Transferring Fairness using Multi-Task Learning without Demographic Information

Anonymous ACL submission

Abstract

Training supervised machine learning systems with a fairness loss can ensure prediction fairness across different demographic groups. However, doing so requires demographic annotations for training data, without which we cannot produce debiased classifiers for most tasks. Drawing inspiration from transfer learning methods, we investigate whether we can utilize demographic data from a related task to improve the fairness of a target task. We adapt a single-task fairness loss to a multi-task setting to exploit demographic labels from a related task in debiasing a target task, and demonstrate that demographic fairness objectives transfer fairness within a multi-task framework. Additionally, we show that this approach enables intersectional fairness by transferring between two datasets with different single-axis demographics. We explore different data domains to show how our loss can improve fairness domains and tasks.

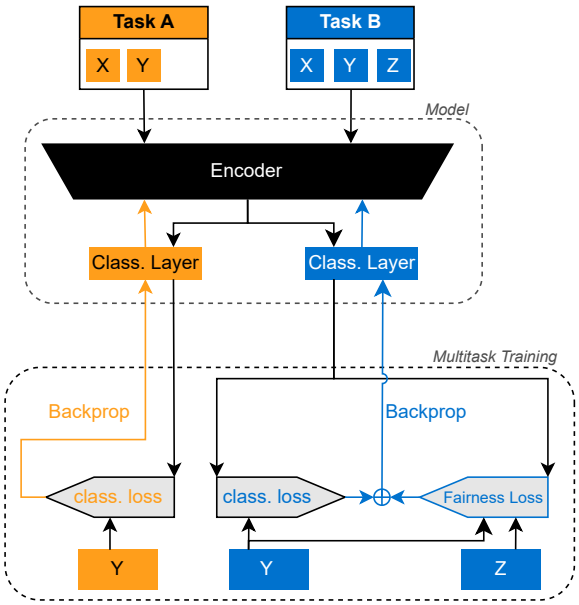

Figure 1: Our approach, *MTL fair*, a multitask method to utilize an auxiliary task (B) to train a fair model for a task (A) without demographic annotations.

1 Introduction

Machine learning models can have disparate performance on specific subpopulations even when they have relatively high performance overall. High overall accuracy can, in fact, mask poor performance for smaller subpopulations. To alleviate disparate performance and biased model behavior, a variety of techniques can make for fairer AI systems, such as additional training objectives to debias models. These training objectives utilize example metadata, such as author demographics of a document, to influence the loss towards fairer model behavior. Unfortunately, these techniques require demographic metadata for training sets which is often unavailable, and thus creating a barrier to training systems that behave fairly.

Transfer learning is a general strategy for learning with limited or no training labels, where annotations from one task are used to produce a model in a related task. Multi-task learning (MTL) utilizes transfer learning between tasks by jointly training a model over several related tasks. We draw inspiration from MTL methods and ask, *can MTL transfer demographic fairness between related tasks?* Suppose we have target labels for two tasks A and B, but demographic labels only for task A; can we transfer fairness learned from task A to task B? We adapt existing MTL and fairness loss methods to achieve the goal of demographic fairness transfer. Figure 1 shows a representation of our method to achieve model fairness given demographic annotations for only one task.

The success of this approach can be adapted to address a limitation in current demographic fairness methods: intersectional fairness. Intersectional fairness means that fairness conditions hold across cross-products of orthogonal attributes and not just within a single attribute. Crenshaw (1989) introduced the term *intersectionality* in the legal

field[1] to describe how anti-discrimination laws failed to protect Black women workers, as employers avoided charges of discrimination by hiring enough Black men and White women to satisfy the single-identity clauses. Similarly, early work in the machine learning field found biases in the performance of vision models at the intersection of gender and skin color (Buolamwini and Gebru, 2018), where facial recognition models performed worse for Black women. Current methods cannot produce intersection fairness unless we have annotations for both attributes on the same instances. This high bar for training data further exacerbates data scarcity since most datasets with demographic attributes only consider single-axis attributes (e.g. race or gender alone.) Therefore, we use our MTL approach to explore how two related tasks, each with different single-axis demographic annotations (i.e. gender *or* race), can produce an intersectionally fair model for both tasks (gender *and* race).

Finally, we explore how the relationship between tasks enables fairness transfer by conducting experiments with different tasks in two domains (clinical and social media) and evaluate the fairness transfer between tasks within and across domains.

We summarize our contributions as follows:

- We transfer fairness across tasks by adapting single-task fairness losses to multi-task settings.
- We enable intersectional fairness by leveraging two tasks with single-axis demographic attributes using a multi-task fairness loss.
- We explore the relationship between task similarity and fairness generalization.

2 Methods

We begin by describing the learning setting shown in Figure 1. Let us assume we desire an unbiased model for task A for which we have input text (X) and associated labels (Y), but no demographic attributes. Instead, we have demographic data for task B, a task related to but distinct from A. Since there exist similarities between tasks A and B, we wish to utilize the demographic attributes (Z) available for task B to obtain a fair classifier for task A. Specifically, by using multi-task training to jointly train a model with both tasks A and B, with an added fairness loss supported by task B alone, we hope to produce a fair model for task A.

[1]The idea can be found in prior sources (Truth, 1851), as described in Costanza-Chock (2020).

Employing a similar idea, we generalize our approach to intersectional fairness. We want to train classifiers for both tasks A and B, which consist of text data and target labels. We have demographic attributes for both A and B, but they are *different* attributes for each task, e.g. task A has gender attributes and task B has race attributes. Since neither task has both attributes, we are unable to utilize an intersectional fairness loss to the tasks individually. Therefore, we propose a multi-task objective to combine attributes from both tasks to obtain intersectional fairness.

This section introduces our fairness definitions and losses, provides formal definitions of our training objectives and describes our training procedure.

2.1 Fairness Loss

We select a fairness definition that supports intersectionality and that is differentiable so that it can be included in model training. We use ϵ-Differential Equalized Odds (ϵ-DEO), a variant of ϵ-DF (Foulds et al., 2020), that applies the equalized odds objective, to ensure that both the recall and specificity rates are equal across demographic groups (Barocas et al., 2019) and intersectional subgroups, and that is learnable and differentiable. Utilizing the equalized odds objective is important– as opposed to others, e.g. *demographic parity*– because it avoids limitations that arise when the labels are correlated with demographic variables, which is the case in many real-world problems and some of the datasets used in our experiments, e.g. the clinical datasets (Hardt et al., 2016). Under ϵ-DEO, perfect fairness would be a score of 0, which would mean that there is no difference in the recall and specificity rates across demographic subgroups. A formal definition is provided in Appendix A.

The standard approach to incorporating fairness metrics into learning objectives uses an additive term. For example, for a deep neural network classifier $M(X)$ with parameters θ, we obtain the *single task* equation in Table 1, where $\epsilon(X;\theta)$ is the ϵ-DEO measure for the classifier, ϵ_t is the desired base fairness (in our experiments 0), and λ is a hyper-parameter that trades between prediction loss and fairness (Foulds et al., 2020). Since the fairness term is differentiable, the model can be trained using stochastic gradient descent on the objective via backpropagation and automatic differentiation. A *burn-in* period and stochastic approximation-based update are adopted following Foulds et al. (2020).

Fairness loss	Objective
single task	$\min_\theta f(X;\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(x_i;\theta) + \lambda[\max(0, \epsilon(X;\theta) - \epsilon_t)]$
MTL	$\min_\theta f(A;B;\theta) \triangleq \frac{1}{\lvert A\rvert\lvert B\rvert}\sum_{i=1}^{\lvert A\rvert}\sum_{j=1}^{\lvert B\rvert}\mathcal{L}(x_{a,i};[\theta_s \cup \theta_a]) + \mathcal{L}(x_{b,i};[\theta_s \cup \theta_b]) + \lambda[\max(0, \epsilon(B;[\theta_s \cup \theta_b]) - \epsilon_t)]$
MTL intersectional	$\min_\theta f(A;B;\theta) \triangleq \frac{1}{\lvert A\rvert\lvert B\rvert}\sum_{i=1}^{\lvert A\rvert}\sum_{j=1}^{\lvert B\rvert}\mathcal{L}(x_{a,i};[\theta_s \cup \theta_a]) + \lambda[\max(0, \epsilon(A;[\theta_s \cup \theta_a]) - \epsilon_t)] + \mathcal{L}(x_{b,i};[\theta_s \cup \theta_b]) + \lambda[\max(0, \epsilon(B;[\theta_s \cup \theta_b]) - \epsilon_t)]$

Table 1: Objectives for adding fairness losses in single task, MTL and MTL intersectional cases.

One optimization challenge that emerges from incorporating fairness is instability due to the representativeness of the mini-batches: a diverse set of examples is needed on which the fairness loss can be meaningfully measured. Following prior work (Foulds et al., 2020), we use a stochastic approximation-based update for $\epsilon(X;\theta)$ by estimating mini-batch noisy expected counts per intersecting demographic group with a hyperparameter ρ, $\tilde{\mathcal{N}}_t = (1-\rho)\tilde{\mathcal{N}}_{t-1} + \rho\mathcal{N}_t$, where $\tilde{\mathcal{N}}_t$ is the approximated count at time t and $\mathcal{N}_t$ is the actual count. Thus ρ controls the smoothness of the approximation of the demographic counts in mini-batches.

2.2 MTL fairness

We train a model jointly on tasks A and B with a fairness loss applied only to task B, as seen in Figure 1 (*MTL fair*.) The MTL training will optimize the shared model parameters (the encoder) to exploit task similarities and improve fairness in task A based on the fairness constraints of task B.

Assume we have a target task A with training instances of input features x_a and task labels y_a, and an auxiliary task B, with training instances of input features x_b, task labels y_b and demographic attributes z_b. Adding the fairness loss with respect to task B in a multi-task objective of a DNN-based classifier $M(X)$ with shared parameters θ_s, task A-specific parameters θ_a and task B-specific parameters θ_b, where $\theta = (\theta_s \cup \theta_a \cup \theta_b)$ becomes *MTL* equation in Table 1, where $\epsilon(B;[\theta_s \cup \theta_b])$ is the ϵ-DEO measure for the classifier on task B. Notably, $\epsilon(B;[\theta_s \cup \theta_b])$ is applied to both task-specific and shared parameters.

2.3 Intersectionality

We formalize the problem of intersectional fairness across tasks using the ϵ-DEO loss across both tasks using MTL training with two fairness losses, one for each task.

Assume we have a target task A, with training instances of input features x_a, task labels y_a, and demographic attributes w_a, and an auxiliary task B with training instances of input features x_b, task labels y_b and demographic attributes w_b. We seek an intersectionally fair classifier on both tasks with respect to $z = w_a \times w_b$. Adding the fairness loss in a multi-task objective of a DNN-based classifier $M(X)$ with shared parameters θ_s, task A-specific parameters θ_a and task B-specific parameters θ_b, where $\theta = (\theta_s \cup \theta_a \cup \theta_b)$ *MTL intersectional* equation in Table 1, where $\epsilon(A;[\theta_s \cup \theta_a])$ and $\epsilon(B;[\theta_s \cup \theta_b])$ are the ϵ-DEO measure for the classifier on task A and B respectively. Notably, both losses update the shared parameters θ_s.

3 Data

Transferring demographic fairness from one task to another can, in principle, be applied to any setting with multiple tasks but where demographic information is available for only one task. However, to evaluate our method we require demographic information for each task's test set, and a dataset with multiple demographic attributes to test intersectional fairness. This makes data selection more challenging. We select datasets in varied domains: clinical text records, online reviews, and social media. A summary of the selected datasets is in Table 2. Appendix C gives a more detailed description of datasets as well as showing in-depth dataset statistics in Table 7.

3.1 Clinical Records

We use the Multiparameter Intelligence Monitoring in Intensive Care (MIMIC-III) dataset (Johnson et al., 2016b,a; Goldberger et al., 2000), a collection of anonymized English medical records that

Data	Task classes	Demog. attributes	Demog. groups
`Clinical notes`			
In-hosp. Mort.	2	gender	2
Phenotyping	28	gender	2
`Online reviews`			
Sentiment	3	gender + age	4
Topic	8	gender + age	4
`Twitter`			
Sentiment	2	race	2
HateXplain	2	race	5

Table 2: Datasets used in our experiments.

include clinical notes drawn from a critical care unit from the Beth Israel Deaconess Medical Center between 2001 and 2012. We select two tasks from those defined by Zhang et al. (2020):

In-hospital Mortality. The task is to predict whether a patient will die in the hospital based on the textual content of all the clinical notes created within the first 48 hours of the hospital stay.

Phenotyping.[2] The task of assigning medical conditions based on the evidence in the clinical record. In our task, we will assign up to 25 acute or chronic conditions from the HCUP CCS code groups (Harutyunyan et al., 2019), labeled with ICD-9 codes, and three extra summary-labels: any, chronic, or acute condition. Therefore, the task is modeled as a set of 28 binary classification tasks, and evaluated as a multi-label problem. We use the same pre-processing pipeline and train-dev-test splits as Zhang et al. (2020).[3]

3.2 Online Reviews

We use the Trustpilot data of Hovy (2015), who provide data from an open review platform that allows users to review a range of products, stores, and services. Each instance is an English language review and a 5-point rating. For our experiments, we utilize the `sentiment` (100k reviews) and `topic` (24k reviews) tasks which share demographics for age – under 35 (U35) and over 45 (O45) years old – and gender – men and women.

Reviews sentiment. Labels assigned based on the stars of the reviews and selected reviews that have both age and gender labels available.

Reviews topic. Labels assigned based on the general topic of the review, e.g. fashion, fitness, etc. using the Trustpilot taxonomy for seller companies and selected the top 5 most popular topics: Fitness & Nutrition (*Fitness*), Fashion Accessories (*Fashion*), Gaming (*Gaming*), Cell phone accessories (*Cell Phone*) and Hotels (*Hotels*)), following Hovy (2015). We perform the same demographic selection criteria as the *sentiment* task. We obtain randomly stratified train-dev-test (60-20-20%) splits ensuring equal representations for both gender and age groups.

3.3 Social Media

Twitter sentiment. We use the twitter sentiment classification task introduced by Elazar and Goldberg (2018). Labels were assigned based on common emojis and demographic variables are based on the dialectal corpus from Blodgett et al. (2016), were race was assigned based on geolocation and words used in the tweet, obtaining a binary AAE (African-American English) and SAE (Standard American English) which we use as proxies for non-Hispanic African-Americans and non-Hispanic Caucasians.

HateXplain. This is a hate speech classification dataset obtained from a combination of Twitter and Gab posts (Mathew et al., 2021). We use the binary version of the task which classifies for toxicity of posts. We select the posts for which there is a majority agreement of annotators for race target groups, and for which we have representation across train-dev-test splits.

For each dataset, we follow the splits provided by Elazar and Goldberg (2018) and Mathew et al. (2021), respectively.

4 Experiments

This section describes baselines and model training. Table 8 in Appendix D shows all combinations of models, training datasets, and fairness attributes.

4.1 Models

We implement our fairness objectives in an MTL setting based on a shared language encoder and task-specific classification heads. We use BERT-style encoders (Devlin et al., 2019) with a domain-specific vocabulary: SciBERT for clinical tasks, pretrained on scientific text (Beltagy et al., 2019), following prior work (Zhang et al., 2020; Amir et al., 2021),[4] RoBERTa for the online reviews tasks (Liu et al., 2019) initialized with the `roberta-base` checkpoint,[5] and BERTweet for the

[2] In a medical record, a phenotype is a clinical condition or characteristic.

[3] https://github.com/MLforHealth/HurtfulWords

[4] https://huggingface.co/allenai/scibert_scivocab_uncased

[5] https://huggingface.co/roberta-base

social media tasks (Nguyen et al., 2020), initialized with the `vinai/bertweet-base` checkpoint.[6] We add a separate linear classification head for each task, with a Softmax output function to allow for multi-class classification or a Sigmoid output function for binary and multi-label classification. The document representation for the classification head is a mean-pooled aggregation across all subword representations of the document taken at the top layer of the network. The training objective is an additive combination of the loss for each of the individual tasks. Models were trained on Nvidia A100 GPUs, using `jiant` (Phang et al., 2020), a multi-task wrapper library.

Fairness methods require a careful tradeoff between the task loss and fairness loss (Islam et al., 2021). To obtain the best performing model, we use a grid search for each task, with a learning rate= $[1e^{-4}, 1e^{-5}, 1e^{-6}]$ with Adam optimizer (Kingma and Ba, 2014), and batch size= $[16, 32, 48]$. We select the best performing model on development data and report test data results.

4.2 Baselines

We establish baselines against which to compare our MTL fairness transfer method.

STL-base. We train a single-task model for each task, i.e. a fine-tuned encoder and classification layer. These models do not include a fairness loss since they represent the classifiers obtained when no demographic attributes are available. We named these models single task learning base (`STL-base`), and they serve as an upper bound in task performance when fairness is not a goal.

STL-fair. Finetuning models without fairness losses can result in unfair classifiers (Lan and Huan, 2017; Zhang et al., 2020), which is known as *no fairness through unawareness* (Barocas et al., 2019). To determine how well we could do with full demographic information, we train single-task models with both a task loss and fairness loss §4.2. For the models trained on the clinical dataset and Twitter datasets, we add a single-attribute fairness loss, with gender and race groups respectively. For the models trained on the online reviews datasets (sentiment and topic), we add an intersectional fairness loss, with age and gender attributes. This allows us to test both single-attribute and intersectional fairness. We call these single task models with fairness objectives `STL-fair`. We performed a grid search on each task, with the same search spaces as before, in addition to the fair-related hyperparameters $\lambda = [.01, .05, .1]$, $\rho = [.01, .1, .9]$, and *burn-in*= $[.5, 1]$ epochs, defined in §2.1.

MTL-base. We next evaluate models trained in a multi-task setting. While MTL can lead to better performance, it often leads to worse results compared to single-task baselines due to task conflict and other optimization challenges (Weller et al., 2022; Gottumukkala et al., 2020). A *dynamic scheduler*, which changes the rate that a task is seen based on the current relative performance, has been shown to improve performance in traditional MTL setups (Gottumukkala et al., 2020). Therefore, we first train MTL models with a dynamic scheduler on mutually related task pairs to avoid a domain mismatch: *In-hospital Mortality* &*Phenotyping* (clinical setting), *reviews sentiment* & *reviews topic* (online reviews domain), and *Twitter sentiment* & *HateXplain* (social media setting). We name these models multi-task baselines `MTL-base`.

BLIND. We also compare our work with other bias removal methods that do not require demographic attributes. Orgad and Belinkov (2023) propose that often classifiers make predictable mistakes when implicit demographic features are used as shorcut features, a bias also known as *simplicity* bias (Bell and Sagun, 2023). *BLIND* trains a success classifier that takes the encoder features and predicts the success of the model on the task. A correct prediction by the success classifier means the model used a shallow, or simple, decision and the sample is down-weighted. We use their implementation of the algorithm[7] and perform a hyperparameter search, $\gamma = [1, 2, 4, 8, 16]$, temp= $[1, 2, 4, 8, 16]$, as suggested by authors (Orgad and Belinkov, 2023). BLIND does not support multi-label tasks so we do not report results for the clinical tasks.

4.3 Our Methods

We propose variations on multi-task learning with a fairness loss in support of our proposed setup.

MTL-fair. We evaluate the fairness loss applied to one of the two tasks for each in-domain task pair: clinical, online reviews, and social media domains. We call these models with an MTL objective and a fairness loss `MTL-fair`. To report a fair comparison, each of the `MTL-fair` models is compared with the task for which no fairness loss

[6] https://huggingface.co/vinai/bertweet-base

[7] code: https://github.com/technion-cs-nlp/BLIND

was added, e.g. for the *In-hospital Mortality* task, we compare the `STL-base` and `STL-fair` trained on *In-hospital Mortality* data only, the `MTL-base` trained on *In-hospital Mortality* and *Phenotyping* (without fairness loss), and the `MTL-fair` trained on *In-hospital Mortality* and *Phenotyping*, with a fairness loss applied to the *Phenotyping* task only. We performed a grid search with the same base search space as in §4.2

MTL-inter. To train intersectionally fair models on two tasks for which we have only a single axis of demographic attributes, we use an MTL objective with two different single-axis fairness losses. We focus on the online reviews datasets, for which we have sufficient demographic data to support this experiment.[8] We call these models that use MTL with intersectionally fair losses `MTL-inter`.

4.4 Evaluation

We utilize established evaluation metrics for all datasets. The clinical datasets are evaluated at the patient level. We use the aggregation function from Zhang et al. (2020) since clinical notes are too long to fit in the context window of models; see §C for more details. We report macro-averaged F1 scores for task performance and ϵ-DEO for fairness metrics. The best model criteria for `STL-base`, `MTL-base` and `BLIND` models is their F1 score on the validation set. We choose `STL-fair`, `MTL-fair` & `MTL-inter` models with the lowest ϵ-DEO and at least 95% performance of the `STL-base` models in the validation set.

So far, it has been assumed that there is an extra dataset that has access to demographic attributes within the same domain. However, due to the scarcity of NLP datasets with access to demographics, it may not be possible to find an eligible dataset within the same domain. To evaluate the robustness of our method, we test the impact of domain mismatch and task similarity on the MTL models with fairness loss. We focus on the *Twitter sentiment* task, as it allows us to pair it with a task within the same domain (*HateXplain*), a similar task but in a different domain (*reviews sentiment*) and other tasks with varied domains and task similarities.

5 Results & Analysis

Table 3 reports performance and fairness scores for within-domain `MTL-fair` experiments. Our baselines perform comparably with prior work (Zhang et al., 2020; Hovy, 2015; Mathew et al., 2021; Elazar and Goldberg, 2018) so we can evaluate the use of multi-task learning methods to debias algorithms with high-performing models. In contrast to the common perception that we must trade off fairness and performance, we observe that the performance of `STL-fair` models is equal to or better in 3/4 tasks compared to the `STL-base` model baselines and produces fairer models based on ϵ-DEO. This confirms recent work suggesting that an extensive grid search of hyperparameters avoids the fairness vs. performance trade-off (Islam et al., 2021).

	Clinical			
	In-hosp Mort.		Phenotyping	
	F1 (%) ↑	ϵ-DEO ↓	F1 (%) ↑	ϵ-DEO ↓
`STL-base`	62.1	0.25	**53.6**	0.28
`STL-fair`	65.1	0.22	52.9	0.26
`MTL-base`	**65.6**	**0.17**	53.3	0.27
`MTL-fair`	64.0	0.19	53.0	**0.21**
	Twitter			
	HateXplain		Sentiment	
	F1 (%) ↑	ϵ-DEO ↓	F1 (%) ↑	ϵ-DEO ↓
`BLIND`	70.4	1.15	**77.6**	0.30
`STL-base`	71.3	1.58	76.4	0.33
`STL-fair`	**71.5**	1.63	76.5	0.28
`MTL-base`	69.9	1.45	76.2	0.37
`MTL-fair`	70.4	**0.80**	75.5	**0.28**

Table 3: Scores of the MTL fairness loss (`MTL-fair`) within-domain experiments. Best per task is **bold**.

Multi-task fairness generalizes to tasks without demographics. We expected the `STL-fair` models to be an upper bound for fairness, and `STL-base` an upper bound for performance compared to the `MTL-fair` models. However, for 3/4 tasks, the `MTL-fair` models are fairer than the `STL-fair` counterparts! In these cases, the performance of the `MTL-fair` models is slightly worse than `STL-fair` models but still comparable to `STL-base`, obtaining models that are fairer while maintaining model performance. This suggests that just as multi-task learning finds representations that are useful for training multiple tasks, multi-task fairness learning corrects model representations to be fairer for both tasks – sometimes finding a fairness minimum that is fairer than it would with access to target task demographic attributes. This technique may be yielding more generalizable and fair representations. Comparing to `BLIND`,

[8] MIMIC has demographic data but is highly skewed, resulting in intersection groups with only a handful of individuals.

we observe that BLIND yields fairer models than STL-base but less fair than STL-fair and our method MTL-fair. This suggests that when we have no demographic attributes, BLIND is better than not attempting fairness, but effectively using demographics, whether internally or in another task, increases the fairness of the models. In all settings, the multi-task fairness loss produced a model that is fairer than the single-task baseline without demographic attributes and with comparable performance.

Multi-task enables intersectional fairness. Table 4 shows the results for the intersectional fairness experiments. The best MTL-inter model performs comparably to the STL-base and is fairer compared to the STL-fair models in both tasks. We obtain an intersectionally fairer model compared to the baselines when only one demographic attribute is available per task. This suggests that the single-attribute fairness losses combine to obtain model representations that are beneficial to the fairness of both protected attributes and their intersectional groups. Compared to prior work, we see fairness benefits when utilizing single-axis demographics, perhaps due to greater loss stability and the ability of MTL setups to integrate all the losses.

Multi-task fairness generalizes across domains and tasks. So far we have assumed access to a task with demographic attributes available within the same domain, exploiting text similarities between the tasks to generalize the fairness across tasks. However, given the scarcity of datasets with demographic attributes, we may wonder whether domain similarity is necessary to transfer fairness. In Table 5 we show the results of the single-task *Twitter sentiment* models as well as applying the MTL fair loss across different datasets. We observe that adding a fairness loss to the MTL settings helps in fairness with tasks across domains and task similarities, except for the clinical *Phenotyping* task. This may be because the performance of the *Phenotyping* task in the MTL system was poor (possibly because of task incompatibility) and the fairness loss might not have actually provided any meaningful change to the model. Regardless, on tasks where we obtain competitive performance for both tasks, the fairness loss was able to generalize fairness, obtaining models that are fairer than the single-task baselines and sometimes fairer than applying a fairness loss to the target task, showing evidence that our method is robust across domains, demographic attributes, and task similarities.

Why does the multi-task fairness loss work? The results in this section suggest that the multi-task fairness loss produces more generalizable and fairer representations. We hypothesize that the combination of (A) the regularizing effect of the fairness loss, as suggested by prior work (Islam et al., 2021), (B) shared parameters across tasks and (C) the simultaneous learning of both tasks allows for positive fairness transfer. First, we note that multi-task learning alone (B & C, MTL-base) or a fairness loss (A, STL-fair) may suffer in performance or fairness (or sometimes both) compared to our method. Further, one could have shared parameters, B, but not train simultaneously by finetuning on individual tasks consecutively rather than simultaneously, a multi-task method also known as STILT (Weller et al., 2022; Phang et al., 2018). In Appendix B we show that when the fairness loss is applied consecutively, rather than simultaneously, the fairness transfer effect is no longer observed. Thus, the MTL objective plus the shared parameters are instrumental in enabling the positive transfer of the fairness loss from one task to another.

6 Related Work

Machine learning methods that seek to transfer fairness to unseen tasks have recently received a lot of focus, some utilizing external datasets to ensure fairness on a target task via MTL (Oneto et al., 2020) and domain-shift transfer methods (Chen et al., 2022; Schrouff et al., 2022b); however, they often rely on strong assumptions of distribution shifts, limiting their impact with real-world applications (Schrouff et al., 2022a) or applicability to NLP methods. In comparison, while our method does not include explicit domain-shift assumptions, it relies on some domain similarities that are well studied for general multi-task setups (Weller et al., 2022). Another solution to debias models is to use proxy variables or inferred demographics in settings where we lack demographic data. However, these methods are dependent on the accuracy of the demographic inference model (Aguirre et al., 2021; Bharti et al., 2023) or the availability of proxy variables, e.g. names (Romanov et al., 2019).

Particularly, within the field of Natural Language Processing (NLP), MTL has become the standard training setting through the use of Large Language Models (LLM) (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020). Unfortunately, studies

	Reviews sentiment						Reviews topic					
			F1 (%) per sub-group ↑						F1 (%) per sub-group ↑			
	F1 (%) ↑	ϵ-DEO ↓	F-U35	F-O45	M-U35	M-O45	F1 (%) ↑	ϵ-DEO ↓	F-U35	F-O45	M-U35	M-O45
BLIND	84.3	1.16	82.7	**85.7**	84.4	83.8	92.0	1.05	**91.7**	86.7	89.7	**89.9**
STL-base	84.5	0.95	**87.1**	83.9	83.1	84.6	91.9	1.42	90.0	85.7	**90.3**	88.5
STL-fair	**85.6**	0.77	86.4	84.8	**84.6**	**86.3**	**92.1**	1.04	90.9	**88.7**	90.2	88.1
MTL-base	84.4	0.89	86.1	84.6	82.9	84.7	91.6	1.52	91.4	85.9	89.4	89.5
MTL-fair	83.6	0.65	85.5	82.7	82.8	83.7	91.2	0.86	90.9	88.3	88.1	89.1
MTL-inter	84.1	**0.58**	86.0	83.7	82.4	84.7	91.6	**0.82**	90.6	86.6	89.4	88.9

Table 4: Scores of the intersectional experiments on the reviews datasets (MTL-inter). Best per task is **bold**.

Method		F1 (%) ↑	ϵ-DEO ↓
BLIND		**77.6**	0.30
STL-base		76.4	0.33
STL-fair		76.5	0.28
MTL-fair:	HateXplain	75.5	0.28
	review sentiment	76.3	**0.23**
	review topic	75.7	0.23
	In-Hosp Mort.	75.8	0.25
	Phenotyping	75.2	0.32

Table 5: Scores of MTL-fair for the Twitter sentiment task paired with different domain and task annotations: same domain, same task, and neither. **Bold** is best.

have found that fine-tuning LLMs often results in unfair models, even when starting from a debiased pre-trained encoder (Lan and Huan, 2017; Zhang et al., 2020). Instead, they conclude that fairness requires applying debiasing methods in fine-tuning for the task of interest, requiring demographic information for each task.

In our work we use a separation-based group-wise definition of fairness, *equalized odds* (Hardt et al., 2016), that was adapted to be differentiable and applied to training procedures inspired by the ϵ-Differential Fairness from Foulds et al. (2020). However, there are many other group-wise definitions of fairness that may be adapted in a similar way for other tasks, e.g. *equalized opportunity* (Hardt et al., 2016), which ensures equal true positive rates (recall) across demographic subgroups. There is also *adversarial* fairness loss, where an adversary is added in the training procedure to predict the demographic attributes from the output of the task classifier. This loss also achieves independence of predictions and demographic attributes, similar to demographic parity, and has found success in similar setups from prior work (Islam et al., 2021; Zhang et al., 2020). Our methods can be easily used with any of these demographic losses in the procedure.

7 Conclusion

We explored whether MTL methods for NLP tasks can transfer demographic fairness from one task to another. To achieve this, we adapted single-task fairness losses to multi-task settings to transfer fairness across tasks. We tested our method in multiple NLP datasets in different domains: clinical notes (Johnson et al., 2016b,a; Goldberger et al., 2000), online reviews (Hovy, 2015) and social media (Mathew et al., 2021; Elazar and Goldberg, 2018). We found that while MTL alone and other consecutive variations of MTL (e.g. STILTS) do not help in fairness and may hurt performance, MTL methods with our fairness loss are able to debias models using the demographic attributes from a secondary task, opening up the possibility for producing fair models for a wide range of tasks that lack demographic data. This finding also informs future work on MTL, suggesting adding regularizers, e.g. fairness losses, can help in performance deficits found in prior work (Weller et al., 2022; Gottumukkala et al., 2020).

Additionally, we showed that MTL methods can debias models for intersectional fairness by leveraging two tasks, each with different demographic attributes, to learn a model that achieves intersectional fairness on both tasks. This finding opens up the integration of intersectional fairness losses to new applications and settings that were previously restricted by limited access to demographic attributes. Finally, we test the ability of the MTL fairness loss to generalize fairness across domains and tasks, we find that the transfer of fairness is not dependent on domain or task similarity, but rather related to the performance of the secondary task. Our methods increase the range of tasks that fairness methods can be applied to in the machine learning and NLP community, by allowing the use of external tasks that have demographic attributes to obtain fairer models.

8 Limitations

Our results suggest that our MTL methods are able to utilize external demographic attributes to achieve better fairness for our target task. However, the selection criteria for the best-performing models require access to demographic attributes for the test set to assess the fairness of the models. A solution to this would be to select the models that are the best performing for our target task with the lowest fairness score for the task that we do have demographic data available. This selection criteria, however, does not guarantee the most optimal model, especially if the demographic attribute distributions or the task domains are different. Our recommendation is to validate the fairness of the models with access to demographic attributes when possible.

9 Ethics Statement

We address intersectionality as intersectional group fairness in the methods and analysis when possible given the data availability, as they enable a practical approach for inquiry of these models. We acknowledge that there are real interlocking systems of power that contribute to causing these disparities in society, and that our dataset capture these. For example, we evaluate models on the clinical domain using the MIMIC-III dataset: the healthcare system has been historically biased against people in groups in many protected attribute axis e.g. socio-economic status, race/ethnicity, gender, and age. The goal of our approach is to address these biases in machine learning models so they are less likely to exacerbate the real-life biases as they are integrated in society.

References

Carlos Aguirre, Keith Harrigian, and Mark Dredze. 2021. Gender and racial fairness in depression research using social media. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2932–2949, Online. Association for Computational Linguistics.

Silvio Amir, Jan-Willem van de Meent, and Byron Wallace. 2021. On the impact of random seeds on the fairness of clinical classifiers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3808–3823, Online. Association for Computational Linguistics.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

Samuel James Bell and Levent Sagun. 2023. Simplicity bias leads to amplified performance disparities. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 355–369.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Beepul Bharti, Paul Yi, and Jeremias Sulam. 2023. Estimating and controlling for equalized odds via sensitive attribute predictors. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. 2022. Fairness transferability subject to bounded distribution shift. In *Advances in Neural Information Processing Systems*.

Sasha Costanza-Chock. 2020. *Design justice: Community-led practices to build the worlds we need*. The MIT Press.

Kimberle Crenshaw. 1989. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In *The University of Chicago Legal Forum*, volume 140.

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

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

James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. 2020. An intersectional definition of fairness. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 1918–1921. IEEE.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Ananth Gottumukkala, Dheeru Dua, Sameer Singh, and Matt Gardner. 2020. Dynamic sampling strategies for multi-task reading comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 920–924, Online. Association for Computational Linguistics.

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

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. 2019. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):96.

Dirk Hovy. 2015. Demographic factors improve classification performance. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 752–762, Beijing, China. Association for Computational Linguistics.

Chia-Chien Hung, Anne Lauscher, Dirk Hovy, Simone Paolo Ponzetto, and Goran Glavaš. 2023. Can demographic factors improve text classification? revisiting demographic adaptation in the age of transformers. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1565–1580, Dubrovnik, Croatia. Association for Computational Linguistics.

Rashidul Islam, Shimei Pan, and James R Foulds. 2021. Can we obtain fairness for free? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 586–596.

Alistair Johnson, Tom Pollard, and Roger Mark. 2016a. Mimic-iii clinical database (version 1.4). *PhysioNet*, 10(C2XW26):2.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016b. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Chao Lan and Jun Huan. 2017. Discriminatory transfer. *ArXiv*, abs/1707.00780.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, Online. Association for Computational Linguistics.

Luca Oneto, Michele Donini, Massimiliano Pontil, and Andreas Maurer. 2020. Learning fair and transferable representations with theoretical guarantees. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 30–39. IEEE.

Hadas Orgad and Yonatan Belinkov. 2023. Blind: Bias removal with no demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8801–8821.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Jason Phang, Phil Yeres, Jesse Swanson, Haokun Liu, Ian F. Tenney, Phu Mon Htut, Clara Vania, Alex Wang, and Samuel R. Bowman. 2020. `jiant` 2.0: A software toolkit for research on general-purpose text understanding models. `http://jiant.info/`.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

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

Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. 2022a. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *arXiv preprint arXiv:2202.01034*.

Jessica Schrouff, Natalie Harris, Oluwasanmi O Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Chrsitina Chen, et al. 2022b. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. In *Advances in Neural Information Processing Systems*.

Sojourner Truth. 1851. Ain't i a woman. *December*, 18:1851.

Orion Weller, Kevin Seppi, and Matt Gardner. 2022. When to use multi-task learning vs intermediate fine-tuning for pre-trained encoder transfer learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 272–282, Dublin, Ireland. Association for Computational Linguistics.

Haoran Zhang, Amy X Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, pages 110–120.

A Fairness Definition

ϵ-Differential Fairness is a demographic-parity based metric, which requires that the demographic attributes are *independent* of the classifier output (Barocas et al., 2019; Foulds et al., 2020). Formally, we assume a finite dataset of size N, with each sample consisting of three attributes: features x (in our datasets these are text sequences), task labels y, and demographic attributes z. Let $s_1, ..., s_p$ be discrete-valued demographic attributes, $z = s_1 \times s_2 \times ... \times s_p$. A model $M(X)$ satisfies ϵ-DF with respect to z if for all x, and $\hat{y} \in Range(M)$,

$$e^{-\epsilon} \leq \frac{Pr(M(x) = \hat{y}|\zeta_i)}{Pr(M(x) = \hat{y}|\zeta_j)} \leq e^{\epsilon},$$

for all $(\zeta_i, \zeta_j) \in z \times z$ where $Pr(\zeta_i) > 0$, $Pr(\zeta_j) > 0$. Smaller ϵ is better with $\epsilon = 0$ meaning perfect fairness (Foulds et al., 2020). Perfect fairness under this definition means that the rates of predicted labels are the same across demographic groups, achieving independence between demographic attributes and predictions.

In short, ϵ-Differential Fairness is an independence-based metric that measures the biggest difference in prediction rates between intersections of demographic attributes. However, independence based fairness definitions, like demographic parity and ϵ-DF, have limitations in settings where the prevalence of the target labels is somehow related to the demographic attributes, e.g. breast cancer is much more common in women than men. In these settings, independence based definitions would require model predictions to be independent of the demographic attributes, which would encourage lower performance on the desired task, e.g. either an increase in the prediction of breast cancer for men and/or a decrease in breast cancer for women which are both not ideal. For these reasons, we favor a separation based metric, like *equalized odds*, that avoids limitations associated with dependence of model predictions on demographics by requiring independence conditioned on the target variable (Hardt et al., 2016), i.e. that both recall and specificity rates are equal across demographic groups.

We apply *equalized odds* on the ϵ-DF framework to obtain a metric that is also differentiable, and call it ϵ-Differential Equalized Odds (ϵ-DEO). Formally, let $s_1, ..., s_p$ be discrete-valued demographic attributes, and $z = s_1 \times s_2 \times ... \times s_p$ the intersectional groups. A model $M(X)$ satisfies ϵ-DEO with respect to z if for all x, $\hat{y} \in \text{Range}(M)$ and $y \in \text{Range}(M)$,

$$e^{-\epsilon} \leq \frac{Pr(M(x) = \hat{y}|\zeta_i, y)}{Pr(M(x) = \hat{y}|\zeta_j, y)} \leq e^{\epsilon}, \qquad (1)$$

for all $(\zeta_i, \zeta_j) \in z \times z$ where $Pr(\zeta_i) > 0$, $Pr(\zeta_j) > 0$; smaller ϵ is better, with $\epsilon = 0$ for perfect fairness. Perfect fairness results from a classifier with the same recall and specificity rates across intersectional groups of demographic attributes.

B STILT and frozen experiments

In this section we test the hypothesis of whether it is important to have shared parameters and simultaneous learning when implementing the multi-task fairness loss.

MTL. We label `MTL` the models that were trained simultaneously, as described in §2.2.

STILT. We label `STILT` the models that were trained consecutively. First, the model is finetuned only for task B with the fairness loss, the task with demographic attributes as seen in Figure 1. This step results in a model similar to `STL-fair` for task B. Second, the model is further finetuned for task A (as seen in Figure 1), with a different classification layer and without a fairness loss. Both steps together result in a model that has been trained with the same data and the same number of parameters as `MTL-fair`, however the tasks are not trained simultaneously.

Frozen. In order to test the importance of parameter sharing, we train a variance of the model where the shared parameters, BERT-based encoder, are frozen during training. In this way, the number of shared parameters, θ_s in Table 1, is empty. First, we train a single-task model with a fairness loss where the encoder is frozen, we label this `STL-fair-frozen`. We also train a STILT model, where we first finetune for the task that has demographic attributes (Task B) with a fairness loss end-to-end, and then we finetune for the task without demographic attributes without a fairness loss and with the encoder frozen. The idea is that the fairness loss will influence the encoder towards a fairer minima that then the classification loss for the second task will be able to exploit.

Table 6 shows the results for `STILT-fair`, and the frozen versions `STL-fair-frozen` and `STILT-fair-frozen`. First we see that the frozen versions of the models drastically underperform compared to the end-to-end models (ΔF1 $\approx$ 10.)

	F1 (%) ↑	ϵ-DEO ↓
STL-base	71.3	1.58
BLIND	70.4	1.15
STL-fair	**71.5**	1.63
-frozen	61.8	0.69
STILT-fair	70.4	1.42
-frozen	63.4	**0.60**
MTL-fair	70.4	0.80

Table 6: Scores for the STILT and frozen version of the model on HateXplain dataset.

while also being more fair. This is a clear example of the accuracy-fairness trade-off, which is expected given the drastically smaller amount of parameters available for training for these frozen models. It is clear that these models are fairer because they perform equally worse for all demographic groups.

When comparing the STILT-fair to our method MTL-fair, we see that while the performance of the models is very similar (both scoring 70.4 F1), the fairness is drastically better in the simultaneous training (MTL-fair ϵ-DEO=.80) vs. consecutively (STILT-fair ϵ-DEO=1.42). This suggests that the MTL objective, which allows for both tasks to influence the learning, is instrumental for the fairness loss on task B to transfer to task A.

C Data Details

In this section, we report dataset statistics, including the number of posts per label and demographic. We select datasets in varied domains: clinical text records, online reviews, and social media, with both single and intersectional demographic attributes, gender, race and gender+age subgroups, and in a variety of classification paradigms: multiclass, binary and multilabel. Table 7 shows the total and percentage for all datasets.

C.1 Clinical Records

It is crucial to implement behavioral fairness measures to secure fair behavior in the critical context of AI applications for medical records. We use the Multiparameter Intelligence Monitoring in Intensive Care (MIMIC-III) dataset (Johnson et al., 2016b,a; Goldberger et al., 2000), a collection of anonymized English medical records that include clinical notes drawn from a critical care unit from the Beth Israel Deaconess Medical Center between

	train	val	test
In-Hosp Mort.	13191	2701	2445
Men	55.4	54.8	55.2
Women	44.6	45.2	44.8
Positive	13.1	13.8	11.5
Phenotyping	13839	2850	2519
Men	57.2	55.8	56.4
Women	42.8	44.2	43.6
Upper Resp.	2.6	2.5	2.6
Lower Resp.	3.5	4.0	3.7
Shock	3.8	3.6	4.2
Any Acute	70.8	69.9	70.6
Any Chronic	77.1	78.5	76.8
Any Disease	89.6	90.6	90.1
reviews sentiment	58259	19420	19420
Men Under 35	23.2	23.2	23.2
Men Over 45	34.7	34.7	34.7
Women Under 35	14.8	14.8	14.7
Women Over 45	27.3	27.3	27.3
positive	84.5	84.5	84.5
neutral	3.5	3.5	3.5
negative	12.0	12.0	12.0
reviews topic	14744	4915	4915
Men Under 35	54.0	54.0	54.0
Men Over 45	14.2	14.2	14.3
Women Under 35	21.1	21.1	21.1
Women Over 45	10.7	10.7	10.6
Fitness	39.6	39.5	39.6
Fashion	16.6	16.6	16.7
Gaming	16.0	16.0	16.0
Cell Phone	14.4	14.4	14.4
Hotels	13.4	13.4	13.4
HateXplain	5376	661	681
African	54.5	54.0	55.1
Arab	18.8	18.8	17.8
Asian	6.2	6.2	6.5
Hispanic	5.4	5.1	5.1
Caucasian	15.1	15.9	15.6
Toxic	81.3	81.2	79.7
twitter sentiment	156000	4000	8000
African American	50.0	50.0	50.0
Caucasian	50.0	50.0	50.0
Happy	50.0	50.0	50.0
Sad	50.0	50.0	50.0

Table 7: Total (first line) and percentage of documents in the splits all the datasets, separated by demographics and then task labels.

2001 and 2012. We select two tasks from those defined by Zhang et al. (2020): in-hospital mortality and phenotyping. We use the same pre-processing pipeline as Zhang et al. (2020)[9] and only use gender demographics since the other attributes are highly imbalanced, resulting in very small subgroups, as noted by prior work (Amir et al., 2021). These tasks should be evaluated at the patient level (Zhang et al., 2020), however, because the clinical notes are too long to fit in the input size of the encoder, we created subsequences using sliding windows. The model predicts a label for each subsequence and at evaluation time we aggregate these predictions to obtain a single prediction for each patient. We use an aggregation function from prior work (Zhang et al., 2020):

$$Pr(y = 1|\hat{Y}) = \frac{\max(\hat{Y}) + mean(\hat{Y})n/c}{1 + n/c},$$

where $\hat{Y}$ are the predictions for all the subsequences from a patient, n is the number of subsequences and c is a scaling factor ($c = 2$ (Zhang et al., 2020).)

In-hospital Mortality. The task of in-hospital mortality is to predict whether a patient will die in the hospital based on the textual content of all the clinical notes created within the first 48 hours of the hospital stay. To avoid low information notes, we limit the notes to "nurse", "nursing/other" and "physician" types. We concatenate all notes available within the specified time period and tokenize the concatenated notes and split them into sliding subsequences of 512 subwords, to fit within the BERT context window (Devlin et al., 2019). We limit the number of subsequences per patient by selecting the last 30 subsequences of the concatenated notes, following Zhang et al. (2020).

Phenotyping. In a medical record, a phenotype is a clinical condition or characteristic. Phenotyping is the task of assigning these conditions based on the evidence in the medical record. In our task, we will assign up to 25 acute or chronic conditions from the HCUP CCS code groups (Harutyunyan et al., 2019), labeled with ICD-9 codes. In addition to those conditions, three summary labels are also added for patients that have any chronic or acute condition. Therefore, the task is modeled as a set of 28 binary classification tasks, and evaluated as a multi-label problem. For this task we select the first note written by a "nurse", "nursing/other" or "physician" within the first 48 hours of the stay, as proposed by Zhang et al. (2020).

For each dataset, we use the train-dev-test splits provided by Zhang et al. (2020). Table 7 shows the final breakdown of the number of subsequences in the datasets.

[9] https://github.com/MLforHealth/HurtfulWords

C.2 Online Reviews

Developing automated NLP methods for online product reviews can help companies understand customer feedback, improve the user experience, and enable market analysis. There are a variety of tasks defined for online reviews, such as sentiment analysis, determining the helpfulness of a review, and the topic of the review. Furthermore, reviews are authored by a diverse population and we seek models that perform fairly across this user population.

We use data from Trustpilot, an open review platform that allows users to review a range of products, stores, and services (Hovy, 2015). Each instance is an English language review selected from the Trustpilot website that consists of a text review and a 5-point star rating, along with item information, such as the seller. The original dataset defined three tasks: `sentiment` (based on the rating of the review), `topic` (the subject of the review), and `attributes` (demographic attributes of the review author). For our experiments, we utilize the `sentiment` (100k reviews) and `topic` (24k reviews) tasks which share demographics for age – under 35 (U35) and over 45 (O45) years old – and gender – men and women.

Reviews sentiment. This is a multiclass task where the labels were assigned based on the stars of the reviews: 1-star reviews were labeled as "negative", 3-star labeled as "neutral" and 5-star labeled as "positive". We selected reviews that have both age and gender labels available with age ranges between 16-35 and 45-70 years old, and discarded reviews with 2 and 4 stars.

Reviews topic. This is a multiclass task where labels are assigned based on the general topic of the review, e.g. fashion, fitness, etc. These concepts were assigned to each review using the Trustpilot taxonomy for seller companies, which summarizes the services and products offered by each company in the corpus with high-level concepts. We selected the top 5 most popular topics: Fitness & Nutrition (*Fitness*), Fashion Accessories (*Fashion*), Gaming (*Gaming*), Cell phone accessories (*Cell*

Phone) and Hotels (*Hotels*)). We perform the same demographic selection criteria as the `sentiment` task, resulting in a multiclass task with 5 labels.

For each dataset, we obtain randomly stratified train-dev-test (60-20-20%) splits ensuring equal representations for both gender and age groups. For each review, we follow prior work (Hung et al., 2023) and set the maximum sequence length to 512 subword tokens, the max input size of BERT-style models (Devlin et al., 2019). Table 7 shows the final breakdown of the number of reviews in the datasets.

C.3 Social Media

Social media platforms host a diverse population, with studies demonstrating NLP system bias on related tasks (Aguirre et al., 2021).

Twitter sentiment. This is a binary sentiment classification task using Twitter data. Sentiment labels were assigned based on common emojis, following the preprocessing procedure of Elazar and Goldberg (2018). The demographic variables are based on the dialectal corpus from Blodgett et al. (2016), where race was assigned based on geolocation and words used in the tweet, obtaining a binary AAE (African-American English) and SAE (Standard American English) which we use as proxies for non-Hispanic African-Americans and non-Hispanic Caucasians.

HateXplain. This hate speech classification dataset combines Twitter and Gab messages (Mathew et al., 2021). We use the binary version of the task which identifies toxicity of posts. We select the posts for which there is a majority agreement of annotators for race target groups, and for which we have representation across train-dev-test splits.

For each dataset, we follow the splits provided by Elazar and Goldberg (2018) and Mathew et al. (2021) respectively. Table 7 shows the number of posts for the *HateXplain* and *Twitter sentiment* datasets respectively.

D Experiment Table

For each dataset, the model setup and their respective training data, fairness loss attribute and which task the fairness loss was applied to. `MTL-fair` are the models with the fairness loss from §2.2, and `MTL-inter` is the model with the intersectional fairness loss discussed in §2.3. * The `MTL-inter` model uses two separate single-attribute fairness losses for each task.

E Results without access to val set demographic attributes

The selection criteria for the best-performing models requires access to demographic attributes for the test set of the target task to assess the fairness of the models. In the absence of this, Table 9 shows the results for the model setting where we select models with the target task performance of at least 95% of `STL-base` and with the lowest fairness score of the auxiliary task. These models are labeled as `MTL-fair no demo`. For all of the datasets, `MTL-fair no demo` are less fair than if we could select models based on the fairness of the target task, `MTL-fair`. In some cases, we obtain models that are less fair than our single-task baselines (`STL-base`, 2/4) and multi-task baselines (`MTL-base`, 3/4). This suggest that while we are able to generalize the fairness loss to other tasks during training, the fairness measures across tasks are not related. For these reasons we recommend that `MTL-fair` models are validated for fairness on the target task.

Table 8: list of experiments

	Review Sentiment		
	training data	fairness loss attributes	fairness loss target task
STL-base	sentiment	no	no
STL-fair	sentiment	gender+age	sentiment
MTL-base	sentiment+topic	no	no
MTL-fair	sentiment+topic	gender+age	topic
	Review Topic		
	training data	fairness loss attributes	fairness loss target task
STL-base	topic	no	no
STL-fair	topic	gender+age	topic
MTL-base	sentiment+topic	no	no
MTL-fair	sentiment+topic	gender+age	sentiment
	In-Hospital Mortality		
	training data	fairness loss attributes	fairness loss target task
STL-base	In-hosp Mort.	no	no
STL-fair	In-hosp Mort.	gender	In-hosp Mort.
MTL-base	In-hosp Mort.+Phenotyping	no	no
MTL-fair	In-hosp Mort.+Phenotyping	gender	Phenotyping
	Phenotyping		
	training data	fairness loss attributes	fairness loss target task
STL-base	Phenotyping	no	no
STL-fair	Phenotyping	gender	Phenotyping
MTL-base	In-hosp Mort.+Phenotyping	no	no
MTL-fair	In-hosp Mort.+Phenotyping	gender	In-hosp Mort.
	Twitter Sentiment		
	training data	fairness loss attributes	fairness loss target task
STL-base	Twitter sentiment	no	no
STL-fair	Twitter sentiment	race	twitter sentiment
MTL-base	HateXplain+Twitter sentiment	no	no
MTL-fair	HateXplain+Twitter sentiment	race	HateXplain
	HateXplain		
	training data	fairness loss attributes	fairness loss target task
STL-base	HateXplain	no	no
STL-fair	HateXplain	race	HateXplain
MTL-base	Twitter sentiment+HateXplain	no	no
MTL-fair	Twitter sentiment+HateXplain	race	Twitter sentiment
	Intersectional Experiments		
	training data	fairness loss attributes	fairness loss target task
STL-base-sentiment	sentiment	no	no
STL-base-topic	topic	no	no
STL-fair-sentiment	sentiment	gender+age	sentiment
STL-fair-topic	topic	gender+age	topic
MTL-base	sentiment+topic	no	no
MTL-inter	sentiment+topic	gender/age*	sentiment/topic*

Table 9: Scores of the multi-task fairness loss experiments. For the Phenotyping task, these are macro-averages over all labels. Bold is best per task.

		method	AUROC (%) ↑	ϵ-DEO ↓	ΔRecall (%) ↓	ΔSpecificity (%) ↓
clinical	In-hosp Mort.	stl-base	77.7	0.22	2.05	5.99
		stl-fair	77.5	0.18	3.46	**3.54**
		mtl-base	**78.1**	0.17	**0.23**	4.45
		mtl-fair	78.1	**0.14**	0.98	3.83
		mtl-fair no demo.	78.4	0.18	1.80	4.02
	Phenotyping	stl-base	69.5	0.24	4.97	3.17
		stl-fair	69.6	**0.21**	**4.63**	2.96
		mtl-base	69.7	0.29	5.47	4.12
		mtl-fair	**69.9**	0.23	5.94	**2.46**
		mtl-fair no demo.	70.9	0.28	6.18	4.25
		method	F1 (%) ↑	ϵ-DEO ↓	ΔF1 (%) ↓	
reviews	sentiment	stl-base	83.9	0.83	3.79	
		stl-fair	**86.1**	0.68	3.05	
		mtl-base	83.5	0.66	4.75	
		mtl-fair	84.4	**0.63**	**1.96**	
		mtl-fair no demo.	83.3	0.89	5.92	
	topic	stl-base	91.9	1.42	4.58	
		stl-fair	**92.1**	1.04	**2.86**	
		mtl-base	91.3	1.10	6.15	
		mtl-fair	91.6	**0.85**	3.22	
		mtl-fair no demo.	91.3	1.11	4.79	