations between input and output. If θ is, in addition, required to be fair, it is thereby required to be fair for all distributions, which prevents θ from encoding any prior beliefs about the output distribution. Note this finding generalizes straight-forwardly to equalized risk, and even to approximate fairness (since even for finite distributions, we can define a $\Delta > 0$, such that preserving absolute privacy would lead to a constant θ).

Theorem 1. *For sufficiently small values of Δ, a fully $(\varepsilon, 0)$-private model θ that is also Δ-fair, will have trivial performance.*

Proof. This follows directly from the above. □

While we do not strictly require an absolute privacy in our experiments (setting $\delta = 10^{-5}$), intuitively, privacy compromises fairness by adding more noise to data points of minority group members than to those of majority groups. Fairness, on the other hand, leads to over-sampling or over-attending to data points of minority group members, more likely compromising their privacy.

Pannekoek and Spigler (2021) show, however, that it *is* possible to learn *somewhat private* and

[1] In practice our protected attributes in §3 will be *age* and *gender*. Both are protected under the Equality Act 2010.

[2] In the fairness literature, approximate fairness is referred to as δ-fairness, but below we will use lower case δ to refer to (ε, δ)-differential privacy, and we refer to Δ-fairness to avoid confusion.

[3] Performance or classification parity has been argued to suffer from statistical limitations in (Corbett-Davies and Goel, 2018), which remind us that when risk distributions differ, standard error metrics are poor proxies of individual equity. This is known as the problem of infra-marginality. Note, however, that this argument does not apply to binary classification problems.

[4] Note this is a different trade-off than the fairness-privacy trade-off which results from the need for collecting sensitive data to learn fair models; the latter is discussed at length in Veale and Binns (2017).

somewhat fair classifiers. They combine differential privacy with reject option classification. Their results nevertheless confirm that privacy and fairness objectives are fundamentally at odds, as fairness decreases with the introduction of differential privacy.

3 Experiments

This section describes the algorithms and datasets involved in our experiments, and presents the results of these.

3.1 Algorithms

Empirical Risk Minimization For a model parameterized by θ, in our baseline Empirical Risk Minimization (ERM) setting, we minimize the expected loss $\mathbb{E}[\ell(\theta(x), y)]$ with data $(x, g, y) \in \mathcal{X} \times \mathcal{G} \times \mathcal{Y}$ drawn from a dataset $\mathcal{D}$:

$$\hat{\theta}_{ERM} = \underset{\theta}{\operatorname{argmin}}\, \mathbb{E}_{\hat{\mathcal{D}}}[\ell(\theta(x), y)] \qquad (1)$$

Here $\hat{\mathcal{D}}$ denotes the empirical training distribution. Note that we disregard any group information in our data. In an overparameterized setting, ERM is prone to overfitting spurious correlations, which are more likely to hurt performance on minority groups (Sagawa et al., 2020b).

Distributionally Robust Optimization Several authors have suggested to mitigate the effects of such overfitting by explicitly optimizing for out-of-distribution mixtures of sub-populations (Hu et al., 2018; Oren et al., 2019; Sagawa et al., 2020a). In this work we focus on Group-aware Distributionally Robust Optimization (Group DRO) (Sagawa et al., 2020a).

Under the assumption that the training distribution $\mathcal{D}$ is a mixture of a discrete number of groups, $\mathcal{D}_g$ for $g \in \mathcal{G}$, we define the worst-case loss as the maximum of the group-specific expected losses:

$$\ell(\theta)_{worst} = \max_{g \in \mathcal{G}} \mathbb{E}_{\hat{\mathcal{D}}_g}[\ell(\theta(x), y)] \qquad (2)$$

In Group DRO – in contrast with ERM – we exploit our knowledge of the group membership of data points (x, g, y). The overall objective is for minimizing the empirical worst-case loss is therefore:

$$\hat{\theta}_{DRO} = \underset{\theta}{\operatorname{argmin}} \left[\hat{\ell(\theta)}_{worst} := \max_{g \in G} \mathbb{E}_{\hat{\mathcal{D}}_g}[\ell(\theta(x), y)] \right] \qquad (3)$$

Note, again, that the knowledge of group membership g is only available at training time, not at test time. Unlike Sagawa et al. (2020a), we do not employ heavy ℓ_2 regularization during our experiments, but rather use it with the same parameters as proposed in Koh et al. (2021).

Differentially Private Stochastic Gradient Descent (DP-SGD) We implement differential privacy (Dwork et al., 2006) using DP-SGD, as presented in Abadi et al. (2016). DP-SGD limits the influence of training samples by (i) clipping the per-batch gradient where its norm exceeds a predetermined clipping bound C, and by (ii) adding Gaussian noise $\mathcal{N}$ characterized by a noise scale σ to the aggregated per-sample gradients. We control this influence with a privacy budget ε, where lower values for ε indicates a more strict level of privacy. DP-SGD has remained popular, among other things because it generalizes to iterative training procedures (McMahan et al., 2018), and supports tighter bounds using the Rényi method (Mironov, 2017).

Differential privacy generally comes at a performance cost, leading to privacy-preserving models performing worse compared to their non-private counterparts (Alvim et al., 2011). However, we follow Kerrigan, Slack, and Tuyls (2020) and *finetune* the private models, which are first pretrained (without differential privacy) on a large public dataset. This protocol generally seems to provide a better trade-off between accuracy and privacy (Kerrigan, Slack, and Tuyls, 2020), leading to better-performing, yet private models. The only exception to this setup is the volatility forecasting task, where our models were trained from scratch, as those rely on PRAAT audio features.

3.2 Tasks and architectures

To study the impact of differential privacy on fairness, in ERM and Group DRO, we evaluate increasing levels of differential privacy across five datasets that span four tasks and three different modalities: speech, text and vision.

Facial Attribute Detection We study facial attribute recognition with the CelebFaces Attributes Dataset (CelebA) (Liu et al., 2015). It contains faces of celebrities annotated with attributes, such as hair color, gender and other facial features. Following Sagawa et al. (2020a), we use the hair color as our target variable, with gender being the demographic attribute (see Figure 1 (left)). The dataset contains $\sim 163K$ datapoints, where the smallest

Figure 1: Examples of the different subgroups that appear in a subset of the datasets we train on. CelebA (left) contains images of celebrities, using hair-color as our target variable and gender as our protected attribute. Blog Authorship Corpus (right) contains text-based blogposts on two topics {Technology, Arts} our targets, using $\mathcal{G}$: {Man, Woman} $\times$ {Young, Old} as our protected subgroups.

group (blond males) only counts 1387. We finetune a publicly pretrained ResNet50, a standard model for image classification tasks, on the CelebA dataset and evaluate model performances as accuracies over 3 individual seeds.

Topic Classification For topic classification, we use the Blog Authorship Corpus (Schler et al., 2006). The Blog Authorship Corpus contains weblogs written on 19 different topics, collected from the Internet before August 2004. The dataset contains self-reported demographic information about the gender and age of the authors. Gender information is binary, and we binarize age, distinguishing between young ($=< 35$) and older (> 35) authors,[5] resulting in four different group combinations (see Figure 1 (right)). We chose two topics of roughly equal size (Technology and Arts), reducing the topic classification task to a binary classification task. For our experiments, we finetune a pretrained English DistilBERT model (Sanh et al., 2019). To reduce the overall added computational cost of DP-SGD, we freeze our model, except for the outer-most Transformer encoder layer as well as the classification layer. We report model performances as F1 scores over 3 individual seeds.

Volatility Forecasting For the stock volatility forecasting task, we use the Earnings Conference Calls dataset by Qin and Yang (2019). This consists of 559 public earnings calls audio recordings for 277 companies in the S&P 500 index, spanning over a year of earnings calls. We obtain the self-reported gender of the CEOs from Reuters,[6] Crunchbase,[7] and the WikiData API.[8] Gender information is binary, with 12.3% of speakers being female and 87.7% of speakers being male, a highly skewed distribution. Since our primary focus with this task is to explore the impact of differential privacy on speech, we use only audio features without the call transcripts. For each audio recording A of a given earning call E, the goal is to predict the company's stock volatility as a regression task. Following Qin and Yang (2019), we calculate the average log volatility τ days (temporal window) following the day of the earnings call. For each audio clip belonging to a given call, we extract 26-dimensional features with PRAAT (Boersma and Van Heuven, 2001). Each audio embedding of the call is fed sequentially to a BiLSTM, followed by an attention layer and two fully-connected layers. The model is trained by optimizing the Mean Square Error (MSE) between the predicted and true stock volatility. For all results, we report MSE on the test set for a 70:10:20 temporal split of the data.The results are averaged over 5 seeds.

Sentiment Analysis For our sentiment analysis task, we use the Trustpilot Corpus (Hovy, Johannsen, and Søgaard, 2015)[9]. It consists of text-based user reviews from the Trustpilot website, rating companies and services on a 1 to 5 star scale. The reviews spans 5 different countries; Germany, Denmark, France, United Kingdom and USA, however, we only consider the English reviews, i.e. UK and US. The Trustpilot contains demographic information about the gender, age and geographic

[5] Older authors tend to be underrepresented in web data

[6] https://www.thomsonreuters.com/en/profiles.html

[7] https://www.crunchbase.com/discover/people

[8] https://query.wikidata.org/

[9] https://bitbucket.org/lowlands/release/src/master/WWW2015/data/

		Performance at ε-Privacy							
		No DP		ε_1		ε_2		ε_3	
		Score	ε	Score	ε	Score	ε	Score	ε
CELEB	ERM	0.954 ± 0.000	-	0.943 ± 0.001	9.50	0.940 ± 0.002	5.17	0.932 ± 0.001	0.99
	DRO	0.953 ± 0.001	-	0.899 ± 0.006	9.50	0.891 ± 0.014	5.17	0.873 ± 0.007	0.99
BLOG	ERM	0.699 ± 0.002	-	0.661 ± 0.003	9.25	0.661 ± 0.003	5.03	0.648 ± 0.005	1.02
	DRO	0.692 ± 0.001	-	0.651 ± 0.001	9.25	0.650 ± 0.005	5.03	0.630 ± 0.003	1.02
VOL.	ERM	0.756 ± 0.036	-	0.778 ± 0.073	9.32	0.794 ± 0.046	6.42	0.778 ± 0.039	0.96
	DRO	0.814 ± 0.061	-	0.798 ± 0.042	9.32	0.815 ± 0.056	6.42	0.833 ± 0.093	0.96
T-UK	ERM	0.933 ± 0.008	-	0.919 ± 0.002	9.39	0.916 ± 0.001	4.94	0.889 ± 0.009	1.02
	DRO	0.931 ± 0.004	-	0.893 ± 0.006	9.39	0.873 ± 0.015	4.94	0.820 ± 0.015	1.02
T-US	ERM	0.894 ± 0.007	-	0.817 ± 0.014	10.71	0.812 ± 0.009	5.10	0.666 ± 0.019	1.01
	DRO	0.899 ± 0.009	-	0.569 ± 0.132	10.71	0.437 ± 0.112	5.10	0.342 ± 0.012	1.01

		Group-disparity at ε-Privacy							
		No DP		ε_1		ε_2		ε_3	
		GD	ε	GD	ε	GD	ε	GD	ε
CELEB	ERM	0.556 ± 0.021	-	0.746 ± 0.032	9.50	0.734 ± 0.025	5.17	0.770 ± 0.013	0.99
	DRO	0.514 ± 0.042	-	0.039 ± 0.018	9.50	0.080 ± 0.031	5.17	$\mathbf{0.056} \pm 0.027$	0.99
BLOG	ERM	0.108 ± 0.013	-	0.149 ± 0.006	9.25	0.140 ± 0.004	5.17	0.136 ± 0.011	0.99
	DRO	0.078 ± 0.009	-	0.056 ± 0.020	9.25	0.070 ± 0.013	5.17	$\mathbf{0.077} \pm 0.027$	0.99
VOL.	ERM	0.302 ± 0.042	-	0.328 ± 0.067	9.32	0.557 ± 0.050	6.42	0.573 ± 0.050	0.96
	DRO	0.221 ± 0.062	-	0.320 ± 0.085	9.32	0.371 ± 0.058	6.42	0.421 ± 0.083	0.96
T-UK.	ERM	0.018 ± 0.005	-	0.022 ± 0.006	9.39	0.020 ± 0.014	4.94	0.037 ± 0.006	1.02
	DRO	0.030 ± 0.008	-	0.030 ± 0.004	9.39	0.039 ± 0.023	4.94	$\mathbf{0.025} \pm 0.010$	1.02
T-US	ERM	0.055 ± 0.006	-	0.048 ± 0.019	10.71	0.054 ± 0.015	5.10	0.109 ± 0.017	1.01
	DRO	0.036 ± 0.007	-	0.118 ± 0.040	10.71	0.078 ± 0.030	5.10	$\mathbf{0.021} \pm 0.030$	1.01

Table 1: Performance (top) and Δ-Fairness (bottom) of ERM and Group DRO across different degrees of differential privacy (ε). ε_1, ε_2 and ε_3 corresponds to ε-values of roughly 10, 5 and 1 respectively (see table for exact values). We report F1 scores for sentiment and topic classification, accuracy for face recognition and MSE for volatility forecasting. Group disparity (GD) is measured by the absolute difference between the best and worst performing sub-group (Δ-Fairness; see Definition 2.1). The performance and corresponding uncertainties are based on several individual runs of each configuration, see §6.2 in the Appendix for further details. Differential privacy consistently hurts fairness for ERM. For Group DRO, we **bold-face** numbers where strict differential privacy (ε_3) *increases* fairness; this happens in 4/5 datasets. We see large increases for face recognition and small increases for topic classification and sentiment analysis.

location of the users, but as with the topic classification task, we only concern ourselves with the gender and age of the users. As with the topic classification task, we finetune DistilBERT on the UK and US English parts of the Trustpilot Corpus, freezing all parameters but the final encoder layer, as well as the classification layer. Classification performance is measured as F1 scores and the results are averaged over 3 seeds.

Our implementation is a PyTorch extension of the WILDS repository[10] (Koh et al., 2021) using the DP-SGD implementation provided by the Opacus Differential Privacy framework[11]. For further details about data and training, see §6.2 in the Appendix. We release the code for our experiments at: https://github.com/anonymized.

3.3 Results

Our results are presented in Table 1. The top half of the table presents standard (average) performance numbers across multiple runs of ERM and Group DRO at different privacy levels. Recall that performance for sentiment analysis as well as topic classification is measured in F1, volatility forecasting is measured in MSE and face recognition is measured in accuracy. The accuracy of our ERM face attribute detection classifier is 0.954 in the non-private setting, for example.

Our first observation is that, as hypothesized earlier, differential privacy hurts model performance. For our smallest text-based dataset (T-US), per-

[10] https://github.com/p-lambda/wilds/
[11] https://opacus.ai/

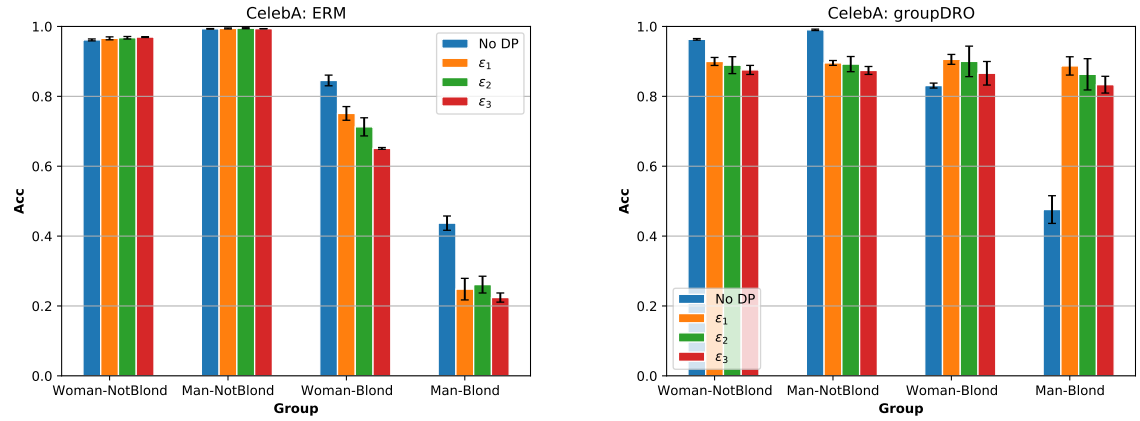

Figure 2: **Face Attribute Detection:** Performance of individual groups of increasing levels of ε. Comparing baseline ERM to Group DRO, we find that Group DRO performance on the minority group (blond males) perform much better under privacy constraints; we return to this in §3.4.

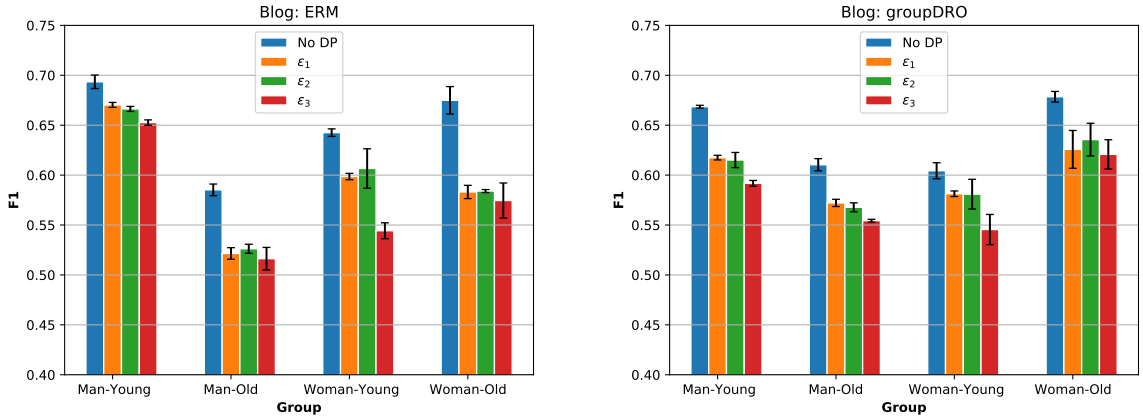

Figure 3: **Topic Classification:** Performance of individual groups of increasing levels of ε. Group DRO, compared to baseline ERM, results in a more balanced performance across all groups, even on a low privacy budget.

formance becomes very poor at the strictest privacy level. This is however associated with a high amount of variance between seeds, see Figure 5 in the Appendix. The above face attribute detection classifier, which had an accuracy of 0.954 in the non-private setting, has a performance of 0.932 at this level.

Differential privacy hurts fairness in ERM The effect on differential privacy on fairness (bottom half of Table 1) is also quite consistent. The gap between the majority group and the minority group (or, more precisely, the best-performing and the worst-performing demographic subgroup) widens with increased privacy. In face recognition, for example, the accuracy gap between the two groups is 0.556 without differential privacy, but 0.770 at the strictest privacy level.

Differential privacy increases fairness in Group DRO For Group DRO, we see the opposite effect. For 4/5 datasets, we see that differential privacy leads to an increase in fairness. For face recognition, for example, the gap goes from 0.514 in the non-private setting to 0.056 in the strictest, basically disappearing. This is also illustrated in the bar plots in Figure 2. See Figure 3 for similar bar plots of the topic classification results; we include similar plots for other tasks in the Appendix. We do also observe that this increase in privacy can be expensive in terms of overall performance (e.g. Trustpilot-US). Note that the increase in fairness at higher privacy levels is seemingly at odds with previous results suggesting that privacy and fairness conflict, e.g., Agarwal (2021). We return to this question in §3.4.

Note also that the only exception to the latter

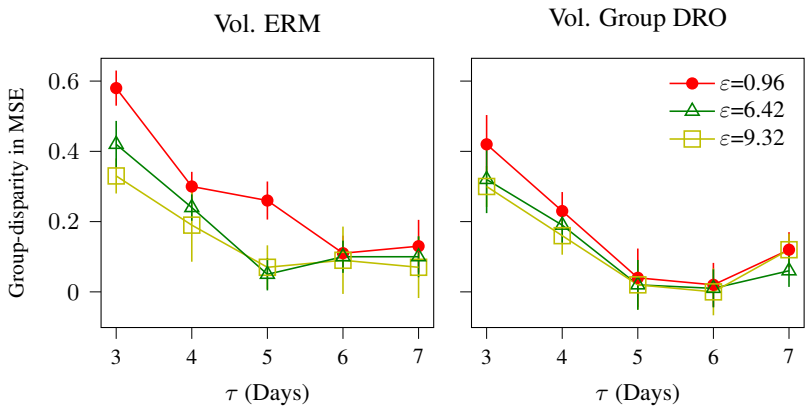

Figure 4: **Volatility Forecasting:** A comparison of group-disparity between subgroups for increasing temporal volatility windows (τ) and privacy budgets (ε), over 5 independent runs.

trend is for volatility forecasting, where differential privacy hurts fairness both in ERM and Group DRO (though Group DRO mitigates the disparity). This speech-based prediction is the only regression task, and the only task for which we do not rely on pretrained models trained on public data.

For this task, we further analyze group disparity for varying temporal windows (τ) used to calculate target volatility values, along with increasingly strict privacy budgets (ε) in Figure 4. The disparity between subgroups widens with stricter privacy guarantees (Bagdasaryan, Poursaeed, and Shmatikov, 2019). This gap is significant for lower values of τ, strengthening the hypothesis that short-term volatility forecasting is much harder than long-term (Qin and Yang, 2019), especially for minority classes due to the disproportionate impact of noise. Comparing ERM and Group DRO, we find Group DRO mitigates this disparity gap. We observe disparity reduces with increasing temporal window, since stock prices over a larger time frame are comparatively more stable (Qin and Yang, 2019). As a consequence, the influence of Group DRO for higher τ $(6, 7)$ is reduced, despite facilitating faster convergence. Most importantly, we observe the power of Group DRO in mitigating the disparity caused by strict privacy safeguards ($\varepsilon = 0.96$) for crucial short term prediction ($\tau = 3$) tasks.

3.4 Discussion

It is well-known that differential privacy comes with a performance cost (Shokri and Shmatikov, 2015).[12] However, recent work has additionally shown that differential privacy is at odds with most, if not all, definitions of fairness, including equalized risk (Ekstrand, Joshaghani, and Mehrpouyan, 2018; Cummings et al., 2019; Bagdasaryan, Poursaeed, and Shmatikov, 2019; Farrand et al., 2020). Our work makes two important contributions: (a) We evaluate and confirm this hypothesis at a larger scale than previous studies for standard empirical risk minimization; and (b) we point out that the opposite holds true in the context of Group Distributionally Robust Optimization: Here, adding differential privacy improves fairness (equalized risk).

While (b) at first seems to contradict the very hypothesis that (a) confirms – namely that privacy is at odds with fairness – we believe the explanation is quite simple, namely that we are observing two opposite trends (at the same time): On one hand, differential privacy adds disproportionate noise to minority group examples; but on the other hand, it adds Gaussian noise which acts as a regularizer to improve robust optimization.

In their evaluation of Group Distributionally Robust Optimization, Sagawa et al. (2020a) observe that robustness is only achieved in the context of heavy regulation; specifically, they show fairness improvements when they add ℓ_2 regularization or early stopping. The ℓ_2 regularization and early stopping did not increase fairness under ERM, but seemed to 'activate' Group DRO. This makes intuitive sense: Since regularized models cannot perfectly fit the training data, heavily regu-

[12] A multitude of algorithmic improvements have been proposed to mitigate the overall accuracy drop caused by the increased privacy protection -– including private sampling from hyperbolic word representation spaces (Feyisetan, Diethe, and Drake, 2019), Gaussian f-differential privacy (Bu et al. 2020), and gradient denoising (Nasr et al., 2020). It is yet to be examined, if the empirical application of such utility preservation techniques affects the disparate impact issue.

larized Group DRO sacrifices average performance for worst-case performance and obtain better generalization. In the absence of regularization, however, Group DRO is less effective.

In our experiments (§3), we add minimal regularization to Group DRO, following the implementation in Koh et al. (2021), but differential privacy, we argue, provides that additional regularization. To see this, remember that DP-SGD works by Gaussian noise injection. Gaussian noise injection is known to be near-equivalent to ℓ_2-regularization and early stopping (Bishop, 1995). DP-SGD simply makes the trade-off more urgent.

4 Related Work

Fair machine learning Early work on mitigating group-level disparities included oversampling (Shen, Lin, and Huang, 2016; Guo and Viktor, 2004) and undersampling (Drumnond, 2003; Barandela et al., 2003), as well as instance weighting (Shimodaira, 2000). Other proposals modify existing training algorithms or cost functions to obtain fairness (Khan et al., 2017; Chung, Lin, and Yang, 2015). In the context of large-scale deep neural networks, Group DRO is a particularly interesting approach to mitigating group-level disparities (Creager, Jacobsen, and Zemel, 2021). See Williamson and Menon (2019) and Corbett-Davies and Goel (2018) for interesting discussions of how fairness has been measured. More recent alternatives to Group DRO include Invariant Risk Minimization (Arjovsky et al., 2020), Spectral Decoupling (Pezeshki et al., 2020) and Adaptive Risk Minimization (Zhang et al., 2021). We ran experiments with both Invariant Risk Minimization and Spectral Decoupling, but they performed much worse than Group DRO.

Fairness and privacy Recent studies suggest that privacy-preserving methods such as differential privacy tend to disproportionately affect minority class samples (Ekstrand, Joshaghani, and Mehrpouyan, 2018; Cummings et al., 2019; Bagdasaryan, Poursaeed, and Shmatikov, 2019; Farrand et al., 2020). Pannekoek and Spigler (2021) show that it *is* possible to learn *somewhat private* and *somewhat fair* classifiers, in their case by combining differential privacy and reject option classification. Jagielski et al. (2019) introduced the so-called DP-oracle-learner, derived from an *oracle-efficient* algorithm (Agarwal et al., 2018), which satisfies equalized odds, an alternative notion of fairness (Williamson and Menon, 2019). Lyu et al. (2020) introduced Differentially Private GANs (DP-GANs), while Tran, Fioretto, and Van Hentenryck (2020) utilize Lagrangian duality to integrate fairness constraints to protected attributes. Group DRO has, to the best of our knowledge, not been studied under differential privacy before.

5 Conclusions

In §2, we summarized previous work suggesting that differential privacy and fairness are at odds. In §3, we then confirmed this hypothesis at scale, across five datasets, spanning four tasks and three modalities, showing that for Empirical Risk Minimization, stricter levels of privacy consistently *hurt* fairness. This holds true even after pretraining on large-scale public datasets (Kerrigan, Slack, and Tuyls, 2020). In the context of Group-aware Distributionally Robust Optimization (Group DRO) (Sagawa et al., 2020a), however, which is designed to mitigate group-level performance disparities (optimizing for equalized risk), we saw the opposite effect: Strict levels of differential privacy were associated with an *increase* in fairness. In §3.4, we discuss how this aligns well with the observation that Group DRO works best in the context of heavy ℓ_2 regularization, keeping in mind that Gaussian noise injection is near-equivalent to ℓ_2 regularization (Bishop, 1995).

References

Abadi, M.; Chu, A.; Goodfellow, I.; McMahan, H. B.; Mironov, I.; Talwar, K.; and Zhang, L. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, 308–318.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, 60–69. PMLR.

Agarwal, S. 2021. Trade-Offs between Fairness and Privacy in Machine Learning. In *IJCAI 2021 Workshop on AI for Social Good*.

Alvim, M. S.; Andrés, M. E.; Chatzikokolakis, K.; Degano, P.; and Palamidessi, C. 2011. Differential privacy: on the trade-off between utility and information leakage. In *International Workshop on Formal Aspects in Security and Trust*, 39–54. Springer.

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2020. Invariant Risk Minimization. arXiv:1907.02893.

Bagdasaryan, E.; Poursaeed, O.; and Shmatikov, V. 2019. Differential Privacy Has Disparate Impact on Model Accuracy. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Barandela, R.; Rangel, E.; Sánchez, J. S.; and Ferri, F. J. 2003. Restricted decontamination for the imbalanced training sample problem. In *Iberoamerican congress on pattern recognition*, 424–431. Springer.

Bertsimas, D.; Farias, V. F.; and Trichakis, N. 2011. The Price of Fairness. *Oper. Res.*, 59(1): 17–31.

Bishop, C. M. 1995. Training with Noise is Equivalent to Tikhonov Regularization. *Neural Computation*, 7(1): 108–116.

Boersma, P.; and Van Heuven, V. 2001. Speak and unSpeak with PRAAT. *Glot International*, 5(9/10): 341–347.

Campbell, N. A. 1978. The Influence Function as an Aid in Outlier Detection in Discriminant Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(3): 251–258.

Chang, H.; and Shokri, R. 2021. On the Privacy Risks of Algorithmic Fairness. arXiv:2011.03731.

Chernick, M.; and Murthy, V. K. 1983. The Use of Influence Functions for Outlier Detection and Data Editing. *American Journal of Mathematical and Management Sciences*, 3: 47–61.

Chung, Y.-A.; Lin, H.-T.; and Yang, S.-W. 2015. Cost-aware pre-training for multiclass cost-sensitive deep learning. *arXiv preprint arXiv:1511.09337*.

Corbett-Davies, S.; and Goel, S. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. arXiv:1808.00023.

Creager, E.; Jacobsen, J.-H.; and Zemel, R. 2021. Environment Inference for Invariant Learning. arXiv:2010.07249.

Cummings, R.; Gupta, V.; Kimpara, D.; and Morgenstern, J. 2019. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 309–315.

Desai, S.; Zhan, H.; and Aly, A. 2019. Evaluating Lottery Tickets Under Distributional Shifts. *CoRR*, abs/1910.12708.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019, Vol. 1*, 4171–4186. Minneapolis, Min.: Association for Computational Linguistics.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J. S.; and Pontil, M. 2018. Empirical Risk Minimization Under Fairness Constraints. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Drumnond, C. 2003. Class Imbalance and Cost Sensitivity: Why Undersampling beats Oversampling. In *ICML-KDD 2003 Workshop: Learning from Imbalanced Datasets*.

Dwork, C.; McSherry, F.; Nissim, K.; and Smith, A. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Proceedings of the Third Conference on Theory of Cryptography*, TCC'06, 265–284. Berlin, Heidelberg: Springer-Verlag. ISBN 3540327312.

Ekstrand, M. D.; Joshaghani, R.; and Mehrpouyan, H. 2018. Privacy for All: Ensuring Fair and Equitable Privacy Protections. In Friedler, S. A.; and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 35–47. New York, NY, USA: PMLR.

Farrand, T.; Mireshghallah, F.; Singh, S.; and Trask, A. 2020. Neither Private Nor Fair: Impact of Data Imbalance on Utility and Fairness in Differential Privacy. In *Proceedings of the 2020 Workshop on Privacy-Preserving Machine Learning in Practice*, 15–19.

Feyisetan, O.; Diethe, T.; and Drake, T. 2019. Leveraging Hierarchical Representations for Preserving Privacy and Utility in Text. In *2019 IEEE International Conference on Data Mining (ICDM)*, 210–219. IEEE Computer Society.

Guo, H.; and Viktor, H. L. 2004. Learning from imbalanced data sets with boosting and data generation: the databoost-im approach. *ACM Sigkdd Explorations Newsletter*, 6(1): 30–39.

Gupta, A.; Thadani, K.; and O'Hare, N. 2020. Effective Few-Shot Classification with Transfer Learning. In *Proceedings of the 28th International Conference on Computational Linguistics*, 1061–1066. Barcelona, Spain (Online): International Committee on Computational Linguistics.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385.

Hovy, D.; Johannsen, A.; and Søgaard, A. 2015. User review sites as a resource for large-scale sociolinguistic studies. In *Proceedings of the 24th international conference on World Wide Web*, 452–461.

Hu, W.; Niu, G.; Sato, I.; and Sugiyama, M. 2018. Does Distributionally Robust Supervised Learning Give Robust Classifiers? In Dy, J.; and Krause, A.,

eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2029–2037. PMLR.

Jagielski, M.; Kearns, M.; Mao, J.; Oprea, A.; Roth, A.; Sharifi-Malvajerdi, S.; and Ullman, J. 2019. Differentially private fair learning. In *International Conference on Machine Learning*, 3000–3008. PMLR.

Kerrigan, G.; Slack, D.; and Tuyls, J. 2020. Differentially Private Language Models Benefit from Public Pre-training. *ArXiv*, abs/2009.05886.

Khan, S. H.; Hayat, M.; Bennamoun, M.; Sohel, F. A.; and Togneri, R. 2017. Cost-sensitive learning of deep feature representations from imbalanced data. *IEEE transactions on neural networks and learning systems*, 29(8): 3573–3587.

Kogan, S.; Levin, D.; Routledge, B. R.; Sagi, J. S.; and Smith, N. A. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 272–280.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; Lee, T.; David, E.; Stavness, I.; Guo, W.; Earnshaw, B. A.; Haque, I. S.; Beery, S.; Leskovec, J.; Kundaje, A.; Pierson, E.; Levine, S.; Finn, C.; and Liang, P. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. arXiv:2012.07421.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 3730–3738.

Lyu, L.; Li, Y.; Nandakumar, K.; Yu, J.; and Ma, X. 2020. How to democratise and protect AI: fair and differentially private decentralised deep learning. *IEEE Transactions on Dependable and Secure Computing*.

McMahan, B.; Andrew, G.; Mironov, I.; Papernot, N.; Kairouz, P.; Chien, S.; and Úlfar Erlingsson. 2018. A General Approach to Adding Differential Privacy to Iterative Training Procedures. Workshop on Privacy Preserving Machine Learning (NeurIPS 2018).

Mironov, I. 2017. Rényi Differential Privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, 263–275.

Oren, Y.; Sagawa, S.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally Robust Language Modeling. In *EMNLP/IJCNLP (1)*, 4226–4236.

Pannekoek, M.; and Spigler, G. 2021. Investigating Trade-offs in Utility, Fairness and Differential Privacy in Neural Networks. arXiv:2102.05975.

Pezeshki, M.; Kaba, S.-O.; Bengio, Y.; Courville, A.; Precup, D.; and Lajoie, G. 2020. Gradient Starvation: A Learning Proclivity in Neural Networks. arXiv:2011.09468.

Qin, Y.; and Yang, Y. 2019. What You Say and How You Say It Matters: Predicting Stock Volatility Using Verbal and Vocal Cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 390–401. Florence, Italy: Association for Computational Linguistics.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, Massachussets: Belknap Press of Harvard University Press, 1 edition. ISBN 0-674-88014-5.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020a. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*.

Sagawa, S.; Raghunathan, A.; Koh, P. W.; and Liang, P. 2020b. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. arXiv:2005.04345.

Sanh, V.; Debut, L.; Chaumond, J.; and Wolf, T. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Schler, J.; Koppel, M.; Argamon, S.; and Pennebaker, J. 2006. Effects of Age and Gender on Blogging. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*.

Shen, L.; Lin, Z.; and Huang, Q. 2016. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, 467–482. Springer.

Shimodaira, H. 2000. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2): 227–244.

Shokri, R.; and Shmatikov, V. 2015. Privacy-preserving deep learning. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, 1310–1321.

Tran, C.; Fioretto, F.; and Van Hentenryck, P. 2020. Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach. *arXiv preprint arXiv:2009.12562*.

Veale, M.; and Binns, R. 2017. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2): 2053951717743530.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 353–355. Brussels, Belgium: Association for Computational Linguistics.

Williamson, R.; and Menon, A. 2019. Fairness risk measures. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 6786–6797. PMLR.

Yuan, M.; Kumar, V.; Ahmad, M. A.; and Teredesai, A. 2021. Assessing Fairness in Classification Parity of Machine Learning Models in Healthcare. arXiv:2102.03717.

Zhang, M.; Marklund, H.; Dhawan, N.; Gupta, A.; Levine, S.; and Finn, C. 2021. Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Distribution Shift. arXiv:2007.02931.

6 Appendix

6.1 Additional Figures

This section contains group-specific bar-plots for the performance on individual groups in the Trustpilot Corpus. For barplots on CelebA and Blog Authorship, see Figure 2 and 3.

6.2 Experimental Details

This section contains additional details surrounding the experiments described in §3.

CelebA We use the same processed version of the CelebA dataset as Sagawa et al. (2020a) and Koh et al. (2021), that is, we use the same train/val/test splits as Liu et al. (2015) with the *Blond Hair* attribute as the target with the *Male* attribute being the spuriously correlated variable. See group distribution in the training data in Table 2.

Non-Blond, Man	Blond, Man	Non-Blond, Woman	Blond, Woman
66874	1387	71629	22880

Table 2: Group distribution in the training set of CelebA

Blog Authorship Corpus In addition to the preprocessing described in §3, we split the data into a 60/20/20 train/val/test split (you can find the exact seed that generates the splits in our code). See group distribution in the training data in Table 3. The Blog Authorship Corpus can be downloaded

Group	Young, Man	Old, Man	Young, Woman	Old, Woman
Count	27222	2295	12750	2435

Table 3: Group distribution in the training set of Blog Authorship corpus

at: https://www.kaggle.com/rtatman/blog-authorship-corpus

Earnings Conference Calls Out of the 559 calls, we only include 535 datapoints that contain self-reported demographic attributes about gender. See Table 4 for group distributions for the training data. The target stock volatility variable is calculated following (Kogan et al., 2009; Qin and Yang, 2019), defined by:

$$v_{[t-\tau,t]} = \ln\left(\sqrt{\frac{\sum_{i=0}^{\tau}(r_{t-i}-\bar{r})^2}{\tau}}\right) \quad (4)$$

Here r_t is the return price at day t and $\bar{r}$ the mean of return prices over the period of $t-\tau$ to t. We refer to τ as the temporal volatility window in our experiments. The return price r_t is defined as $r_t = \frac{P_t}{P_{t-1}} - 1$ where P_t is the closing price on day t.

Group	Man	Woman
Count	333	42

Table 4: Group distribution in the training set of Earnings Conference Calls

Trustpilot We only include the datapoints that contains complete demographic attributes, i.e. the gender, age and location, but as with our topic classification experiments, we only study the group that we can define based on age and gender. All attributes are self-reported. For training we divide the reviews into the four resulting groups (*Old-Man*, *Young-Woman*, etc.) and downsample the largest groups to match the size of the smallest group. For validation as well as testing, we withhold 200 samples from each demographic with an even distribution among the ratings (1 to 5). The review scores are then binarized by grouping positive (4 and 5 stars) and negative (1 and 2 stars) and discarding neutral ones (3 stars). For a similar use of this binarization scheme, see Gupta, Thadani, and O'Hare (2020) and Desai, Zhan, and Aly (2019). See the group distributions for the training data in Table 5 and 6 for the US and UK tasks respectively.

Group	Young, Man	Old, Man	Young, Woman	Old, Woman
Count	7242	7210	7222	7255

Table 5: Group distribution in the training set of Trustpilot-US

Group	Young, Man	Old, Man	Young, Woman	Old, Woman
Count	18464	18693	18554	18693

Table 6: Group distribution in the training set of Trustpilot-UK

BiLSTM The BiLSTM model was trained using a Nvidia Tesla K80 GPU. We use a learning rate of $1e^{-2}$ and train using DP-SGD for 30 epochs using a virtual batch size of 32. The average sequence length of the audio embeddings is 159. We set the maximum sequence length to 150 as we did not observe a performance increase for higher values. We run 5 individual seeds for each configuration.

In our differentially private experiments with the BiLSTM (i.e Earnings Conference Calls), we fix the gradient clipping C to 0.8. By specifying various approximate target levels of $\varepsilon \in \{1, 5, 10\}$

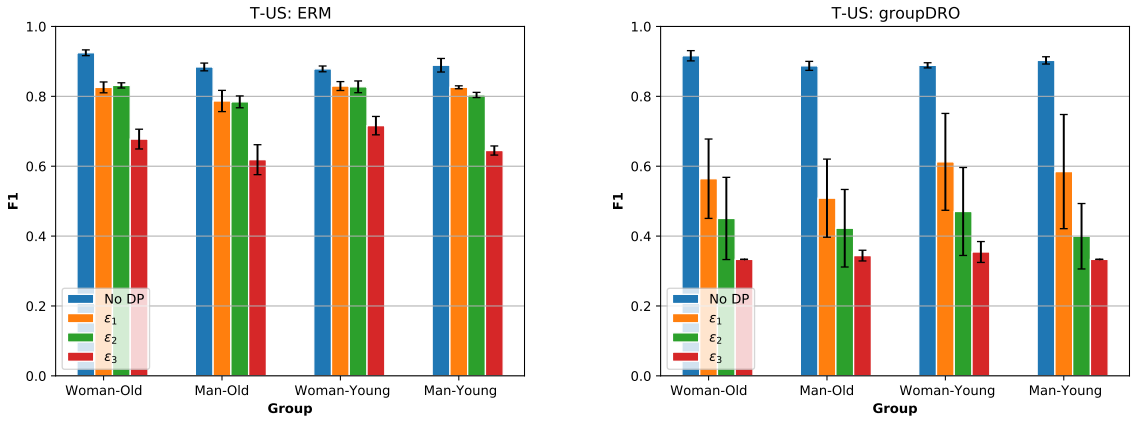

Figure 5: Performance of individual groups of increasing levels of ε for the Trustpilot-US corpus. Error bars show standard deviation over 3 individual seeds.

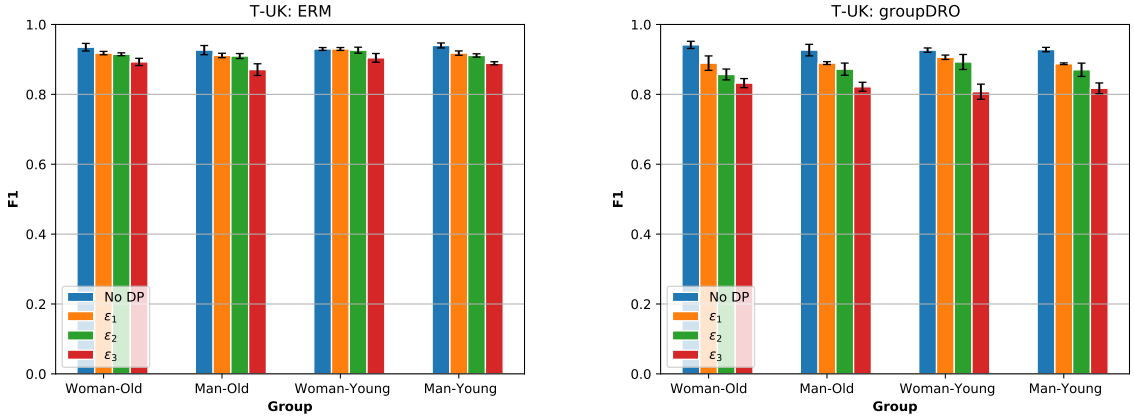

Figure 6: Performance of individual groups of increasing levels of ε for the Trustpilot-UK corpus. Error bars show standard deviation over 3 individual seeds.

a corresponding noise multiplier σ is computed with the Opacus framework, based on the batch size and number of training epochs.

DistilBERT DistilBERT is a small Transformer model trained by distilling BERT (Devlin et al., 2019) (bert-base-uncased). It has 3/5th of the parameters of bert-base-uncased, runs 60% faster, while preserving over 95% of the performance of bert-base-uncased, as measured on the GLUE language understanding benchmark (Wang et al., 2018).

We finetune DistilBERT on the Trustpilot corpus and Blog Authorship corpus for 20 epochs each, using a batch size of 8, accumulating gradient for a total virtual batch size of 16 using the built in Opcaus functionality. We limit the number of tokens in a sequence to 256 and use a learning rate of $5e^{-4}$ with the AdamW optimizer in addition to a weight decay of 0.01. Otherwise we use the default parameters defined in the Huggingface Transformers python package (version 4.4.2). The models are trained using a single Nvidia TitanRTX GPU and each configuration takes between 5 and 14 hours to run, depending on the size of that dataset and if DP is used or not. We run 3 individual seeds for each configuration.

In our differentially private experiments with DistilBERT (i.e. Blog Authorship and Trustpilot), we fix the gradient clipping C to 1.2 and by specifying various target levels of $\varepsilon \in \{1, 5, 10\}$ a corresponding noise multiplier σ is computed with the Opacus framework, based on the batch size and number of training epochs.

Resnet50 ResNet50 is a variant of the ResNet model (He et al., 2015), which has 48 convolution layers along with 1 max pooling and 1 average

pooling layer. It has 3.8 x 10^9 floating points operations.

We finetune our Resnet50 model on the CelebA dataset for 20 epochs using a batch size of 64. We optimize the model using standard stochastic gradient descent (SGD) with a learning rate of $1e^{-3}$, momentum of 0.9 and no weight decay. We train our models using a single Nvidia TitanRTX GPU and each configuration takes between 6 and 8 hours to run, depending on if DP is used or not. We run 3 individual seeds for each configuration.

As with the differentially private DistilBERT experiments, we also here fix the gradient clipping C to 1.2 and by specifying various target levels of $\varepsilon \in \{1, 5, 10\}$ a corresponding noise multiplier σ is computed with the Opacus framework, based on the batch size and number of training epochs.