Does Differential Privacy Impact Bias in Pretrained Language Models?

Md. Khairul Islam[1], Andrew Wang[1], Tianhao Wang[1], Yangfeng Ji[1],
Judy Fox[1], Jieyu Zhao[2]
[1] University of Virginia, [2] University of Southern California
{mi3se, ajw7uhj, tianhao, yj3fs, cwk9mp}@virginia.edu, jieyuz@usc.edu

Abstract

Differential privacy (DP) is applied when fine-tuning pre-trained language models (LMs) to limit leakage of training examples. While most DP research has focused on improving a model's privacy-utility tradeoff, some find that DP can be unfair to or biased against underrepresented groups. In this work, we extensively analyze the impact of DP on bias in LMs. We find differentially private training can increase the model bias against protected groups w.r.t AUC-based bias metrics. DP makes it more difficult for the model to differentiate between the positive and negative examples from the protected groups and other groups in the rest of the population. Our results also show that the impact of DP on bias is affected by both the privacy protection level and the underlying distribution of the dataset.

1 Introduction

In a data-driven world, an appropriate dataset is critical for fair and responsible decision-making and analysis. However, inequalities in the world or limitations in the collection process may restrict the database coverage across different minority groups and representations [26]. Identifying these discriminations [28] and insufficient coverages [29] helps reduce the database bias which can impact the models trained on them. Social media and online conversation platforms are places where millions of user relies on such text databases everyday. The user's confidentiality and fair decisions are very crucial for these natural language processing (NLP) tasks.

Pretrained transformer-based language models such as BERT [1] have led to significant advancements in research. Much of the success of natural language models (LMs) ultimately derives from the vast data used to train these models. However, the use of a large training dataset raises concerns about data privacy, where the model can be used to detect the presence of sensitive information in the training data. To defend against these attacks, Differentially private (DP) training techniques [2, 3] have been used during the model training or fine-tuning process [4]. These techniques ensure that a model does not leak sensitive training data. Otherwise, an attacker can extract the dataset [5] using inference attacks.

However, recent works in data privacy indicate that DP training may cause machine learning models to become more biased [6, 7, 8]. However, most of these works focus on the computer vision domain or tabular datasets. With the wide usage of NLP models and the urgency to realize trustworthy NLP, we need to understand

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

whether we can obtain an NLP model equipped with both privacy and fairness, especially for the pre-trained language models. An NLP model is considered biased when the model is unable to perform on protected social groups equally as well as on others. For example, prior research has demonstrated a coreference resolution model can behave very differently for different demographic groups [9, 10]. DP may introduce bias because it steers a model away from relying on a select few data points, causing that model to attend poorly to social groups that are underrepresented in the training data.

In this work, we explore the impact of differential privacy on model bias in the pre-trained BERT language model. The degree of which can be tuned by adjusting the privacy budget parameter. We train the model with different privacy budgets and measure the bias across six identity subgroups using multiple metrics. We consider bias in the context of the toxic language detection task, which has been shown to produce biased models [11]. We choose two popular datasets, Jigsaw Unintended Bias [12] and the Measuring Hate Speech from UCBerkeley [13]. We use both prediction and probability-based bias metrics to analyze the effect of DP on the bias from different perspectives. We then investigate them in each identity group for any discriminatory behavior against them.

Contributions: In this work, we present a detailed analysis of the impact of DP training on bias in fine-tuned language models. We present our results on two popular hate speech datasets by training our models at different privacy levels and analyzing how it affects the model bias. We show that DP training makes the model more biased in AUC-based metrics. DP also negatively affects the model's utility when adopted to pertained language models. Our findings will give new insights into the privacy and bias trade-off, which can help NLP researchers incorporate DP into their works. The code for our work can be found at: https://github.com/khairulislam/DP-on-NLP-Bias.

2 Related Work

Fairness is critical to data-driven systems [26] to make more accurate decisions at a larger scale while managing millions of user data. Misrepresentations of minorities and insufficient data coverage across different modalities [29] introduce bias in modern database systems. [26] proposed data-centric approaches to identify and resolve these issues in data. [28] formalized the discriminations as a database repair problem and provided sufficient conditions to train fair classifiers. Their solution correctly captures subtle fairness violations in the data and provides provable fairness guarantees about classifiers trained on them. [29] proposed an efficient approach to identify regions with insufficient coverage across different groups over multiple relational tables in a database. The solution can efficiently identify inequalities in the collected data and help with a more fair solution.

Prior research has shown from a theoretical perspective that DP has a detrimental effect on model fairness [6, 14]. [6] assume the conditions of "pure DP" [2], and demonstrate that such a model cannot achieve perfect equal opportunity between social groups. [14] finds that model fairness has a disproportionately negative impact on accuracy for certain social groups.

In computer vision, recent works have empirically investigated the effects of DP on model fairness in models with more realistic privacy settings. Empirical analyses have found that DP can worsen accuracy for certain subgroups in image recognition tasks [7, 8] and synthetic data generation tasks [15]. [16] showed both DP-SGD and PATE have a disparate impact on the under-represented groups, but PATE has a significantly less disproportionate impact on utility compared to DP-SGD.

Bagdasaryan and Shmatikov [7] found that DP can worsen model bias in sentiment analysis. However, they only considered a single bias metric (accuracy degradation between privileged and unprivileged groups) on a single dataset using a glove-based model. In this work, we analyze pre-trained BERT models on multiple datasets using multiple bias metrics. Balancing between fairness and privacy can significantly impact using private models in practice.

Private-FairNR [6] algorithm approximately satisfies fairness for a private learner sampling hypothesis. [17] formally guaranteed the privacy of extracted text representation, while also helping model fairness. They aimed

to protect the test phase privacy of end users while adopting local DP (LDP) with the Laplace mechanism.

3 Model Bias in NLP

Evaluating biases in NLP models requires a metric over some demographic groups. In this section, we describe the terminology for those groups and the metrics for bias evaluation.

3.1 Terminology

Protected attributes refer to sensitive attributes such as gender and race that should not be used to discriminate against individuals [18]. *Bias* occurs when a model experiences a degradation in performance when inferring examples pertaining to certain social groups implied by a protected attribute such as gender or race. In our calculations of bias, we refer to a *subgroup* as the social group whose bias we are measuring and *background* as the rest of the evaluation set [12]. *Prediction-based bias metrics* calculate the bias against the protected attributes using the model's predicted label (e.g. positive/negative), whereas *Probability-based bias metrics* use the prediction probability to calculate bias. These definitions of bias metrics are done following [19].

3.2 Protected Attributes

Bias in NLP has been well studied within the protected attributes of *gender* [20] and *race* [11]. Following this, we examined bias for sensitive attributes *gender* and *race*. In *gender* attribute the identity subgroups are *male/men*, *female/women*, and *transgender*. For *race* attribute the identity subgroups are *white*, *black*, and *asian*.

3.3 Bias Evaluation Metrics

A *degradation in performance* indicative of model bias can be measured in different ways. We consider metrics such as equality of odds metrics because of their prolific use in other NLP model fairness literature [18, 12, 21] and Bias-AUC because of its use as the benchmark in the Jigsaw Unintended Bias competition. We summarize all the different bias evaluation metrics we consider in Table 11. The implementations follow [18, 12] and [21]. More details about these metrics are in Appendix 6.

Bias Metric	**Formulation**	**Short form**
Demographic Parity [18]	$1 - \lvert p(\hat{Y} = 1 \lvert A = 1) - p(\hat{Y} = 1 \lvert A = 0) \rvert$	parity
Equality of Opportunity (w.r.t $Y = 1$)	$1 - \lvert p(\hat{Y} = 1 \lvert Y = 1, A = 1) - p(\hat{Y} = 1 \lvert Y = 1, A = 0) \rvert$	EqOpp1
Equality of Opportunity (w.r.t $Y = 0$)	$1 - \lvert p(\hat{Y} = 1 \lvert Y = 0, A = 1) - p(\hat{Y} = 1 \lvert Y = 0, A = 0) \rvert$	EqOpp0
Equality of Odds [18]	$0.5 \times [EqOpp0 + EqOpp1]$	EqOdd
Protected Accuracy [21]	$p(\hat{Y} = y \lvert Y = y, A = 1), y \in \{0, 1\}$	p-acc
Subgroup AUC [12]	$AUC(D_g^- + D_g^+)$	
Background Pos, Subgroup Neg [12]	$AUC(D^+ + D_g^-)$	BPSN
Background Neg, Subgroup Pos [12]	$AUC(D^- + D_g^+)$	BNSP

Table 11: X, Y, A denotes the input, label, and sensitive attribute (e.g. male, female). $\hat{Y}$ and p are the model's prediction and the output probability. All metrics are in the range $[0 - 1]$ and a higher value is better (less bias). D_g^+ and D_g^- are the set of positive and negative examples in the identity subgroup g. D^+ and D^- are the set of positive and negative examples outside g.

4 Differential Privacy

Differential privacy (DP) [2] aims to preserve privacy using a quantifiable protection guarantee and acceptable utility in the context of statistical information disclosure. It is the *de facto* definition for privacy. In the context of our work, we use the notion of (ε, δ)-privacy. Following [2], if we have some arbitrary operation $\mathcal{A}$ with output space S and two datasets D, D' that differ in only a single record, then we can formulate (ε, δ)-privacy as $\Pr(\mathcal{A}(D) \in S) \leq e^{\epsilon} \Pr(\mathcal{A}(D') \in S) + \delta$.

By limiting any effect due to the inclusion of one individual's data (by the parameter ϵ), the DP notion approximates the effect of "opting-out": whether an individual's data is included or not does not influence the result much, thus the fact that the individual participated in the data release is protected.

To satisfy DP, noise is added to the aggregated-level results such that an individual's information disclosure is bounded. Our implementation in this paper uses the Gaussian mechanism [22] to guarantee (ϵ, δ)-DP.

DP in machine learning: When training models with DP, perturbations are added to the gradients (i.e., clipping the gradients and then adding Gaussian noise) [3]. More specifically, during the t-th iteration the optimizer will compute noisy gradients as:

$$g^t = \frac{1}{|B|}\Big(\sum_{x_i \in B} \hat{g}_i^t + \mathcal{N}\left(0, \sigma^2 C^2 I\right)\Big),$$

where B is a subsampled batch used to compute the gradients, w^{t-1} is the current model before t-th iteration, σ is noise multiplier,

$$\hat{g}_i^t = \nabla f(x_i; w^{t-1}) \min\{1, \frac{C}{\|\nabla f(x_i; w^{t-1})\|_2}\}$$

(i.e., each gradient is clipped by C, so that $\sum \hat{g}_i^t$ has bounded ℓ_2-sensitivity and we can use the Gaussian mechanism to ensure DP), and g^t is the (noisy) gradient used to update the model.

Training a model requires multiple training epochs. Our formulation of DP is amenable to this practice. If we have k operations that satisfy some ε privacy constraint, we can combine those operations and maintain DP for $O(\sqrt{k}\epsilon)$. So we refer to ϵ as the *privacy budget* of a privacy-preserving algorithm.

Impact of DP Methods on Fairness: Gaussian Mechanism introduces enough noise so that the contribution of individual data points to model decision-making is limited. However, a byproduct of this approach is that the distinguishing features of underrepresented social groups within the dataset can be "smoothed over." Thus, we conjecture that the DP model attends disproportionately worse to the underrepresented social groups and is thus biased. In what follows, we present evidence supporting the fact that DP negatively impacts the model fairness.

5 Datasets

We choose two popular toxicity detection datasets for our study, Jigsaw Unintended Bias [12] and UCBerkeley Hate Speech [13]. Both datasets (1) have target labels so that we can use supervised learning, (2) are for text classification using NLP techniques, and (3) have annotated social groups for all examples.

5.1 Jigsaw Unintended Bias

The Jigsaw Unintended Bias dataset was developed to learn and minimize any unintended bias against different identities that a machine learning model learns to predict toxicity [27] from online comments. Here toxicity is defined as anything rude or disrespectful that can make someone leave a discussion. The dataset collected by Jigsaw and Google has annotations for demographic groups by disability, gender, race or ethnicity, religion, and

Group	Jigsaw				UCBerkeley			
	Train		Test		Train		Test	
	class 0	class 1	class 0	class 1	class 0	class 1	class 0	class 1
Male	3187	3375	1792	320	2361	796	502	171
Female	3950	3639	2252	350	4852	2305	1042	511
Transgender	158	287	103	26	882	244	196	51
White	1507	3612	825	353	1694	643	378	132
Black	901	2369	515	246	2103	1568	483	337
Asian	358	282	196	21	831	207	195	53
Total	144334	72167	89543	7777	19376	7618	4142	1643

Table 12: Distribution of identities in both datasets. Total is the class distribution in the dataset after pre-processing. Class 1 is for toxic and 0 for non-toxic.

sexual orientation. The complete dataset has about 2 million examples. We report the label distribution for each identity in Table 12.

Train/Validation/Test split: We undersampled the training dataset using a 2:1 ratio between the non-toxic and toxic labels. Due to computing resource limitations, we halved the training set, preserving the 2:1 label distribution. This yielded a training set with 144,334 non-toxic examples and 72,167 toxic examples. We use the pre-existing splits from the original source for the validation and test data. This yielded a test and a validation set each with 97,320 examples.

5.2 UCBerkeley Hate Speech

This dataset [1] is a collection of online comments from three major social media platforms (YouTube, Twitter, and Reddit), labeled by human annotators through crowd-sourcing [13]. It provides a unique way to measure hate speech at eight theorized qualitative from genocidal hate speech to counter speech. The dataset comes with annotations for the targeted group in the comment text.

Pre-processing: The original dataset has 135,556 comments and the annotations for 'hatespeech' contain 3 classes: 0 for neutral or counter speech, 1 when the annotator is unclear, and 2 for hate speech. For the simplicity of the experiment, we dropped comments with label 1, converting the task to a binary classification where hate speech is a positive class and non-hate speech is negative. The dataset also had multiple annotations per comment. We aggregated the annotations for each comment. If any comment had the same label at least from 50% of the annotators, then it was chosen as true, otherwise false. After aggregation, we had 38,564 comments left. Additionally, the dataset contains transgender identity labels split into multiple groups (transgender_men, transgender_women, transgender_unspecified). We combined them together in a single transgender column for bias calculation.

Train/Validation/Test split: We randomly split the aggregated data into train, validation, and test sets using a 70:15:15 ratio.

[1] https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech

6 Experimental Setup

Model: We use the pre-trained BERT-base-uncased model from HuggingFace [2] to perform all our experiments in this section. For training the model on downstream tasks we choose only to train the last three layers (final encoder layer, pooler, classifier). The rest of the layers were frozen, yielding 7.6 M trainable parameters out of a total of 109M. We choose to train only these layers because: 1) DP is more effective when applied to fewer layers, and 2) we can utilize BERT's rich pre-trained embeddings.

Input texts were tokenized using the BERT-base-uncased tokenizer from HuggingFace. The comment texts were generally not very lengthy, so we kept the maximum sequence length to 128 across both datasets. The batch size was set to 64.

Optimization: We use the Adam optimizer with cross-entropy loss and learning rate 10^{-3}. We train each model for a maximum of 10 epochs. At each epoch, the trained model is evaluated on the validation set and saved if the F1 score improves. Early stopping patience was 3. We also used a learning rate scheduler (ReduceLROnPlateau) to reduce the learning rate by a factor of 0.1 if the validation F1 score does not improve for more than one epoch.

Privacy: We use the Pytorch Opacus library [23]. It provides a privacy engine to train models with DP-SGD [3]. DP-SGD was chosen since it is the most widely used one in the related works [7, 14, 24], supports iterative training process and is available as a framework. We use the *make_private_with_epsilon* method offered by the library, which takes as input the model to be trained, optimizer, training data, number of epochs, target ϵ, target δ and maximum gradient norm. The target epsilon is the privacy budget we want to achieve. For a reasonable privacy guarantee, ϵ should be set below 10 [3] and this setting has been followed in other applications of DP on NLP [7, 24, 17]. For our task, we experimented with five different target epsilons 0.5, 1.0, 3.0, 6.0, and 9.0. The smaller the value the more private the model is. This will show us the change in model behavior at different privacy levels.

Evaluation: We tune the training process using the F1 score on the validation set, then checkpoint the best model based on that, and finally use that model to evaluate the test set. We have presented the final test results in Section 7. Each experiment is run three times with arbitrarily chosen random seeds 2022, 42, and 888. The average score is reported in Table 13.

Bias Evaluation Metrics We use the following metrics for calculating bias during our experiments:

- *Equality of Odds (EqOdd) [18]:* Widely used to measure unequal treatments against protected groups in the dataset. The metric combines the disparity in false positive and true positive rates for two social groups in the same protected class.
- *Demographic Parity (Parity) [18]:* Enforces the model's prediction to be independent of the protected attribute. The metric computes the difference in likelihood between unprotected or protected examples to be classified as positive.
- *Subgroup AUC, BPSN, and BNSP [12]:* The Subgroup AUC, BPSN, and BNSP metrics measure the unintended bias in the dataset based on the AUC metric. AUC is threshold agnostic, unlike equality of odds or other prediction-based metrics that require converting model predictions into positive or negative classes using some threshold. The choice of threshold can change the results and provide misleading measurements. These metrics can be used to find new and potentially subtle biases in models.

[2] https://huggingface.co/bert-base-uncased

7 Results

7.1 Overall Results

Here we present the impact of adding DP on the overall model utility for both datasets. Table 13 shows that the model utility decreases with stricter privacy (smaller ϵ). However, for the UCBerkeley dataset, the false positive rate increases, and the recall drops significantly. This shows the model predicts fewer positive cases with added privacy. The recall drop is also significant for Jigsaw. This decrease in overall performance also impacts the performance of the identity subgroups.

Metric	Jigsaw - Privacy Budget (ϵ)						UCBerkeley - Privacy Budget (ϵ)					
	∞	≤ 9.0	≤ 6.0	≤ 3.0	≤ 1.0	≤ 0.5	∞	≤ 9.0	≤ 6.0	≤ 3.0	≤ 1.0	≤ 0.5
Acc	0.911	0.887	0.886	0.884	0.871	0.870	0.807	0.787	0.787	0.785	0.779	0.772
F1	0.593	0.522	0.518	0.508	0.459	0.440	0.647	0.554	0.559	0.539	0.523	0.480
AUC	0.946	0.920	0.918	0.913	0.886	0.872	0.855	0.813	0.819	0.814	0.802	0.790
FPR	0.080	0.102	0.103	0.105	0.113	0.111	0.120	0.086	0.089	0.079	0.082	0.069
TPR	0.809	0.768	0.763	0.751	0.686	0.642	0.623	0.469	0.476	0.443	0.427	0.371

Table 13: Overall model performance. The results are best for the non-DP training ($\epsilon \to \infty$) and worst at the most strict privacy budget, $\epsilon \leq 0.5$.

7.2 Prediction Based Metrics

Equality of Odds, parity, and protected accuracy are prediction-based bias metrics. They calculate the bias score based on the model's prediction. We present the results in Table 14. For each identity, we report the best and the worst results, and the privacy budget that achieves that result. The closer these scores are to 1, the less the bias is.

Group		Jigsaw			UCBerkeley		
		EqOdd	**parity**	**p-acc**	**EqOdd**	**parity**	**p-acc**
Male	min	0.894 (6.0)	0.852 (1.0)	0.741 (1.0)	0.955 (∞)	0.763 (∞)	0.765 (1.0)
	max	0.928 (0.5)	0.872 (∞)	0.801 (∞)	0.983 (0.5)	0.868 (0.5)	0.799 (∞)
Female	min	0.932 (9.0)	0.851 (1.0)	0.785 (9.0)	0.937 (∞)	0.890 (∞)	0.717 (0.5)
	max	0.940 (0.5)	0.872 (∞)	0.822 (∞)	0.957 (3.0)	0.929 (0.5)	0.756 (∞)
Transgender	min	0.818 (9.0)	0.842 (1.0)	0.674 (∞)	0.910 (∞)	0.740 (∞)	0.815 (9.0)
	max	0.952 (0.5)	0.863 (∞)	0.785 (0.5)	0.962 (0.5)	0.848 (0.5)	0.839 (∞)
White	min	0.734 (9.0)	0.853 (1.0)	0.588 (6.0)	0.917 (9.0)	0.752 (∞)	0.769 (9.0)
	max	0.842 (0.5)	0.875 (∞)	0.647 (0.5)	0.940 (1.0)	0.851 (0.5)	0.800 (∞)
Black	min	0.777 (∞)	0.847 (1.0)	0.636 (9.0)	0.812 (3.0)	0.836 (∞)	0.761 (0.5)
	max	0.901 (0.5)	0.871 (∞)	0.697 (0.5)	0.855 (∞)	0.924 (0.5)	0.821 (∞)
Asian	min	0.916 (∞)	0.842 (1.0)	0.814 (9.0)	0.871 (∞)	0.737 (∞)	0.823 (0.5)
	max	0.976 (0.5)	0.863 (∞)	0.859 (∞)	0.894 (0.5)	0.844 (0.5)	0.847 (∞)
Trend $\epsilon \downarrow$		$\uparrow$	$\downarrow$	$\uparrow\downarrow$	$\uparrow$	$\uparrow$	$\downarrow$

Table 14: Prediction Based Bias (Jigsaw). The privacy budget (ϵ) for each metric is mentioned in the parentheses. **The trends are not monotonic and can be mixed.** Smaller ϵ means stricter privacy.

Table 14 shows several trends depending on the dataset and metric. The equality of odds always improves with a strict privacy budget (small ϵ). However, this is due to a significant drop in recall (Figure 2) for most

groups. They are reduced to a smaller score range. Thus the TPR difference becomes smaller, improving the EqOpp1.

The trend in demographic parity is the opposite in both datasets. With a stricter privacy budget, parity decreased in the Jigsaw (2-3%) but increased in the UCBerkeley dataset (4-11%). An increase in this value indicates that the model's decision of whether the comment is toxic or not, is more independent of the protected group [18]. We show in Section 8.4 that DP increases positive predictions in Jigsaw and decreases them in UCBerkeley. More positive predictions increase the probability of disparity among different subgroups of Jigsaw. Similarly in UCBerkeley dataset, since there are fewer positive predictions from the model, the disparity based on positive outcomes decreases too.

The protected accuracy has mixed trends in the Jigsaw dataset, changing in either direction. In the UCBerkeley dataset, there is a 2-5% drop with DP training. The detailed plots for these metrics at each privacy budget and for each identity are available in our GitHub repo.

7.3 Probability Based Metrics

This section presents the bias calculated using the metrics presented by [12]. These metrics are dependent on the model's prediction probability, hence better representing the bias in the model's confidence. They are also threshold agnostic, unlike prediction-based metrics.

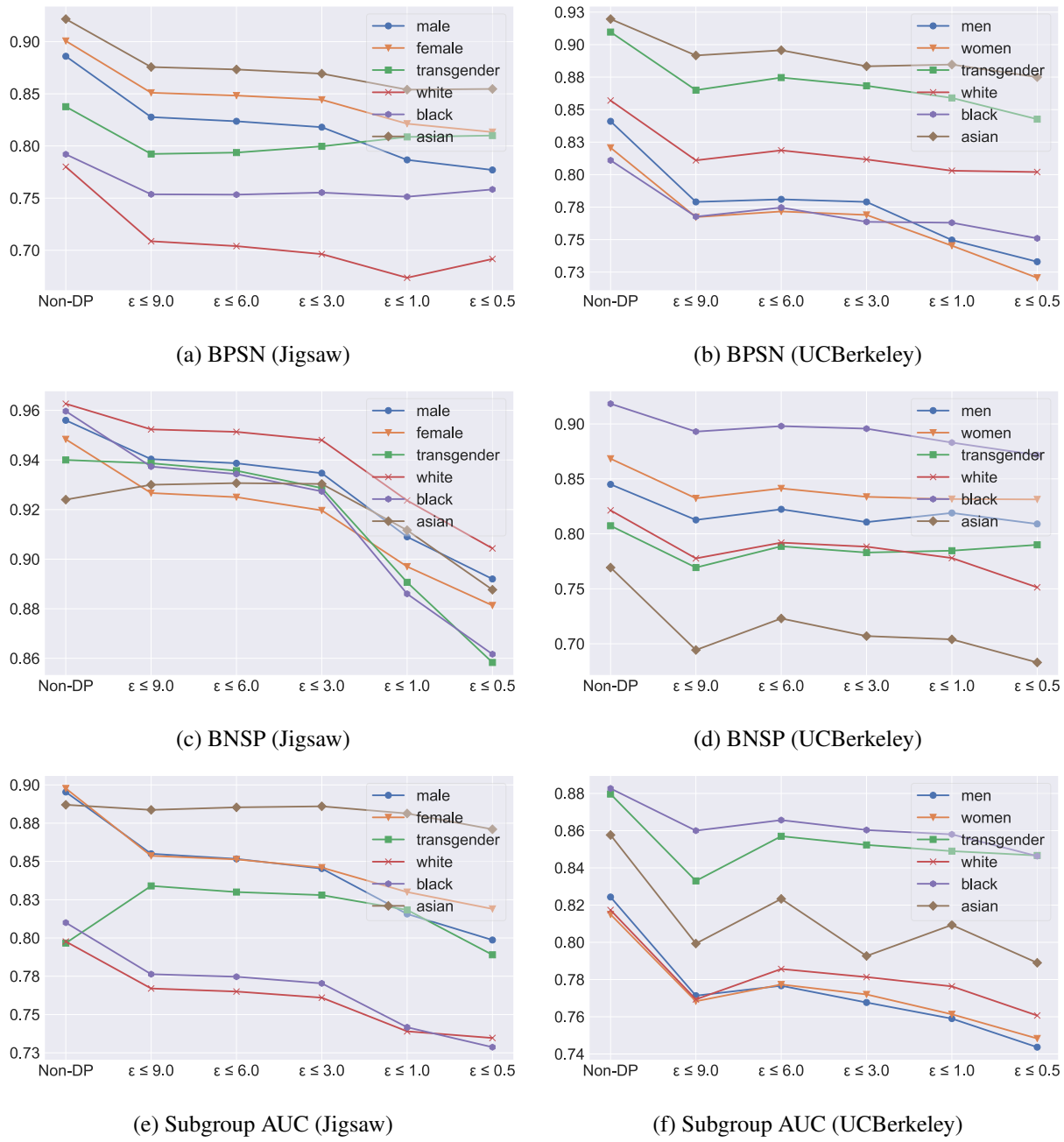
Figure 1 shows that for stricter privacy (smaller ϵ), both BNSP and BPSN drop significantly for most identities. A drop in BNSP means the scores for positive examples in these subgroups are lower than the scores for other negative examples in the background data. These examples would likely appear as false negatives within the subgroup at many thresholds [12].

Similarly, a drop in BPSN means scores for negative examples in these subgroups are higher than scores for other positive examples in the background. These examples would likely appear as false positives within these subgroups at many thresholds [12]. A decrease in the subgroup AUC score shows that the model can not understand and separate the positive and negative examples within the subgroup. These drops between non-DP training and training with DP at $\epsilon \leq 0.5$ are highlighted in Table 15, showing an increase in bias at stricter privacy budgets, compared to non-DP training.

	Jigsaw			UCBerkeley		
Group	**Δ Subgroup AUC**	**Δ BPSN**	**Δ BNSP**	**Δ Subgroup AUC**	**Δ BPSN**	**Δ BNSP**
Male	**0.097**	0.064	**0.109**	**0.081**	0.036	**0.108**
Female	0.079	0.067	0.087	0.067	**0.037**	0.100
Transgender	0.008	**0.082**	0.028	0.033	0.017	0.067
White	0.063	0.058	**0.088**	0.057	0.070	0.055
Black	**0.081**	**0.098**	0.034	0.036	0.047	**0.060**
Asian	0.016	0.036	0.067	**0.069**	**0.086**	0.045

Table 15: Decrease in probability-based bias from non-DP training to training with $\epsilon \leq 0.5$. The biggest drop along each metric column for each sensitive attribute (race, gender) is in bold. The DP model is 4-11% more biased in several identity groups.

Figure 1 shows some interesting cases. In the Jigsaw dataset, white and black identities have much lower AUC and BPSN scores compared to others. Similarly in the UCBerkeley dataset, men and women have much lower AUC and BPSN scores than other identities. This shows that the DP models more often tend to label non-toxic comments mentioning these identities as toxic, compared to the non-DP models. Additionally, DP amplifies the difference in the AUC gap between white and Asian subgroups in Jigsaw and white and black subgroups in UCBerkeley. The non-DP model already had a gap in AUC between them, but DP increases it.



(a) BPSN (Jigsaw)

(b) BPSN (UCBerkeley)

(c) BNSP (Jigsaw)

(d) BNSP (UCBerkeley)

(e) Subgroup AUC (Jigsaw)

(f) Subgroup AUC (UCBerkeley)

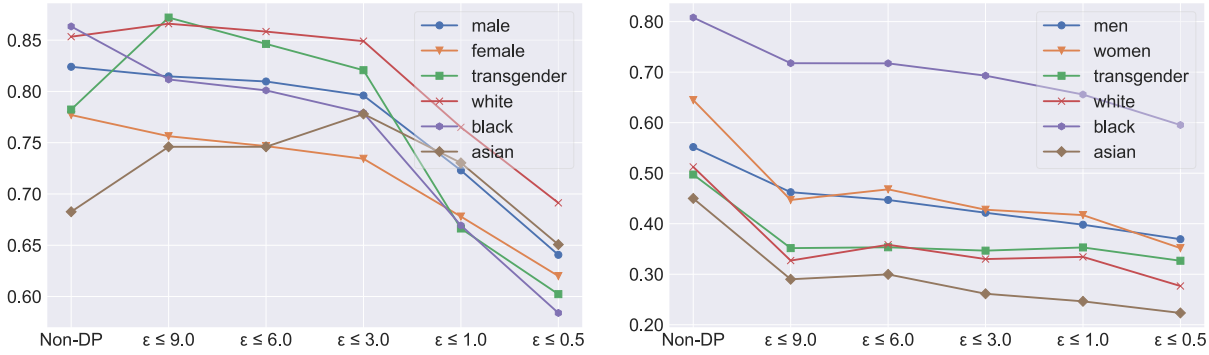
Figure 1: AUC based bias [12]. BNSP for Jigsaw and BPSN for UCBerkeley have a significant drop in value with a much smaller ϵ. The larger the drop, the more biased the model w.r.t that metric.

8 Discussion

8.1 DP's Positive Impact on Equality of Odds and Opportunities.

Equality of odds is a function of relative true positive and false positive rates between a subgroup and the background population. As such we investigate why the addition of noise does not decrease relative TPR (recall) and FPR. DP adds noise in the training phase, adversely affecting overall model performance (Table 13). The model experiences a degradation in the recall for all social groups as the privacy setting increases. So recall values

grow more similar. This minimizes the difference in TPR between a subgroup and a background population, contributing to an overall improvement in equality of odds. However, such a trend does not necessarily indicate a decrease in bias but instead indicates that a model is losing its ability to differentiate between the positive and negative classes. Figure 2 shows that the recall drops significantly for private training.



(a) Jigsaw - Recall

(b) UCBerkeley - Recall

Figure 2: Recall for each subgroup at different privacy budgets. Differential private training significantly reduces the model's ability to predict target toxic comments.

8.2 DP's Impact on Probability-based Bias.

The fall in overall model AUC scores also affects the subgroup AUC, BPSN, and BNSP, as shown in Table 15. The model makes more mistakes in differentiating the positive and negative examples between the subgroups and the background data, even within the subgroup itself. Thus introducing substantial bias against those subgroups at different prediction thresholds. [12] showed these subtle biases in the toxicity datasets might not be captured by prediction-based metrics like EqOdd, which depends on prediction thresholds. So we have prioritized the AUC-based bias metrics (subgroup AUC, BPSN, BNSP) over the other ones to investigate any potential bias.

8.3 Bias Gap between Groups

In this section, we show how much DP affects the gap between bias metrics of a pair of groups from the same attribute (race, gender). Table 16 shows the results in terms of AUC-based bias metrics for a non-private ($e\epsilon \rightarrow \infty$) and a private ($\epsilon \leq 0.5$)model. The results show that in many cases DP significantly widens the gap between bias metrics of different groups. For most other cases the gap changes slightly. And in rare occasions, there is a drop in the gap.

8.4 Predicted Label Distribution.

We found DP has opposite effects on the two datasets about total toxic comments being predicted, as shown in Table 17. In Jigsaw, increasing privacy in the training increases the number of toxic predictions. In the UCBerkeley dataset, the toxic predictions decrease with an increased privacy budget.

It can be attributed to how the dataset is distributed. [13] targeted an even distribution of labeled comments across different hate intensity levels, focused on finding more hate speech examples, whereas in Jigsaw there was no such filtering when creating the dataset. So the model trained on UC Berkeley is more skewed toward hate comments, whereas with Jigsaw it is the opposite. Adding DP introduces both noise and gradient clipping during the training and thus reduces this skewness. Finally, raises the plausibility of predicting opposite examples.

Subgroup		Jigsaw						UCBerkeley					
		Δ BPSN		Δ BNSP		ΔAUC		Δ BPSN		Δ BNSP		Δ AUC	
		$\epsilon \to \infty$	$\epsilon \leq 0.5$	$\epsilon \to \infty$	$\epsilon \leq 0.5$	$\epsilon \to \infty$	$\epsilon \leq 0.5$	$\epsilon \to \infty$	$\epsilon \leq 0.5$	$\epsilon \to \infty$	$\epsilon \leq 0.5$	$\epsilon \to \infty$	$\epsilon \leq 0.5$
Male	Female	0.015	0.036	0.008	0.011	0.003	0.020	0.020	0.012	0.023	0.022	0.009	0.004
Male	Trans.	0.048	0.033	0.016	0.034	**0.098**	0.010	0.069	**0.110**	0.380	0.190	0.056	**0.103**
Female	Trans.	**0.063**	0.003	0.008	0.023	**0.101**	0.030	0.089	**0.122**	0.061	0.041	0.065	**0.099**
White	Black	0.012	**0.066**	0.003	0.042	0.012	0.006	0.046	0.051	0.097	0.012	0.066	**0.085**
White	Asian	0.142	0.163	0.039	0.016	0.089	**0.136**	0.063	0.073	0.052	0.068	0.041	0.028
Black	Asian	0.130	0.097	0.036	0.026	0.077	**0.142**	0.109	0.129	0.149	**0.189**	0.025	0.057

Table 16: Difference in AUC-based bias metrics between groups of the same attribute (race, gender). Cases where the gap between bias changed significantly are in bold.

Budget	**Jigsaw**		**UCBerkeley**	
(ϵ)	True	False	True	False
∞	0.138	0.862	0.227	0.737
9.0	0.155	0.845	0.195	0.805
6.0	0.156	0.844	0.200	0.801
3.0	0.156	0.845	0.183	0.817
1.0	0.159	0.841	0.180	0.820
0.5	0.153	0.847	0.155	0.845
Trend $\epsilon \downarrow$	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$

Table 17: Predicted Label Distribution

8.5 Limitations

We make our observations based on toxicity and hate speech detection tasks. However, bias in NLP has also been investigated in other tasks like coreference resolution [20], sentiment analysis [25], and question answering. Whether trends found in our results persist in those tasks too, is something to be explored for future works. We consider six diverse identity subgroups across two protected attributes (race, and gender). There exist more sensitive attributes in the dataset like religion and sexual orientation which are not explored here but can be explored in future works. We saw similar trends in bias across both selected attributes and the identity subgroups. Even with new attributes or subgroups, the trends should persist similarly.

9 Conclusion

In this work, we explore how differential privacy affects the bias in NLP models. We found DP increases model bias and the impact of that increase varies across different identities. We perform our empirical analysis on two hate/toxic language detection datasets. We evaluated the gender and racial bias of the model using different bias metrics for models trained at different privacy budgets (ϵ). We found that (Table 15) stronger privacy budgets cause the model to have more difficulty distinguishing between the positive/negative examples in the identity subgroup from negative/positive examples in other subgroups at different prediction thresholds [12]. We also observe an increase in equality of odds at a much stricter privacy level, mainly because the recall drops significantly for each group, reducing the difference between them. However, the protected accuracy also drops in most cases. Our overall observations confirm that DP increases bias in the NLP models for hate speech detection, and the researchers need to be aware of this bias when adding privacy to NLP models.

References

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[2] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In Theory of cryptography conference, pages 265–284. Springer, 2006.

[3] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pages 308–318, 2016.

[4] Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, et al. Differentially private fine-tuning of language models. arXiv preprint arXiv:2110.06500, 2021.

[5] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21), pages 2633–2650, 2021.

[6] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization, UMAP'19 Adjunct, page 309–315, New York, NY, USA, 2019. Association for Computing Machinery.

[7] Eugene Bagdasaryan and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. CoRR, abs/1905.12101, 2019.

[8] Amartya Sanyal, Yaxi Hu, and Fanny Yang. How unfair is private learning ? 06 2022.

[9] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 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, Volume 2 (Short Papers), pages 15–20, 2018.

[10] Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 8–14, 2018.

[11] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In Proceedings of the Third Workshop on Abusive Language Online, pages 25–35, Florence, Italy, August 2019. Association for Computational Linguistics.

[12] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In Companion proceedings of the 2019 world wide web conference, pages 491–500, 2019.

[13] Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. arXiv preprint arXiv:2009.10277, 2020.

[14] Cuong Tran, My H. Dinh, and Ferdinando Fioretto. Differentially private deep learning under the fairness lens. CoRR, abs/2106.02674, 2021.

[15] Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. Robin hood and matthew effects–differential privacy has disparate impact on synthetic data. arXiv preprint arXiv:2109.11429, 2021.

[16] Archit Uniyal, Rakshit Naidu, Sasikanth Kotti, Sahib Singh, Patrik Joslin Kenfack, Fatemehsadat Mireshghallah, and Andrew Trask. Dp-sgd vs pate: Which has less disparate impact on model accuracy? arXiv preprint arXiv:2106.12576, 2021.

[17] Lingjuan Lyu, Xuanli He, and Yitong Li. 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, pages 2355–2365, 2020.

[18] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. Advances in neural information processing systems, 29, 2016.

[19] Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. Transactions of the Association for Computational Linguis-

tics,9:1249–1267, 11 2021.

[20] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez,and Kai-Wei Chang. Gender bias in contextualized word embeddings. 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 629–634, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[21] Charan Reddy, Deepak Sharma, Soroush Mehri, Adriana Romero-Soriano, Samira Shabanian, and Sina Honari. Benchmarking bias mitigation algorithms in representation learning through fairness metrics. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.

[22] Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(34):211–407, 2014.

[23] Ashkan Yousefpour, Igor Shilov, Alexandre Sablayrolles, Davide Testuggine, Karthik Prasad, Mani Malek, John Nguyen, Sayan Ghosh, Akash Bharadwaj, Jessica Zhao, et al. Opacus: User friendly differential privacy library in pytorch. arXiv preprint arXiv:2109.12298, 2021.

[24] Rohan Anil, Badih Ghazi, Vineet Gupta, Ravi Kumar, and Pasin Manurangsi. Large-scale differentially private bert. arXiv preprint arXiv:2108.01624, 2021.

[25] Svetlana Kiritchenko and Saif Mohammad. 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, 2018.

[26] Shahbazi, Nima, Mahdi Erfanian, and Abolfazl Asudeh. "Coverage-based Data-centric Approaches for Responsible and Trustworthy AI." Data Engineering (2024): 3.

[27] cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, nithum. (2019). Jigsaw Unintended Bias in Toxicity Classification. Kaggle. https://kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification.

[28] Salimi, Babak, Bill Howe, and Dan Suciu. "Database repair meets algorithmic fairness." ACM SIGMOD Record 49.1 (2020): 34-41.

[29] Lin, Yin, Yifan Guan, Abolfazl Asudeh, and H. V. Jagadish. "Identifying insufficient data coverage in databases with multiple relations." Proceedings of the VLDB Endowment 13, no. 11 (2020)..