

Fair Text Classification via Transferable Representations

Thibaud Leteno

THIBAUD.LETENEO@UNIV-ST-ETIENNE.FR

*Université Jean Monnet Saint-Etienne,
CNRS, Institut d'Optique Graduate School,
Laboratoire Hubert Curien UMR 5516, F-42023,
Saint-Etienne, France*

Michael Perrot

MICHAEL.PERROT@INRIA.FR

*Univ. Lille, Inria, CNRS, Centrale Lille,
UMR 9189 - CRISTAL, F-59000, Lille, France*

Charlotte Laclau

CHARLOTTE.LACLAU@TELECOM-PARIS.FR

*LTCI, Télécom Paris
Institut Polytechnique de Paris, France*

Antoine Gourru

ANTOINE.GOURRU@UNIV-ST-ETIENNE.FR

*Université Jean Monnet Saint-Etienne,
CNRS, Institut d'Optique Graduate School,
Laboratoire Hubert Curien UMR 5516, F-42023,
Saint-Etienne, France*

Christophe Gravier

CHRISTOPHE.GRAVIER@UNIV-ST-ETIENNE.FR

*Université Jean Monnet Saint-Etienne,
CNRS, Institut d'Optique Graduate School,
Laboratoire Hubert Curien UMR 5516, F-42023,
Saint-Etienne, France*

Editor: Manuel Gomez-Rodriguez

Abstract

Group fairness is a central research topic in text classification, where reaching fair treatment between sensitive groups (e.g., women and men) remains an open challenge. We propose an approach that extends the use of the Wasserstein Dependency Measure for learning unbiased neural text classifiers. Given the challenge of distinguishing fair from unfair information in a text encoder, we draw inspiration from adversarial training by inducing independence between representations learned for the target label and those for a sensitive attribute. We further show that domain adaptation can be efficiently leveraged to remove the need for access to the sensitive attributes in the data set we cure. We provide both theoretical and empirical evidence that our approach is well-founded.

Keywords: natural language processing, fairness, text classification, domain adaptation, transfer

1. Introduction

Machine learning algorithms have become increasingly influential in decision-making processes that significantly impact our daily lives. One of the major challenges that has emerged in research, both academic and industrial, concerns the fairness of these models, that is,

their ability to treat individuals and groups equitably without causing prejudice or discrimination. As more researchers work to overcome these shortcomings, the first problem is to define what *fairness* is. This definition may hardly be consensual (Han et al., 2023) or is at least difficult to establish, as it depends on situational and cultural contexts (Fiske, 2017). In this work, we focus on group fairness (that we will refer to as fairness for simplicity), which prevents predictions related to individuals from being based on sensitive attributes such as gender or ethnicity. We then adopt common metrics for assessing group fairness in practice, which are based on the notion of disparate impact referenced in legal frameworks across several countries.¹ This type of metric considers a predictive model fair if its outcomes remain consistent across groups of individuals defined by sensitive attributes.

In this article, we focus on the problem of fairness in the domain of Natural Language Processing (NLP) (Li et al., 2023; Chu et al., 2024) and more specifically for text classification as it is one of the most ubiquitous tasks in our society, with prominent examples in medical and legal domains (Demner-Fushman et al., 2009) or human resources (Jatobá et al., 2019), to name a few. For more general overviews of fairness in machine learning systems, we refer the interested readers to Caton and Haas (2024); Barocas et al. (2023). Initially, works in text classification rely on text encoders, which are parameterized and learned functions that map tokens (arbitrary text chunks) into a latent space of controllable dimension, usually followed by a classification layer. Built upon the Transformers architecture (Vaswani et al., 2017), popular Pre-trained Language Models (PLMs) such as BERT (Devlin et al., 2019) leverage self-supervised learning to train the text encoder parameters. These PLMs are further fine-tuned for the supervised task at hand. More recently, with the advent of powerful decoder-based models, practitioners started to prompt those models for classification tasks (Dubey, 2024; Ruan et al., 2024). While many studies already report biases in NLP systems (Sun et al., 2019; Hutchinson et al., 2020; Tan and Celis, 2019; Liang et al., 2021; Bender et al., 2021), these issues become even more significant with the advent of public-ready AI-powered NLP systems. As mentioned above, recent developments in NLP, such as prompting-based models, raise questions about ensuring fairness in text classification. Atwood et al. (2024) highlight the limitations of prompting for fairness control, whereas regularization-based methods achieve better fairness-performance trade-offs. Meanwhile, Roccabruna et al. (2024) evaluate multiple large decoder-based models alongside RoBERTa (Liu et al., 2019) on temporal relation classification, finding that RoBERTa outperforms all the decoder-based models for this task. However, other approaches leverage powerful decoder models to generate embeddings for various tasks, including text classification, as seen with SFR-Embedding-2_R (Meng et al., 2024) or NV-Embed-v2 (Lee et al., 2024) both built on Mistral-7B (Jiang et al., 2023). While some recent works adopt this embedding-based strategy (Yang, 2024), others continue to rely on encoder-only architectures (Sturman et al., 2024). For fairness control in text classification, this leaves two main approaches: incorporating fairness constraints into prompts or debiasing the model during fine-tuning. Our work is part of this latter setting.

1. for example, GDPR, Article 22 (European Parliament and Council of the European Union, 2016) and AI Act (European Parliament and Council of the European Union, 2024), Recital 27 in the European Union, Title VII of the 1964 Civil Rights Act (Act, 1964) in the United States of America.

Contributions This paper extends our work on Wasserstein independence for text classification (Leteno et al., 2023) to mitigate bias in text classifiers. We introduce an extensive theoretical analysis and present additional experimental results. Our approach addresses bias directly in the latent space, making it applicable to any text encoder or decoder (e.g., BERT or Mistral). To proceed, we disentangle the neural signals encoding bias from those used for predictions. Disentanglement-based methods have primarily focused on images or tabular data (Jang and Wang, 2024; Locatello et al., 2019). In this paper, we introduce an approach tailored to NLP and capable of handling less-explored scenarios, including continuous sensitive attributes and regression tasks. Our method overcomes a major shortcoming of prior studies that rely on access to the sensitive attributes during training - regulations, such as GDPR (European Parliament and Council of the European Union, 2016), impose more stringent requirements for the collection and use of protected attributes, which can, in certain cases, pose constraints on some methodologies. In the following, we demonstrate that our approach tackles this issue by learning from simple data sets, such as toy data sets, to transfer knowledge and enable fair classification even when sensitive attributes are not available in the deployment data.

In a nutshell, our goal is to reduce the dependency between predictions and sensitive attributes to improve fairness. To achieve this, we minimized the Wasserstein Dependency Measure (Ozair et al., 2019) between the hidden representations of two neural networks: one for the end-task classification and one for predicting the sensitive attributes. This requires approximating several measures relative to the initial objective of independence between the classifier and the sensitive attribute. In this paper, we establish the theoretical validity of these approximations. First, we examine the relation between the chosen dependency measure and various fairness metrics. Second, we derive an upper bound on the transfer of sensitive attributes, supporting the use of predicted sensitive attributes when the real ones are unavailable. Finally, we justify the use of latent representations and provide guarantees on this approximation. We further validate our approach empirically by comparing it to state-of-the-art methods and evaluating different variations of our architecture.

Organization of the paper The rest of this paper is organized as follows. Section 2 presents recent advances related to our proposition. Section 3 discusses our motivation, provides the background knowledge to understand our contributions, and presents our first results that establish the relation between fairness and the Wasserstein Dependency Measure. Section 4 proceeds with the theoretical framework of the proposed approach and its analysis. Section 5 provides the description of the proposed approach and the algorithmic details of the implementation. Section 6 introduces the setting of our experiments, and Section 7 presents the experiments and their interpretations. We present our conclusions and research perspectives in Section 8 and end the paper with a section dedicated to the limitations of our contributions.

2. Related Works

Recent work on fairness in NLP has focused on fair text classification with adversarial methods (Beutel et al., 2017; Zhang et al., 2018; Elazar and Goldberg, 2018; Madras et al., 2018; Torres, 2024) being widely investigated. Han et al. (2021b,a) suggest using multiple

discriminators, each learning distinct hidden representations or applying adversarial training across domains. Other contributions enforce fairness through balanced training (Han et al., 2021c), batch selection (Roh et al., 2021), or by integrating fairness metrics, such as Equality of Opportunity, directly into the objective function (Shen et al., 2022a,b). However, these methods rely on access to sensitive attribute annotations during training, limiting their practical applicability. In this work, we overcome this constraint while providing strong theoretical guarantees.

Next, we focus on related work that considers settings where sensitive attributes are unavailable, followed by fairness approaches based on dependency measures and theoretical guarantees.

Sensitive attribute access for fairness mitigation To address their absence, proxy models have been proposed to enhance fairness. Other approaches circumvent the use of sensitive attributes during training or inference by leveraging related features (Zhao et al., 2022), knowledge distillation (Chai et al., 2022), adversarial reweighted learning (Lahoti et al., 2020), proxy features (Gupta et al., 2018), or perturbations (Awasthi et al., 2020). However, Kenfack et al. (2023) recently highlighted the risks associated with proxy-sensitive attributes, which may exacerbate the fairness-accuracy trade-off. Domain adaptation has also been explored as a means to address fairness in data sets lacking demographic information. Schumann et al. (2019) employ adversarial learning to enforce fairness in the source domain while predicting domain membership, while Coston et al. (2019) propose loss reweighting to mitigate the absence of sensitive attributes in either domain. Our approach follows this line of research, specifically addressing the lack of sensitive attributes in the target domain. By working in the representation space to minimize divergence between domains, we aim to ensure that the classifier trained on the source domain treats both domains equivalently.

Fair classification with dependency measures The Wasserstein distance has been increasingly used to enforce fairness constraints in machine learning. For instance, Risser et al. (2022) and Jiang et al. (2020) apply it to measure the discrepancy between the distributions of predictions conditionally on groups defined by the sensitive attribute. Although effective, these approaches are limited to categorical sensitive attributes and mainly favor conditional independence. In contrast, we propose to exploit the Wasserstein dependency measure, which captures the dependence between the joint distribution of the hidden output representations and the sensitive attribute, and the product of their marginals. This distinction allows us to assess and mitigate bias at a more fundamental level, ensuring that the learned representations themselves do not encode sensitive information. Our approach is inspired by Ozair et al. (2019), which uses Wasserstein’s dependency measure to improve representation learning for images. However, while their work focuses on improving feature representations for downstream tasks, we incorporate sensitive attributes into the estimation process to promote fairness.

Another related approach in NLP is proposed by Cheng et al. (2021), which maximizes the mutual information between sentence representations and their augmented counterparts to remove sensitive information from inputs. However, as noted by Shen et al. (2022b) and Cabello et al. (2023), this does not guarantee the independence between predictions and

sensitive attributes. Our method differs by explicitly minimizing the dependency between representations of the same sentence processed by two different encoders, ensuring that predictions remain unaffected by sensitive attributes.

Additionally, our work shares conceptual similarities with Nam et al. (2020), which addresses bias in image data. However, instead of focusing on reweighting samples to counteract biases in a secondary model, we employ the Wasserstein distance to quantify and minimize the dependency between the representations learned by two models. More recently, Iskander et al. (2024) also seeks to mitigate disparities but relies on task-specific representations and KL divergence to enforce distributional uniformity across groups.

Theoretical guarantees in fairness Most fairness mitigation techniques are evaluated on test sets that may not fully represent real-world deployment scenarios (Dunkelau, 2020; Hort et al., 2024). This highlights the need for theoretical guarantees to ensure the reliability of mitigation approaches with respect to fairness metrics. Several works provide such guarantees, often focusing on post-training corrections. For instance, Woodworth et al. (2017) propose a post-hoc correction method with guarantees on classifier performance and prediction disparities across sensitive attributes. Denis et al. (2024) derive distribution-free fairness guarantees, while Chzhen et al. (2020) establish fairness bounds dependent only on the dimensionality of the unlabeled data set.

On the other hand, Celis et al. (2019) develop a meta-learning framework to obtain an optimally fair classifier with respect to algorithmic complexity, and McNamara et al. (2017) show that learned representations can satisfy both group and individual fairness criteria. Finally, a closely related work is Gupta et al. (2021), who consider Mutual Information to measure the dependency between representations, providing fairness guarantees based on this latter. They derive an upper bound on the Demographic Parity measure via the Mutual Information between latent representations and the sensitive attributes, as well as bounds on the Mutual Information between classification labels and conditional latent representations. However, unlike our approach, they do not provide guarantees on the dependency between the classification labels and sensitive attributes.

3. Wasserstein Dependency Measure and Group Fairness

This section introduces the notations used throughout the paper, along with the definitions of key fairness metrics and the Wasserstein Dependency Measure (I_W). We then present our first result, establishing a link between two popular group fairness metrics and I_W .

3.1 Notations

We consider a corpus of n triplets $\{(x_i, y_i, a_i)\}_{i=1}^n$, where $x_i \in \mathcal{X}$ is a short document or a sentence, $y_i \in \mathcal{Y}$ is a label and $a_i \in \mathcal{A}$ is either a *sensitive* attribute, such as gender, ethnicity or age, or represents intersectional groups of several sensitive attributes. In this paper, we assume that \mathcal{Y} and \mathcal{A} are discrete spaces, and we will often abuse notations such that $y \in \mathcal{Y}$ and $a \in \mathcal{A}$ represent either a target label or a vector representation obtained through one-hot encoding. The embeddings (or representations) are obtained thanks to an encoding function, Enc , that maps words into numeric values. The objective is to predict outcomes y for a given input x by estimating the conditional distribution $p(Y|X=x)$. To this end, we learn a scoring function $\pi_y : \mathcal{X} \rightarrow \mathcal{P}(\mathcal{Y})$ where $\mathcal{P}(\mathcal{Y})$ is the set of probability

distributions over \mathcal{Y} . Given $\pi_y(x)$, the actual prediction is denoted by \hat{y} and corresponds to the label predicted as most likely. For instance, in a social network context, one can learn a classifier to predict whether a message is toxic. This prediction could inform decisions such as banning the message or its author from the platform.

In modern NLP applications, deep classification often follows a two-step approach: the scoring function π is expressed as $\pi_y = h_y \circ \text{Enc}$, where $\text{Enc}(x) \in \mathbb{R}^d$ maps a text x into a low-dimensional embedding space, and h_y , typically a simple neural network layer with a softmax activation serves as the classification layer.

3.2 Group Fairness

Our goal is to learn fair models and we focus on two main definitions of fairness. On the one hand, we consider demographic parity (Hardt et al., 2016) which is defined, for a desirable outcome y and a sensitive attribute a , as

$$\mathbf{DP}_{a,y} = \mathbb{P}(\hat{Y} = y \mid A = a) - \mathbb{P}(\hat{Y} = y). \quad (1)$$

On the other hand, we consider equality of opportunity (Hardt et al., 2016) which is defined, for an outcome y and a sensitive attribute a , as

$$\mathbf{EO}_{a,y} = \mathbb{P}(\hat{Y} = Y \mid Y = y, A = a) - \mathbb{P}(\hat{Y} = Y \mid Y = y). \quad (2)$$

3.3 Wasserstein Dependency Measure

Mutual Information (MI) is an information-theoretic metric that measures the statistical dependence or the amount of information shared between two variables. For two random variables $U \sim p(U)$ and $V \sim p(V)$ that takes values in \mathcal{U} and \mathcal{V} , respectively, the MI is defined as the KL-divergence between the joint distribution $p(U, V)$ and the product of the marginal distributions $p(U)p(V)$

$$\text{MI}(U, V) = \text{KL}(p(U, V) \parallel p(U)p(V)).$$

Early works in fair classification introduced the idea that fairness can be improved by reducing the Mutual Information (MI) between the classifier’s output, \hat{Y} , and the sensitive attribute, A (Kamishima et al., 2012; Zemel et al., 2013). Specifically, enforcing Demographic Parity (DP) corresponds to minimizing the MI between these two random variables, ensuring that \hat{Y} is independent of A . Similarly, Equalized Odds (EO) can be formulated as minimizing the MI between A and \hat{Y} conditionally on the true label Y , ensuring that predictions remain independent of the sensitive attribute within each outcome class.

However, MI is known to be intractable for most real-life scenarios and has strong theoretical limitations as outlined by McAllester and Stratos (2020). Notably, it requires an exponential number of samples in the value of the MI to build a high confidence lower bound, and it is sensitive to small perturbations in the data sample. To overcome this issue, Ozair et al. (2019) propose a theoretically sound dependency measure, the *Wasserstein Dependency Measure* (I_W), based on the Wasserstein 1-distance

$$I_W(U, V) = W_1(p(U, V), p(U)p(V)).$$

Using the Kantorovich-Rubinstein duality, it can also be expressed as

$$I_W(U, V) = \sup_{\|f\|_L \leq 1} \mathbb{E}_{U, V \sim p(U, V)}[f(U, V)] - \mathbb{E}_{U \sim p(U), V \sim p(V)}[f(U, V)], \quad (3)$$

where $\|f\|_L \leq 1$ is the set of all 1-Lipschitz functions. The Wasserstein distance has been efficiently used in many machine learning applications (Frogner et al., 2015; Courty et al., 2014; Torres et al., 2021) and a particularly interesting one is that of fair machine learning (Jiang et al., 2020; Silvia et al., 2020; Gordaliza et al., 2019; Laclau et al., 2021). We present the Wasserstein distance in Appendix A along with technical lemmas relevant for our proposition.

3.4 Connection with Group Fairness

In this section, we show a connection between the Wasserstein Dependency Measure and the two group fairness measures we consider. Hence, in the next lemma, we show that a linear combination of Demographic Parity or Equality of Opportunity for all possible values of a and y are equivalent to the Wasserstein Dependency Measure between well-chosen random variables. This result is reminiscent of the result of Gupta et al. (2021), who showed a connection between group fairness and mutual information.

Lemma 1 (Group fairness and Wasserstein Dependency Measure.) *Let I_W be the Wasserstein dependency measure, and A, Y, \hat{Y} be random variables corresponding to the sensitive attribute, the true label, and the predicted label, respectively. Let $\|\cdot\|_p$ be the ground metric for the Wasserstein 1-distance. We have that*

$$I_W(\hat{Y}, A) = \frac{\sqrt[p]{2}}{2} \sum_{a \in \mathcal{A}} \mathbb{P}(A = a) \sum_{y \in \mathcal{Y}} |\mathbf{DP}_{a,y}|, \\ I_W((\hat{Y} = Y)|Y = y, A|Y = y) = \frac{\sqrt[p]{2}}{2} \sum_{a \in \mathcal{A}} \mathbb{P}(A = a|Y = y) |\mathbf{EO}_{a,y}|,$$

with $|\cdot|$ denoting the absolute value.

Proof The proof is provided in Appendix B. ■

This lemma shows that minimizing the Wasserstein Dependency Measure between well-chosen random variables is a sound way to minimize Demographic Parity or Equality of Opportunity. This motivates the regularization of a learning algorithm by $I_W(\hat{Y}, A)$ to improve the fairness of text classifiers.

4. Predictive and Sensitive Information Approximations

To improve classifier fairness, we aim to minimize the Wasserstein Dependency Measure (I_W) between the sensitive attribute A and the label predictions \hat{Y} . However, this optimization presents several challenges, notably having access to the sensitive attributes and requiring to differentiate a signal that went through an argmax function to obtain the label predictions.

To address these, we first approximate the sensitive attribute labels using their predicted values, \hat{A} , obtained from a neural network. Then, instead of working directly with \hat{Y} and \hat{A} , we use their hidden representations, denoted as Z_y and Z_a , from the corresponding neural networks to overcome the non-differentiability of the argmax function. We also provide guarantees on these approximations. This leads to the following optimization objective for learning a fair text classifier:

$$\arg \min \mathcal{L}(Y, h_y(Enc(X_y))) + \beta I_W(Z_y, Z_a), \quad (4)$$

where $I_W(Z_y, Z_a) = W_1(p(Z_y, Z_a), p(Z_y)p(Z_a))$. Here, Z_y and Z_a represent the hidden representations from two Multi-Layer Perceptrons (MLPs): one for classification and one for the proxy model introduced in Section 4.1. The function \mathcal{L} ensures the classifier achieves high accuracy on Y (e.g., we consider the cross-entropy for binary classification), while the second term encourages fairness by constraining the learned representations. The hyperparameter $\beta \in \mathbb{R}^+$ controls the balance between accuracy and fairness, as the two objectives may converge at different speeds.

We refer to this approach as Wasserstein Fair Classification (WFC). Details on its implementation are provided in Section 5.

4.1 Definition of the Demonic Model

In the following, we use a surrogate model, referred to as the *demonic* model, for predicting the sensitive attribute A without requiring explicit observation of attributes at training time. To proceed, we assume a similar architecture as for predicting the labels: we learn a scoring function $\pi_a = h_a \circ Enc$ which, given an example x , outputs a probability distribution over \mathcal{A} , with h_a a fixed classification function predicting A . The predicted sensitive attribute is then \hat{a} and corresponds to the most likely sensitive attribute according to π_a . Consequently, we propose to consider $I_W(\hat{Y}, \hat{A})$ instead of $I_W(\hat{Y}, A)$ to approximate the dependency between the predictions and the sensitive attributes. In the next theorem, we study this approximation and show that it is close to the original measure while being dependent on the *demonic* model performance.

Lemma 2 *Let \hat{Y}, \hat{A}, A be random variables that correspond to the predicted label, predicted sensitive attribute, and true sensitive attribute, respectively. Let $\|\cdot\|_p$ be the ground metric for the Wasserstein 1-distance. Then, we have that*

$$I_W(\hat{Y}, A) \leq I_W(\hat{Y}, \hat{A}) + 2\sqrt[2]{2}\mathbb{P}(A \neq \hat{A}).$$

Proof The proof is provided in Appendix C. ■

This lemma shows that replacing A by \hat{A} is sound when the latter is an accurate estimate of the former, that is, when $\mathbb{P}(A \neq \hat{A})$ is small. In the next theorem, we combine this result with a standard generalization result to show that this remains valid in the finite sample regime. The proof is provided in Appendix C.1.

Theorem 3 *Let $\hat{A}, A \in \{0, 1\}$, and \mathcal{H} be a hypothesis space of VC-dimension d . Let $\|\cdot\|_p$ be the ground metric for the Wasserstein 1-distance. Assume that we have access to a training*

set of m i.i.d. examples. Then, with probability at least $1 - \delta$, we have $\forall h \in \mathcal{H}$

$$I_W(\hat{Y}, A) \leq I_W(\hat{Y}, \hat{A}) + 2\sqrt[3]{2} \left(\hat{\varepsilon} + \sqrt{\frac{4}{m} \left(d \log \frac{2em}{d} + \log \frac{4}{\delta} \right)} \right),$$

with e , the base of the natural logarithm and $\hat{\varepsilon}$ the empirical risk of the demonic model.

Remark This bound indicates that minimizing $I_W(\hat{Y}, \hat{A})$ allows to minimize $I_W(\hat{Y}, A)$. However, it is tight when the *demonic* model is accurately predicting the sensitive attributes. In other words, with an accurate *demonic* model, the bound on the error rate is low and the bound tends to the estimate $I_W(\hat{Y}, \hat{A})$. In the perfect case, where the *demonic* model achieves perfect predictions, the bound is simply $I_W(\hat{Y}, \hat{A})$. Moreover, with input data of sufficient size, the bound on the error rate ε gets lower. We will consider the case where the *demonic* model is trained on data out of the domain (transfer learning scenario) later in Section 4.2. Note that we can easily generalize to multi-label sensitive attributes by considering the Natarajan dimension (Natarajan, 1989) instead of the VC-dimension. Moreover, multiple sensitive attributes can be considered by looking at the groups' intersections to cast the problem as a multi-label one. Finally, continuous sensitive attributes can be handled by binning. However, finding the relevant threshold for binning is a non-trivial problem. We present the different scenarios and solutions in Appendix E.

4.2 Demonic Model in Cross-Domain Settings

Recall that \hat{A} and the latent representations Z_a are obtained through a proxy neural network trained to predict the sensitive attribute to tackle the lack of sensitive attribute annotation. As it, one can train h_a on a different data set from the end-task one.

Let us consider two data sets, the end-task data set (or target) $\mathcal{D}_{\mathcal{T}}$ and the side data set (or source) $\mathcal{D}_{\mathcal{S}}$. $\mathcal{D}_{\mathcal{T}} = \{x_{\mathcal{T},i}, y_{\mathcal{T},i}\}_i^{n_{\mathcal{T}}}$ is composed of a set of features and labels, while $\mathcal{D}_{\mathcal{S}} = \{x_{\mathcal{S},i}, a_{\mathcal{S},i}\}_i^{n_{\mathcal{S}}}$ is composed of a set of features and sensitive attributes. We assume that we are in the context of covariate shift: the feature distributions are different but the sensitive attribute distributions are similar ($\mathcal{A}_{\mathcal{T}} \approx \mathcal{A}_{\mathcal{S}}$).

Then, we want to learn a mapping $\phi : X_{\mathcal{S}} \rightarrow X_{\mathcal{T}}$ and train the *demonic* model classification layer h_a on the mapped $X_{\mathcal{S}}$:

$$\min_{h_a, \phi} \mathcal{L}(h_a(Enc(X_{\mathcal{S}})), A_{\mathcal{S}}) + \Lambda(\phi(Enc(X_{\mathcal{S}})), Enc(X_{\mathcal{T}})),$$

with $\Lambda(\phi(Enc(X_{\mathcal{S}})), Enc(X_{\mathcal{T}}))$ the measure of divergence between the embeddings of $\mathcal{X}_{\mathcal{T}}$ and $\mathcal{X}_{\mathcal{S}}$. Note that the encoder Enc has to be the same for the source and target domains.

We provide experimental details in Section 5.2. Moreover, Theorem 3 can be adapted to this setting, only the approximation of the error rate of the *demonic* model changes.

Theorem 4 Assuming that $\hat{A}, A \in \{0, 1\}$. Assume that $\mathcal{D}_{\mathcal{S}}$ and $\mathcal{D}_{\mathcal{T}}$ are a source and a target distribution such that $\mathbb{P}_{\mathcal{D}_{\mathcal{S}}}(X = x) \neq \mathbb{P}_{\mathcal{D}_{\mathcal{T}}}(X = x)$ and $\mathbb{P}_{\mathcal{D}_{\mathcal{S}}}(A = a|X = x) = \mathbb{P}_{\mathcal{D}_{\mathcal{T}}}(A = a|X = x)$, that is assume a covariate-shift. Let $\|\cdot\|_p$ be the ground metric for the Wasserstein 1-distance. Assume that $I_W(\hat{Y}, A)$ and $I_W(\hat{Y}, \hat{A})$ are computed on the target

distribution and let $\varepsilon_S = \mathbb{P}_{\mathcal{D}_S}(\hat{A} \neq A)$, $\varepsilon_T = \mathbb{P}_{\mathcal{D}_T}(\hat{A} \neq A)$, then we have that

$$I_W(\hat{Y}, A) \leq I_W(\hat{Y}, \hat{A}) + 2\sqrt[p]{2} \left(\varepsilon_S + \frac{1}{2} d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) + \lambda \right),$$

where $d_{\mathcal{H}\Delta\mathcal{H}}(\tilde{\mathcal{D}}_S, \tilde{\mathcal{D}}_T)$ is the $\mathcal{H}\Delta\mathcal{H}$ -divergence between the marginal feature distributions $\tilde{\mathcal{D}}_S$ and $\tilde{\mathcal{D}}_T$ and $\lambda = \lambda_S + \lambda_T$ with λ_S and λ_T the errors of $h^* = \operatorname{argmin}_{h \in \mathcal{H}}(\varepsilon_T(h), \varepsilon_S(h))$ with respect to \mathcal{D}_S and \mathcal{D}_T respectively.

Proof This is a direct application of Ben-David et al. (2010, Theorem 2). ■

Remark We can draw similar conclusions as for Theorem 3. However, in this case, one must also consider the divergence between the domains, determinant to the success of the approximation. The closer the two domains are, the tighter the bound is. Therefore, if the *demonic* model decreases in accuracy due to the divergence between the source and target domains, the bound gets looser. Note that the $\mathcal{H}\Delta\mathcal{H}$ -divergence between the source and target domains is restricted to the binary setting. To generalize to categorical sensitive attributes (and by extension to multiple or continuous sensitive attributes as in Theorem 3), one could consider a generalization of the $\mathcal{H}\Delta\mathcal{H}$ -divergence as in (Sicilia et al., 2022). Recall that we detail the different scenarios and solutions in Appendix E.

4.3 Using Latent Representations

In the previous section, we explained why using the Wasserstein Dependency Measure between the predicted labels and sensitive attributes, $I_W(\hat{Y}, \hat{A})$, instead of between the predicted labels and the true sensitive attributes, $I_W(\hat{Y}, A)$. Nevertheless, as such, we cannot consider this measure to regularize any training algorithms since the argmax operation producing the hard predictions (\hat{Y}) following the classification layer is not differentiable. Thus, instead of considering the network's final output, one can overcome this limitation by minimizing the I_W between the latent representations of the networks h_y and h_a , respectively referred to as Z_y and Z_a . In Theorem 5, we show that the I_W between the neural networks' representations is an upper bound of the I_W between the predictions.

Theorem 5 *Let \hat{Y}, \hat{A} be random variables that correspond to the predicted label and predicted sensitive attribute, respectively. Assume that $h_y = \sigma_\lambda(f(Z_y))$ and $h_a = \sigma_\lambda(g(Z_a))$ where σ_λ is the softmax function with temperature λ , f and g are both L -lipschitz with respect to the p -norm, and Z_y and Z_a are latent representations of the examples. Let $\|\cdot\|_p$ be the ground metric for the Wasserstein 1-distance. For a given example x with predicted label \hat{y} and predicted sensitive attribute \hat{a} , let $\xi_y(x) = f(Z_y)_{\hat{y}} - \max_{y' \neq \hat{y}} f(Z_y)_{y'}$ and $\xi_a(x) = g(Z_a)_{\hat{a}} - \max_{a' \neq \hat{a}} g(Z_a)_{a'}$ be positive margins. Let $\delta = 1 - \mathbb{P}(\xi_y(X) \geq \xi, \xi_a(X) \geq \xi)$ with $\xi > 0$. Let $\alpha = \sqrt[p]{2} \left\| \left(\frac{|\mathcal{Y}|}{|\mathcal{A}|} \right) - 1 \right\|_p (1 - \delta)$ and $\iota = L(|\mathcal{Y}| + |\mathcal{A}|)^{\frac{1}{2} - \frac{1}{p}}$. Then, setting $\lambda = \frac{1}{\xi} \log \left(\frac{2\xi\alpha}{\iota I_W(Z_y, Z_a)} \right)$, we have that*

$$I_W(\hat{Y}, \hat{A}) \leq 2I_W(Z_y, Z_a) \frac{\iota}{\xi} \left[1 + \log \left(\max \left(4, \frac{2\xi\alpha}{\iota I_W(Z_y, Z_a)} \right) - 1 \right) \right] + \sqrt[p]{2} \left\| \left(\frac{|\mathcal{Y}|}{|\mathcal{A}|} \right) - 1 \right\|_p \delta.$$

Proof The proof of a slightly sharper result, in particular when $I_W(Z_y, Z_a)$ is large, is provided in Appendix D. We present this simpler version here for better readability. ■

Remark This result suggests that minimizing $I_W(Z_y, Z_a)$ is a sound way to minimize $I_W(\hat{Y}, \hat{A})$. The tightness of the bound depends mainly on the error introduced by the softmax and, more specifically, on two terms: ξ and δ . The margin $\xi_y(x)$ (resp. $\xi_a(x)$) measures how dominant the predicted class is relatively to the others, i.e., it is large when \hat{Y} in one-hot encoded form and $\sigma_\lambda(f(Z_y))$ are close. In other words, $\xi_y(x)$ (resp. $\xi_a(x)$) represents the confidence level of the classification model and ξ represents the minimum expected confidence. The term δ is the proportion of examples for which this minimum confidence is not obtained by the model. We note that there is a trade-off between the first and the second term in the bound, depending on the value of ξ , as a high value of ξ is likely to imply a large δ and vice versa.

This result also indicates that for a given model, there is an optimal softmax temperature for inference, but, since the models are assumed given, it does not help us finding the optimal softmax temperature at training time. Furthermore, since the softmax is followed by an argmax function, the optimal temperature at inference has a limited impact. Finally, the way it is used here, temperature scaling is mainly seen as a technical tool and it might be possible to derive a similar result without it, thus it might not be necessary. For all these reasons, we do not further investigate this term experimentally.

5. Implementation of Wasserstein Fair Classification

In this section, we present both the overall architecture of WFC and the implemented training strategy.

5.1 Architecture of WFC

The overall architecture of WFC is composed of three components: two classifiers and a critic (see Figure 1). We recall that the architecture aims to minimize the loss function described in Equation 4.

Learning Z_y and Z_a Given a batch of documents along with their sensitive attribute, we start by generating a representation of each document using a pre-trained language model (PLM). These representations serve as input to two MLPs, which are trained to predict A and Y , respectively. The first model, referred to as the *demonic* model, is pre-trained. The prediction \hat{Y} outputted by the second MLP (in green in Figure 1) is directly used to compute the first term of our objective function (see Equation 4). Additionally, from a given hidden layer in each of the MLPs, we extract the hidden representation vectors, Z_y and Z_a , which capture intermediate features relevant to their respective tasks.

Computing $I_W(Z_y, Z_a)$ The second term of the loss is the I_W between Z_y and Z_a . To compute this latter, we use the following approximation (Arjovsky et al., 2017)

$$\max_{\omega, \|C_\omega\|_L \leq 1} \mathbb{E}_{Z_y, Z_a \sim p(Z_y, Z_a)} [C_\omega(Z_y, Z_a)] - \mathbb{E}_{Z_y \sim p(Z_y), Z_a \sim p(Z_A)} [C_\omega(Z_y, Z_a)]. \quad (5)$$

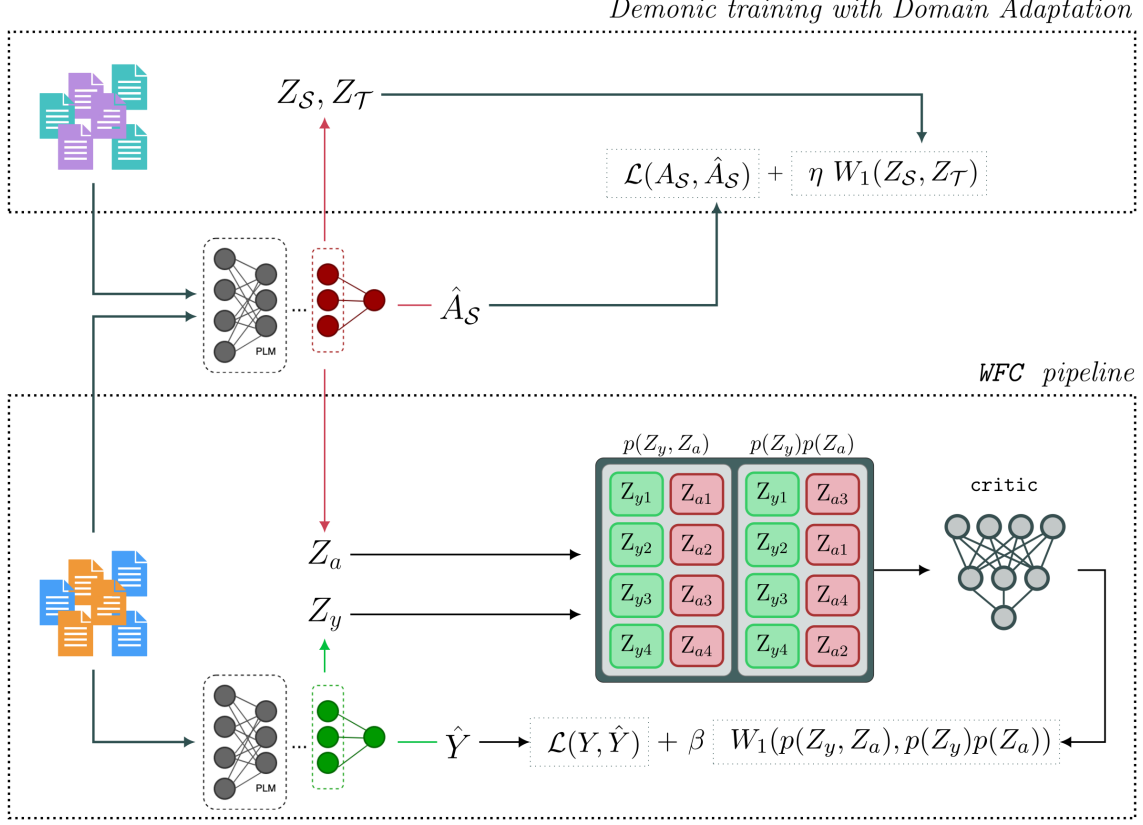


Figure 1: Architecture of our method. The top part illustrates the pre-training of the *demonic* model (red) with domain adaptation. The model is trained to predict the sensitive attribute on the source domain (A_S) while minimizing the divergence between the hidden representations from the source and target domains (Z_S and Z_T). The bottom part describes the WFC pipeline for a batch of size 4, the demonic model is then frozen. The data representation on the right shows how we enforce dependency or independence between Z_y and Z_a . During inference, only the trained classifier (green) is retained to predict Y .

where C_ω is called the critic. Initially proposed for Wasserstein GAN (Generative Adversarial Network) (Arjovsky et al., 2017), the critic is a neural network used as an alternative to the GAN’s discriminator to overcome this latter unstable training (Arjovsky and Bottou, 2017). It estimates the Wasserstein distance between real and fake distributions by outputting scores (the Wasserstein distance) instead of classifying samples as probabilities. The underlying idea is to use a neural network with a Lipschitz constraint on the weights to approximate the Kantorovich-Rubinstein formulation of the Wasserstein distance (Equation 3). The Lipschitz constraint is induced by weights clipping or gradient penalty (Gulrajani et al., 2017). We follow Ozair et al. (2019) who use it to compute the I_W with this differentiable estimate of the Wasserstein distance. To enforce the Lipschitz constraint, we

clamp the weights to given values ($[-0.01, 0.01]$) at each optimization step.² For a batch of documents, the critic takes as input the concatenation of Z_y and Z_a , and the concatenation of Z_y and Z_a randomly drawn from the data set (equivalent to $Z_y \sim p(Z_y), Z_a \sim p(Z_a)$). We then follow the training procedure introduced by Arjovsky et al. (2017), which alternates maximizing Equation 5 in the `critic` parameters for n_c iterations and minimizing Equation 4 for n_d iterations in the h_y classifier parameters. We add a comparison to WFC_{eo} , where we compute and minimize the I_W between instances that were well classified during the training. This allows us to compare optimizing directly DP vs. EO.

Overall The overview of the training process is detailed in Appendix F.1. The details of the MLPs used to parameterize each component, including their architecture, are given in Appendix F. We evaluate and optimize the hyperparameters for our models on a validation set, focusing on the MLP and Critic learning rates, the value of n_d (number of batches used to train the main MLP), the layers producing Z_a and Z_y , the value of β and the value used to clamp the weights to enforce the Lipschitz constraint. The values allowing us to obtain the optimal trade-off between accuracy and fairness (DTO, cf. Section 6.1) during this process are presented in Appendix F.2. In all our experiments, and if not mentioned otherwise, the value of β is set to 1. Our implementation is available on Github: https://github.com/LetenoThibaud/wasserstein_fair_classification.

5.2 Pre-training the *Demonic* Model

Overview We pre-train the *demonic* model, an MLP with a similar architecture as the previous classifier, to predict the sensitive attributes. Note that we do not update the *demonic* weights during the training phase of the main model. The benefits are twofold. First, unlike previous works (Caton and Haas, 2020), we require only limited access to sensitive attribute labels during training, and we do not need access to the sensitive attributes at inference. This makes `WFC` highly compatible with recent regulations (e.g., US Consumer Financial Protection Bureau). Second, the *demonic* model can be trained in a few-shot fashion if some examples of the training set are annotated with sensitive attributes.

Learning with a related data set However, when no sensitive attributes are available in the training set, we replace the training data of the *demonic* model with data from another domain (e.g., another data set) containing sensitive information for the same attribute. For example, for gender, we can leverage generated data sets, like the EEC data set (Kiritchenko and Mohammad, 2018). This enables knowledge transfer between data sets, promoting fairness autonomy regardless of whether sensitive attributes are present in the data, as long as another data set with similar sensitive attributes exists. Finally, in most cases, sensitive attribute knowledge transfers easily between data sets without additional adjustments. However, when data set divergence is significant, domain adaptation techniques can be applied to ensure transfer quality.

Learning with domain adaptation If the training data set differs significantly from the end-task data set, we add a regularization term to the loss of the *demonic* model to train it with

2. We also tested some more recent improvements of Lipschitz constraint enforcement (Gulrajani et al., 2017; Wei et al., 2018). Interestingly, all lead to poorer performance.

a double objective: 1) predicting the sensitive attribute and 2) generating representations from the source and target domains that are both close and informative for classification. In practice, under the covariate shift assumption, we use the Wasserstein distance between the representations of the source and target data sets as a measure of divergence. For domain adaptation, as for **WFC**, a critic model estimates the Wasserstein distance between the source and target representations. We use this measure for domain adaptation as done in Shen et al. (2018). Note that while in **WFC** the Wasserstein distance is computed between the joint and the product of the marginal distributions of the representations to compute a measure of independence, here we compute it between the representations themselves. Specifically, we compute the Wasserstein distance between the last hidden states of the model for both sets of representations (source and target). Therefore, if we consider the source and target domains, respectively $\mathcal{D}_S = \{x_{S,i}, a_{S,i}\}_i^{n_S}$ and $\mathcal{D}_T = \{x_{T,i}, y_{T,i}\}_i^{n_T}$, with X_S, X_T the sets of input texts, A_S the sensitive attributes. The objective of the *demonic* model, h_a , can be written as follows

$$\arg \min \mathcal{L}(A_S, h_a(\text{Enc}(X_S)) + \eta W_1(Z_S, Z_T), \quad (6)$$

where \mathcal{L} is the loss function aiming at maximizing the accuracy of h_a on predicting A , and Z_S, Z_T are the hidden representations of the model respectively for X_S and X_T .

6. Experimental Framework

In this section, we present the setting of our experiments, that is, the data sets and metrics we consider.

6.1 Evaluation Metrics

In this section, we introduce the metrics used to evaluate the performance of the models. For utility, we consider the balanced accuracy (Bal. Acc.) to handle class imbalance in the data. For fairness, we recall in Section 3.1 the Equality of Opportunity (cf. Equation 2). In our experiments, we consider binary sensitive attributes ($\mathcal{A} = \{0, 1\}$). For multi-class objectives (e.g. $\mathcal{Y} = \{1, \dots, C\}$), one can aggregate EO scores over classes (formulated as the difference of true positive rates across sensitive groups). This measure is the TPR-parity (or TPR-GAP) score (De-Arteaga et al., 2019; Ravfogel et al., 2020) defined as follows

$$\text{TPR-parity} = \sqrt{\frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} (\text{TPR}_{1,c} - \text{TPR}_{0,c})^2},$$

with $\text{TPR}_{0,c}$ and $\text{TPR}_{1,c}$ the true positive rates for class c and respectively sensitive group 0 and 1. For clarity in the results' comparison with the accuracy score, we consider the following

$$\text{Fairness} = (1 - \text{TPR-parity}) * 100.$$

The Fairness score indicates a perfectly fair model when equal to 100, and unfair when equal to 0. Additionally, as fairness often requires determining a trade-off such that reaching equity does not degrade the general classification performance, Han et al. (2021c) proposed the Distance To Optimum (**DTO**) score. It measures the accuracy-fairness trade-off by

computing the Euclidean distance from a model to an *Utopia point* (point corresponding to the best accuracy and best fairness values across all the baselines). The goal is to minimize the DTO. Let consider the *Utopia point* with coordinates $\{\mathbf{accuracy}_u, \mathbf{fairness}_u\}$ and the performance of a model at a given epoch $\{\mathbf{accuracy}_m, \mathbf{fairness}_m\}$, then we have

$$\mathbf{DTO} = \sqrt{(\mathbf{fairness}_u - \mathbf{fairness}_m)^2 + (\mathbf{accuracy}_u - \mathbf{accuracy}_m)^2}.$$

Finally, we consider the **Leakage** metric, which measures the accuracy of a classification model trained to predict the sensitive attribute A from the latent representations (Z) of another model. Let us consider two models, a classification model h that we want to evaluate and another model $h_{leakage}$ trained to retrieve the sensitive information A from the latent representations of h , Z_h . We consider a test set of size n such that

$$\mathbf{Leakage} = \left(\frac{1}{n} \sum_{i=0}^n \mathbb{1}_L(Z_{hi}) \right) * 100 \text{ with } \mathbb{1}_L(Z_h) = \begin{cases} 1 & \text{if } h_{leakage}(Z_h) = A, \\ 0 & \text{if } h_{leakage}(Z_h) \neq A. \end{cases}$$

It measures the fairness of the latent representations themselves and demonstrates representation unfairness when close to 100. We use the architecture presented in Shen et al. (2022b), that is an MLP with one hidden layer of size 100, a ReLU activation function, and a constant learning rate of 0.001. The optimizer used is Adam.

6.2 Data Sets

We employ two widely-used data sets to evaluate fairness in the context of text classification, building upon prior research (Ravfogel et al., 2020; Han et al., 2021b; Shen et al., 2022b). Both data sets are readily available in the FairLib library (Han et al., 2022).

Bias in Bios (De-Arteaga et al., 2019). This data set, referred to as “Bios data set” in the rest of the paper, consists of brief biographies from the common crawl associated with occupations (a total of 28) and genders (male or female). As per the partitioning prepared by Ravfogel et al. (2020), the training, validation, and test sets comprise 257,000, 40,000, and 99,000 samples, respectively.

Moji (Blodgett et al., 2016). This data set contains tweets written in either “Standard American English” (SAE) or “African American English” (AAE), annotated with positive or negative polarity. We use the data set prepared by Ravfogel et al. (2020), which includes 100,000 training examples, 8,000 validation examples, and 8,000 test examples. The target variable Y represents the polarity, while the protected attribute corresponds to the ethnicity, indicated by the AAE/SAE attribute.

7. Results and Discussion

In this section, we consider three experimental axes to illustrate our method: 1) in-domain experiments compared to state-of-the-art methods, 2) cross-domain experiments, 3) analysis of the WFC method.

Model	Bal. Acc. \uparrow	Fairness \uparrow	DTO \downarrow	Leakage \downarrow
*CE	$72.3 \pm 0.5^\dagger$	$61.2 \pm 1.4^\dagger$	31.0	87.9 ± 3.3
INLP + BERT _{ft}	$73.3 \pm 0.0^\dagger$	$85.6 \pm 0.0^\dagger$	8.49	86.7 ± 0.6
Adv + BERT _{ft}	75.6 ± 0.4	90.4 ± 1.1	4.03	$78.8 \pm 6.0^\dagger$
Gate + BTEO + BERT _{ft}	$76.2 \pm 0.3^\dagger$	90.1 ± 1.30	3.55	$100.0 \pm 0.0^\dagger$
FairBatch + BERT _{ft}	$75.1 \pm 0.6^\dagger$	90.6 ± 0.5	4.47	$88.4 \pm 0.4^\dagger$
EO _{GLB} + BERT _{ft}	75.2 ± 0.2	$90.1 \pm 0.4^\dagger$	4.49	85.7 ± 1.2
DAFair + BERT _{ft}	$79.5 \pm 0.2^\dagger$	$73.1 \pm 1.1^\dagger$	18.3	-
Adv	$74.5 \pm 0.3^\dagger$	$81.5 \pm 2.0^\dagger$	11.1	-
Gate + BTEO	$74.9 \pm 0.2^\dagger$	$86.2 \pm 0.3^\dagger$	6.94	-
Con _{dp}	$75.8 \pm 0.3^\dagger$	$88.1 \pm 0.6^\dagger$	4.96	$54.2 \pm 0.9^\dagger$
Con _{eo}	$74.1 \pm 0.7^\dagger$	$84.1 \pm 3.0^\dagger$	9.08	$80.1 \pm 4.2^\dagger$
WFC	75.2 ± 0.1	91.4 ± 0.3	4.29	86.9 ± 0.2
WFC _{eo}	75.1 ± 0.1	91.0 ± 0.8	4.39	85.9 ± 0.2
WFC + BTEO	75.3 ± 0.1	91.1 ± 0.3	4.21	87.2 ± 0.5

Table 1: Results on Moji. For baselines, results are drawn from Shen et al. (2022b). We report the mean \pm standard deviation over 5 runs. * indicates the model without fairness consideration, and - indicates that we cannot access the result. The best results are in bold, results in blue indicate the best results without fine-tuning BERT. \dagger indicates statistical significant difference with WFC based on the Student’s t-test.

7.1 Comparison with State-of-the-Art Methods

Firstly, we compare our approach with state-of-the-art methods and different text encoders.

Baselines First, we consider *CE, that is, the architecture without any form of regularization. Then, the baselines are INLP (Ravfogel et al., 2020), the ADV method (Han et al., 2021b), FairBatch (Roh et al., 2021), GATE (Han et al., 2021c), EO_{GLB} (Shen et al., 2022a) and Con, displaying the *dp* and *eo* versions (Shen et al., 2022b). Shen et al. (2022b) extend some of the methods by rebalancing classes during training (+ BTEO) or fine-tuning a BERT model in addition to the trainable MLP (+ BERT_{ft}). We also consider DAFair (Iskander et al., 2024) in our baselines due to the proximity with our work as indicated in Section 2, and rerun their experiments with similar settings as the authors on respective data sets to ensure comparable results with regards to the splits and seeds. If not mentioned otherwise, the results of the other baselines are drawn from Han et al. (2022) and Shen et al. (2022b). We did not rerun it since we built our code on the Fairlib library made available by the authors. As such, the base architectures, evaluation protocols, seeds, and data splits are similar.

Setting To compare our method against state-of-the-art approaches, we first use the representation generated by a base BERT model as an input to the MLPs. For Bios, the

Model	Bal. Acc. \uparrow	Fairness \uparrow	DTO \downarrow	Leakage \downarrow
*CE	82.3 ± 0.2	$85.1 \pm 0.8^\dagger$	5.87	$98.0 \pm 0.0^\dagger$
INLP + BERT _{ft}	82.3 ± 0.0	$88.6 \pm 0.0^\dagger$	2.61	$97.6 \pm 0.1^\dagger$
Adv + BERT _{ft}	$81.9 \pm 0.2^\dagger$	$90.6 \pm 0.5^\dagger$	1.81	$88.6 \pm 4.6^\dagger$
Gate + BTEO + BERT _{ft}	$83.7 \pm 0.2^\dagger$	$90.4 \pm 0.9^\dagger$	0.40	$100.0 \pm 0.0^\dagger$
FairBatch + BERT _{ft}	$82.2 \pm 0.1^\dagger$	89.5 ± 1.3	1.98	$98.0 \pm 0.3^\dagger$
EO _{GLB} + BERT _{ft}	$81.7 \pm 0.4^\dagger$	88.4 ± 1.0	3.12	97.2 ± 0.5
DAFair + BERT _{ft}	$83.7 \pm 0.1^\dagger$	$86.4 \pm 0.3^\dagger$	4.40	-
Adv	$81.1 \pm 0.1^\dagger$	$87.3 \pm 0.9^\dagger$	4.36	-
Gate + BTEO	$79.4 \pm 0.1^\dagger$	$90.8 \pm 0.2^\dagger$	4.30	-
Con _{dp}	$82.1 \pm 0.2^\dagger$	$84.3 \pm 0.8^\dagger$	6.69	$76.3 \pm 1.5^\dagger$
Con _{eo}	$81.8 \pm 0.3^\dagger$	$85.2 \pm 0.4^\dagger$	5.91	$84.9 \pm 3.4^\dagger$
WFC	82.4 ± 0.1	89.0 ± 0.3	2.22	96.5 ± 0.5
WFC _{eo}	$82.1 \pm 0.2^\dagger$	89.0 ± 0.2	2.42	$97.4 \pm 0.3^\dagger$
WFC + BTEO	82.3 ± 0.2	89.1 ± 0.3	2.20	96.7 ± 0.5

Table 2: Results on Bios. For baselines, results are drawn from Shen et al. (2022b). We report the mean \pm standard deviation over 5 runs. * indicates the model without fairness consideration, - indicates that we do not have access to these results. The best results are in bold, results in blue indicate the best results without fine-tuning BERT. \dagger indicates statistical significant difference with WFC based on the Student’s t-test.

demonic MLP is trained on 1% of the training set and obtains 99% accuracy for predicting the sensitive attributes on the test set. Similarly, the *demonic* MLP obtains 88.5% accuracy on Moji. Except for the standard cross-entropy loss without a fairness constraint (CE) and the DAFair baseline, which we run ourselves, we report results from Shen et al. (2022b); Han et al. (2022) as mentioned in §Baselines. In our approach, embedding representations are derived from a fixed BERT model, with only the MLP weights being adjusted. We also evaluate the quality of our method under balanced training as in Shen et al. (2022b).

Discussion We compare WFC with text classification baselines. For Moji (Table 1), the accuracy of WFC is higher than the accuracy of CE, and it is equivalent to competitors. Considering the fairness metrics, we outperform all baselines. Note that DAFair, related to our work with the KL-divergence as dependency measure, outperforms all baselines in terms of accuracy with a limited gain of Fairness. For Bios (Table 2), our method is competitive with the other baselines and ranks 4 out of 12 with BTEO and 5 without it in terms of accuracy-fairness trade-off (DTO). Especially, WFC has the second-best accuracy compared to baselines. Moreover, on both data sets, we obtain similar Leakage as the comparable baselines, we further discuss this score in Section 7.3.1.

Note that BERT is not fine-tuned during our training pipeline. This decision is based on several factors: first, fine-tuning BERT increases training complexity and may hinder convergence. Additionally, it makes our method flexible to any encoder or decoder architec-

Model	Bal. Acc. \uparrow	Fairness \uparrow	DTO \downarrow	Leakage \downarrow
*CE	$85.5 \pm 0.09^\dagger$	$86.1 \pm 0.36^\dagger$	6.63	97.9 ± 0.41
GATE	85.3 ± 0.23	$83.5 \pm 0.60^\dagger$	9.22	$100.0 \pm 0.01^\dagger$
GATE + BTEO	$84.4 \pm 0.14^\dagger$	$92.7 \pm 0.67^\dagger$	1.10	$99.9 \pm 0.13^\dagger$
ADV	84.8 ± 0.72	90.3 ± 0.40	2.49	$89.1 \pm 7.96^\dagger$
ADV + BTEO	$84.3 \pm 0.07^\dagger$	$91.4 \pm 0.41^\dagger$	1.74	$86.2 \pm 6.05^\dagger$
WFC	85.2 ± 0.02	90.0 ± 0.21	2.74	97.8 ± 0.41
WFC + BTEO	$85.1 \pm 0.06^\dagger$	90.0 ± 0.25	2.75	97.8 ± 0.34

Table 3: SFR-Embeddings-2_R Results Bios. We report the mean \pm standard deviation over 5 runs. * indicates the model without fairness consideration. The best results are in bold. \dagger indicates statistical significant difference with WFC based on the Student’s t-test.

ture, regardless of size. However, among the baselines without BERT fine-tuning, we reach the lowest DTO, comparable to those obtained with methods that fine-tune BERT.

When comparing the versions of WFC optimizing EO or DP and rebalancing classes, we report close results on the three approaches. Noting a slightly better DPO on the version optimizing DP (WFC), we consider this version in the other experiments. Despite the better DTO of WFC + BTEO, we do not choose it for the experiments in Sections 7.2 and 7.3 to evaluate the method without external influence.

Ultimately, compared to the baselines, our method demonstrates notable advantages, particularly its ability to achieve competitive performance without access to sensitive attributes in the training set. We assess this capability in the section 7.2. In the next subsection, we explore an alternative model for generating the representations used by the classifier.

7.1.1 USING RECENT DECODER-BASED MODEL

Setting State-of-the-art baselines use BERT representations. However, recent PLMs have surpassed BERT’s performance. Additionally, many modern embedding models are based on a decoder architecture. Therefore, we assess the robustness of our method using representations from SFR-Embedding-2_R model³ (Meng et al., 2024) built on the Mistral model (Jiang et al., 2023). This model is ranked first on the MTEB benchmark⁴ (Muennighoff et al., 2022) on July 8th, 2024, notably for the classification task. We realize this set of experiments on the Bios data set and exclude the Moji data set since we do not have access to the raw text, and that the embeddings depend on the DeepMoji model (Felbo et al., 2017). The *demonic* MLP is also trained on SFR-Embedding-2_R’s representations. We compare our approach to the cross-entropy without regularization (CE), as well as the best baselines on BERT concerning fairness and accuracy (respectively, GATE and ADV). The approaches are evaluated with and without balanced training (BTEO). We realize hyperparameter tuning for all methods as described in Appendix F.3.

3. https://huggingface.co/Salesforce/SFR-Embedding-2_R

4. <https://huggingface.co/spaces/mteb/leaderboard>

Data	Bal. Acc. \uparrow	Fairness \uparrow	DTO \downarrow	Leakage \downarrow	Demonic Bal. Acc. \uparrow
Bios 1%	82.4 \pm 0.1	89.0 \pm 0.3	2.22	96.5 \pm 0.5	99.0
EEC	82.2 \pm 0.4	88.9 \pm 0.4	2.42	97.5 \pm 0.3	98.1
MP	82.4 \pm 0.3	88.9 \pm 0.4	2.30	96.4 \pm 0.5	98.4

Table 4: Comparison between several scenarios for training the *demonic* model for prediction on Bios. We report the mean \pm standard deviation over 5 runs.

Discussion We evaluate the efficiency of our architecture on recent decoder-based models to generate the embedding representations and compare them with the best baselines on the BERT-encoding results. We perform this evaluation on the Bios data set as explained above and present results in Table 3. We observe an improvement of both accuracy and fairness for all methods compared to the results with a BERT encoder. However, in this experiment, improving fairness comes at the cost of performance compared to the model without regularization (*CE). Among all baselines, ours enhances fairness while minimizing performance the less. In contrast, other baselines that improve fairness (GATE + BTEO, ADV, and ADV + BTEO) lead to a performance drop of up to one point.

7.2 Cross-domain WFC

We consider two experiments to assess the transfer of sensitive attributes: with and without the domain adaptation procedure. We conduct these experiments on Bios, as other data sets with gender annotations are already available, unlike AAE/SAE data sets for Moji.

The main objective of this section is to evaluate the performance of WFC when the *demonic* is trained on other sources than the task data set.

7.2.1 ZERO-SHOT CROSS-DOMAIN DEMONIC TRAINING

Setting We consider two source data sets to train the *demonic* MLP without domain adaptation. The EEC data set (Kiritchenko and Mohammad, 2018) consists of 8,640 synthetic sentences in English for Sentiment Analysis. The Marked Personas (MP) data set (Cheng et al., 2023) is composed of 2,700 descriptions of individuals obtained using a generative procedure: we consider the dv2 version. We then evaluate the WFC pipeline with those *demonic* MLP. When training on the EEC data set, we obtain, on average over 5 runs, 98.1% of accuracy, and 98.4% on the MP data set.

Discussion Table 4 shows that when the source and target data sets are similar, we achieve results comparable to those obtained when pre-training is performed using the same data set. The average loss in accuracy and fairness is minimal, with the standard deviation causing the measurements to overlap. These results are promising for improving fairness, especially in situations where collecting sensitive data is not feasible or when only partial information is available. In the next subsection, we investigate when the divergence between the source and target is higher and consider domain adaptation to train the *demonic* model.

Method	Accuracy on \mathcal{S}	Accuracy on \mathcal{T}
baseline	65.3 ± 3.23	75.3 ± 13.9
$\eta = 0.5$	75.0 ± 0.00	96.5 ± 0.94
$\eta = 1$	81.3 ± 0.67	98.0 ± 0.24
$\eta = 2$	75.0 ± 0.00	95.9 ± 0.27

Table 5: Performance of the *demonic* model trained with domain adaptation. Performance on the source is given for the best corresponding performance on the target set.

7.2.2 DEMONIC TRAINING WITH DOMAIN ADAPTATION

Setting and protocol We begin by considering a variant of the MP data set for this experiment. A set of gendered words (listed in Appendix F.3.4) is removed from the texts to increase the divergence with the Bios data set. Next, we train a *demonic* model on this data set with the values of regularization $\eta \in \{0.5, 1, 2\}$ on the domain adaptation term in Equation 6 recalled below

$$\arg \min \mathcal{L}(A_{\mathcal{S}}, h_a(Enc(X_{\mathcal{S}})) + \eta W_1(Z_{\mathcal{S}}, Z_{\mathcal{T}}).$$

We run the pipeline for 15000 epochs; at each epoch, the critic is trained on 20 batches and the model on 5 batches. We assess different values for the learning rate on a validation set and obtain the following optimal learning rate: $1e^{-5}$. We also compare to the baseline, which consists of training the *demonic* for 20 epochs on the source data set only, without any adaptation. For the baseline, the *demonic* model is optimized with the following objective

$$\arg \min \mathcal{L}(a, h_a(z_{source}))$$

Finally, we run the WFC pipeline with the *demonic* obtained as in the previous set of experiments.

Cross-domain demonic performance As shown in Table 5, the domain adaptation procedure significantly improves the performance of the *demonic* model on the sensitive attributes predictions when the domains diverge. Note that the value of η matters; with a lower η , the adaptation may be too weak to align the domains, whereas with a higher η , the regularization term may overly influence the classification term in the loss.

Interestingly, for the case of gender, when the most common expressions of gender are removed from the source but remain in the target domain, the procedure also helps to improve the performance of the *demonic* model on the source domain. Furthermore, we note better scores on the target domain than on the source. We hypothesize that the presence of the gender markers in the target domain makes the task easier, inducing better results.

Finally, it is interesting to note the variance on the *baseline demonic*: while in some cases domain adaptation will not be necessary, the procedure ensures an efficient *demonic* model without regard to the initial conditions of the optimization.

Model	Bal. Acc. \uparrow	Fairness \uparrow	DTO \downarrow	Leakage \downarrow	Demonic accuracy \uparrow
*CE	82.3 ± 0.20	85.1 ± 0.80	5.87	98.0 ± 0.00	-
Baseline	82.5 ± 0.05	86.8 ± 0.50	4.19	97.1 ± 0.44	75.3 ± 13.9
$\eta = 0.5$	82.5 ± 0.02	87.4 ± 0.21	3.57	96.6 ± 0.36	96.5 ± 0.94
$\eta = 1.0$	82.4 ± 0.09	88.7 ± 0.47	2.50	96.7 ± 0.31	98.0 ± 0.24
$\eta = 2.0$	82.5 ± 0.06	87.2 ± 0.15	3.79	96.7 ± 0.10	95.9 ± 0.27

Table 6: Results on Bios with a *demonic* trained with domain adaptation. We report the mean \pm standard deviation over 5 runs. * indicates the model without fairness consideration.

WFC results with cross-domain demonic Table 6 reports the results of the WFC pipeline on the Bios data set when using domain adaptation during the *demonic* training. We note that thanks to the improvement of the accuracy of the *demonic* model, the fairness on the end-task is improved compared to both the pipeline without fairness consideration and the pipeline where the *demonic* model is trained without adaptation. With domain adaptation, the improved performance of the *demonic* model is reflected in the enhanced fairness. This experiment highlights the importance of an accurate *demonic* model and the advantages of considering domain adaptation when training it on data sets diverging from the end-task data set.

7.3 WFC Architecture Components Investigation

In this section, we first investigate the hyperparameter β , its impact on the representation fairness (Leakage), and on fairness metrics related to the bounds from Lemma 1. Next, we study the use of the representations from different layers in the two MLPs (classifier and demonic models). Finally, we explore the use of the predicted sensitive attributes instead of the hidden representations of the demonic model as previously done.

7.3.1 IMPACT OF THE HYPERPARAMETER β

Setting In this experiment, we investigate the impact of the hyperparameter β associated with the regularization term. Recall that our objective is the following

$$\arg \min \mathcal{L}(Y, h_y(Enc(X_y))) + \beta I_W(Z_y, Z_a),$$

where β controls the impact of the Wasserstein Dependency Measure on the loss. We train the model over 5 seeds for different values of β . Specifically, $\beta \in \{0.1, 1, 2, 5, 10, 100\}$.

Discussion First, we note in Table 7 that with a higher β , the Leakage decreases, meaning the sensitive attribute is harder to retrieve from the latent representations. Although we initially aim to improve the Fairness while maintaining the Accuracy of the model, our method can be used to improve the Leakage by increasing the value of β in Equation 4. In other words, we give more importance to the Wasserstein regularization in the loss; as observed in Figure 2 where increasing the importance of the regularization term allows

β	Acc. \uparrow	Fair. \uparrow	DTO \downarrow	Leak. \downarrow	Acc. \uparrow	Fair. \uparrow	DTO \downarrow	Leak. \downarrow
	Bios				Moji			
0.1	82.8 \pm 0.1	87.2 \pm 0.4	3.75	98.1 \pm 0.2	50.4 \pm 0.7	99.5 \pm 1.1	27.08	85.7 \pm 0.1
1.0	82.4 \pm 0.1	89.0 \pm 0.3	2.22	96.7 \pm 0.5	75.2 \pm 0.1	91.4 \pm 0.3	1.00	86.9 \pm 0.2
5.0	81.8 \pm 0.2	88.9 \pm 0.2	2.69	91.8 \pm 1.4	71.4 \pm 0.5	93.7 \pm 0.4	5.38	81.1 \pm 0.5
10.0	81.6 \pm 0.2	88.6 \pm 0.2	3.06	86.1 \pm 0.8	70.1 \pm 0.6	92.7 \pm 0.4	6.21	82.5 \pm 0.8
100.0	81.2 \pm 0.4	87.9 \pm 0.4	3.84	77.7 \pm 1.7	67.9 \pm 1.4	94.7 \pm 1.1	8.9	83.0 \pm 0.5

Table 7: Study of the impact of β . We report the mean \pm standard deviation over 5 runs.

having a lower $I_W(Z_y, Z_a)$.⁵ However, on both data sets, the Accuracy that we want to preserve decreases and the trade-off worsens as the Leakage gets better. In other words, reducing the Leakage makes it more challenging to retrieve sensitive attributes, but could result in unintended information loss needed for the classification task, affecting the performance. Ultimately, we want to enhance fairness while keeping a good performance, and this objective may not necessarily match with a strong Leakage improvement (Shen et al., 2022b). Indeed, Lohaus et al. (2022) show that training neural networks to satisfy fairness constraints (e.g., demographic parity) is often done by making the model more aware of the sensitive attributes, and that stronger fairness constraints make these attributes more recoverable from their internal states. As such, combining privacy and fairness remains an open challenge.

Finally, note that on the Moji data set, the performance for $\beta = 0.1$ is surprisingly low, this is due to the selection criterion used: the DTO. Indeed, when looking at the best results for this setting, we have an accuracy of 73 ± 0.0 for a fairness of 68.5 ± 0.0 . This can be explained by the fact that the fairness regularization term is too low to improve fairness on this data set, then the results for the best accuracy are close to the *CE*-baseline results (cf. Table 1). However, at initialization with an inaccurate classifier, the fairness is very high, thus the optimal trade-off is obtained with these values.

In the next subsection, we investigate the relation of the Wasserstein Dependency Measure between the latent representations with the fairness metrics for different values of β .

7.3.2 RELATION BETWEEN FAIRNESS METRICS AND THE REGULARIZATION TERM

In this section, we empirically show the validity of the bounds from Lemma 1 on two data sets, Bios and Moji.

Setting We train the *WFC* pipeline with different β values as in Section 7.3.1, and report in Figure 2, the $I_W(Z_y, Z_a)$ on the training data, and the EO (2a,b) and DP (2c,d) on the test set for every training epoch.

Discussion We note that the more the loss is constrained by the regularization term, the lower $I_W(Z_y, Z_a)$ is, as well as the fairness metrics. However, after a certain threshold value for β (5.0 in the experiments), the fairness metrics converge. Finally, while in most cases $I_W(Z_y, Z_a)$ is greater than the considered metrics, as expected from Lemma 1 and Theorems 3 and 5, the contrary happens on a few epochs. This discrepancy with the results

5. Note that the values are computed exactly using the POT library (Flamary et al., 2021).



Figure 2: $I_W(Z_y, Z_a)$ and averaged fairness metrics over classes across training epochs. The values are averaged over 5 runs.

expected from the theoretical relation arises because we only plot $I_W(Z_y, Z_a)$ rather than the full right-hand term.

7.3.3 USE OF REPRESENTATIONS FROM DIFFERENT LAYERS

Setting In the previous experiments, following approaches presented in Han et al. (2022), the Wasserstein distance is approximated using the last hidden representations of the 3-layer MLP. In this section, we explore the use of representations from different layers of the MLPs. We compare this approach, on both data sets, with the use of the first hidden representations of the MLPs and with the output logits (before argmax), shown in Figure 3. For the latter, the Wasserstein distance is estimated between distributions of different dimensions. For example, for Bios, the *demonic* MLP predicts 2 labels while the classification MLP predicts 28 labels.

Layer	Bal. Acc. \uparrow	Fairness \uparrow	DTO \downarrow	Leakage \downarrow
Bios				
Last hidden	$82.4 \pm 0.1^*$	$89.0 \pm 0.3^*$	2.06^*	96.5 ± 0.5
First hidden	81.9 ± 0.2	86.7 ± 0.4	4.29	96.5 ± 0.6
Output layer	82.1 ± 0.6	87.5 ± 0.3	3.49	$87.0 \pm 1.1^*$
Moji				
Last hidden	$75.2 \pm 0.1^*$	$91.4 \pm 0.3^*$	1.17^*	86.9 ± 0.2
First hidden	74.3 ± 0.1	80.8 ± 1.0	11.4	85.6 ± 0.6
Output layer	73.5 ± 0.0	70.2 ± 0.2	21.9	$64.5 \pm 0.1^*$

Table 8: Comparison between the use of representations of different MLP layers to compute the Wasserstein.

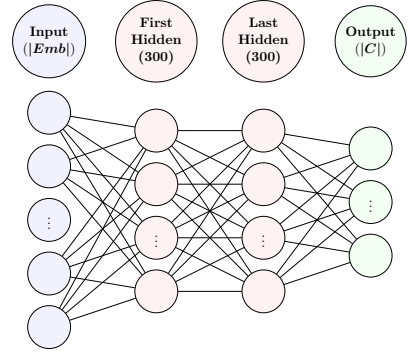


Figure 3: Representation of the MLP's layer.

Labels	Bal. Acc. \uparrow	Fairness \uparrow	DTO \downarrow	Leakage \downarrow
Bios				
Representations	82.4 ± 0.1	$89.0 \pm 0.3^*$	2.06^*	96.5 ± 0.5
Hard labels	82.6 ± 0.2	87.5 ± 0.2	3.28	$92.0 \pm 0.2^*$
Moji				
Representations	$75.2 \pm 0.1^*$	$91.4 \pm 0.3^*$	1.17^*	86.9 ± 0.2
Hard labels	72.2 ± 0.1	65.0 ± 0.0	27.3	$81.0 \pm 0.8^*$

Table 9: Comparison between the use of representations Z_a and hard sensitive attributes to compute the Wasserstein distance.

Discussion On both data sets (Table 8), accuracy is rather stable regardless of the layers used to compute the Wasserstein distance. Still, the best results are obtained using the last hidden representations. However, while we note a slight decrease in fairness on Bios when using representations from other layers, the decrease becomes much more significant on Moji. Thus, using the last hidden layer is the best strategy.

7.3.4 INDEPENDENCE WITH PREDICTED HARD SENSITIVE ATTRIBUTES

Setting To assess the impact of using the representation Z_a , we replace Z_a with the sensitive attributes predicted by the *demonic* MLP, \hat{A} . We consider the setting using the embeddings from BERT, with the Bios and Moji data sets. Then, we obtain the following regularization term: $I_W(Z_y, \hat{A}) = W_1(p(Z_y, \hat{A}), p(Z_y)p(\hat{A}))$. Note that we do not encounter a problem with the non-differentiability for \hat{A} (with the argmax operation as for \hat{Y} as mentioned in Section 4.3) since the *demonic* model is pre-trained.

Discussion We report the results of this experiment in Table 9. When we replace Z_a by the predicted \hat{A} to compute the Wasserstein distance, we observe, on average, a slight improvement of the accuracy on Bios, and a slight decrease of the accuracy on Moji. However, while the decrease in Fairness is not significant for Bios, we observe a substantial drop for

Moji. As a result, using \hat{A} instead of Z_a seems to have a neutral impact at best; this may also result, in some cases, in a reduction of both accuracy and fairness.

8. Conclusion

We extend WFC, a method enforcing fairness constraints using a pre-trained neural network on the sensitive attributes and Wasserstein regularization. We show that minimizing the Wasserstein Dependency Measure (I_W) improves fairness by reducing the statistical dependence between predictions and sensitive attributes, linking it to key metrics such as Demographic Parity and Equality of Opportunity.

Instead of directly optimizing I_W between predictions and sensitive attributes, we apply it to the latent representations of two models: one predicting classification labels and the other sensitive attributes. We prove that this formulation provides an upper bound on the dependency measure between predictions and true sensitive attributes while ensuring computational feasibility. Specifically, the I_W between latent representations upper-bounds the I_W between predicted labels and sensitive attributes, which in turn upper-bounds the I_W between predictions and true sensitive attributes. Our method does not require sensitive attribute annotations at both training and inference time. We obtain competitive results on the Bios data set and outperform baselines on fairness metrics while maintaining comparable accuracy on the Moji data set. The approach is also compatible with both encoder-based and decoder-based architectures.

We also extend our method to settings where sensitive attributes are unavailable, leveraging a domain adaptation approach to enable training under this constraint. We provide theoretical guarantees, inspired by domain adaptation results, to assess its generalization to other data sets.

Perspectives Overall, this work could be extended in numerous ways. The bound presented in Theorem 4 relies on a covariate shift assumption. However, in the literature on causal vs. anticausal learning (Schölkopf et al., 2012), many NLP tasks can be viewed as anticausal, where the target label (e.g., an occupation) generates the observed text (e.g., the description of the occupation) (Jin et al., 2021). In such cases, a prior probability shift assumption, where the label distribution changes across domains but the feature distribution conditioned on the label remains invariant, would arguably be a natural setting to consider. When the sensitive attribute itself is considered as the label (e.g., gender), as is the case for our demonic, one could also argue for an anticausal relation, but in a weaker sense: the sensitive attribute influences some parts of the text (e.g., pronouns, names) without fully determining its content. Studying fairness-aware transfer under such partial forms of anticausality would be an interesting direction for future work.

In Theorem 5, we discussed the use of an optimal softmax temperature to obtain a sharper bound. It would be interesting to investigate whether having such a given temperature is an artifact of our proof technique or whether it is a requirement. Furthermore, if it turns out to be necessary, a possible avenue of research would be to draw connections with other research domains that rely on temperature scaling. Calibration is a prime example of such a field where it is a popular baseline that has been shown to perform well in different settings (Park et al., 2020; Wang et al., 2020; Chen and Su, 2023; Guo et al., 2017). Hence,

it would be interesting to study in which cases the temperature derived in Theorem 5 would positively or negatively impact the level of calibration of the model.

Finally, although we did not explore this direction, the approach could be extended beyond text classification to tasks such as regression or unsupervised learning, or to other types of data such as images.

9. Limitations

The proposed approach is flexible and can handle various types of sensitive attributes. However, due to the lack of available data sets, we were unable to evaluate its performance on continuous sensitive attributes, such as age. Additionally, while gender can be represented as an n -ary variable, our experiments were limited to a binary classification (men vs. women) due to data availability.

Finally, our experiments demonstrate the effectiveness of our approach in transferring sensitive attributes to improve fairness. However, our theoretical results indicate that the success of this transfer depends on its quality; a poor transfer could, in theory, lead to a decrease in fairness.

Acknowledgments

This work is funded by the French National Research Agency (ANR) in the context of the grant ANR-21-CE23-0026 (Project DIKÉ). Michael Perrot is supported by the ANR through the grant ANR-23-CE23-0011-01 (Project FaCTor). Charlotte Laclau is supported by the ANR through the grant ANR-23-CE23-0026 (Project ReFAIR) and through the PEPR IA FOUNDRY (ANR-23-PEIA-0003). Our experiments use the previously mentioned Fairlib framework. We would like to express our gratitude to Xudong Han for his availability and assistance in using it.

Appendix A. Wasserstein Distance

Finding correspondences between two sets of points is a longstanding issue in machine learning. The optimal transport (OT) (Monge, 1781) problem offers an efficient solution to this issue by calculating an optimal one-to-one transport map between the two sets, taking into account the geometrical proximity of the points. Let $\hat{\mu}_0$ and $\hat{\mu}_1$ be measures supported on the point sets $X_0 = \{x_0^{(i)} \in \mathbb{R}^d\}_{i=1}^{N_0}$ and $X_1 = \{x_1^{(j)} \in \mathbb{R}^d\}_{j=1}^{N_1}$, respectively. We consider the Monge-Kantorovich formulation of the original OT problem (Kantorovich, 1942), where the goal is to find a coupling γ defined as a joint probability distribution over $X_0 \times X_1$ with marginals $\hat{\mu}_0$ and $\hat{\mu}_1$. This amounts to minimizing the cost of transport w.r.t. some metric $l_p = \|\cdot\|_p : X_0 \times X_1 \rightarrow \mathbb{R}^+$, the l_p -norm. This problem admits a unique solution γ^* and defines a metric on the space of probability measures called the Wasserstein distance (also known as the Earth-Mover Distance) as follows

$$W_1(\hat{\mu}_0, \hat{\mu}_1) = \min_{\gamma \in \Pi(\hat{\mu}_0; \hat{\mu}_1)} \langle M, \gamma \rangle_F,$$

where $\langle \cdot, \cdot \rangle_F$ is the Frobenius dot product, M is a dissimilarity matrix, i.e., $M_{ij} = l(x_0^{(i)}, x_1^{(j)})$, defining the cost of associating $x_0^{(i)}$ with $x_1^{(j)}$ and $\Pi(\hat{\mu}_0, \hat{\mu}_1) = \{\gamma \in \mathbb{R}_+^{N_0 \times N_1} | \gamma \mathbf{1} = \hat{\mu}_0, \gamma^T \mathbf{1} = \hat{\mu}_1\}$ is a set of doubly stochastic matrices.

In the following, we will rely on the following technical lemma on the Wasserstein distance between discrete distributions.

Lemma 6 *Let $U \sim p(U)$ and $V \sim p(V)$ be two discrete random variables respectively taking values in u_1, \dots, u_k and v_1, \dots, v_k . Assume that $\|u_i - v_j\|_p = \begin{cases} 0 & \text{if } i = j \\ \sqrt[p]{2} & \text{otherwise} \end{cases}$, then, we have that*

$$W_1(p(U), p(V)) = \frac{\sqrt[p]{2}}{2} \sum_{i=1}^k |\mathbb{P}(u_i) - \mathbb{P}(v_i)|$$

Proof From Gibbs and Su (2002, Theorem 4), we have that

$$\min_{u \neq v} \|u - v\|_p TV(p(U), p(V)) \leq W_1(p(U), p(V)) \leq \max_{u, v} \|u - v\|_p TV(p(U), p(V)),$$

where $TV(p(U), p(V)) = \frac{1}{2} \sum_{i=1}^k |\mathbb{P}(u_i) - \mathbb{P}(v_i)|$ is the total variation. Noticing that, in our case, $\min_{u \neq v} \|u - v\|_p = \max_{u, v} \|u - v\|_p = \sqrt[p]{2}$ concludes the proof. ■

Lemma 7 *Let $U \sim p(U)$, $V \sim p(V)$, and $W \sim p(W)$ be discrete random variables taking values in \mathcal{U} , \mathcal{V} , and \mathcal{W} respectively and such that $\|u - u'\|_p = \|v - v'\|_p = \|w - w'\|_p = \begin{cases} 0 & \text{if } u = u', v = v' \text{ or } w = w' \\ \sqrt[p]{2} & \text{otherwise} \end{cases}$, then, we have that*

$$\begin{aligned} W_1(p(U, W), p(U)p(W)) &= \sum_w W_1(p(U|W=w), p(U)) \mathbb{P}(W=w) \\ W_1(p(U, W), p(V, W)) &= \sum_w W_1(p(U|W=w), p(V|W=w)) \mathbb{P}(W=w) \\ W_1(p(U)p(W), p(V)p(W)) &= \sum_w W_1(p(U), p(V)) \mathbb{P}(W=w) \end{aligned}$$

Proof The cost matrix associated with $W_1(p(U, W), p(U)p(W))$ is of size $|\mathcal{U}||\mathcal{W}| \times |\mathcal{U}||\mathcal{W}|$. Assuming that we order the pairs (u, w) by varying the values of u first, that is $(u_1, w_1), (u_2, w_1), \dots$, the cost matrix contains blocks of size $|\mathcal{U}| \times |\mathcal{U}|$. The diagonal blocks have value $\sqrt[p]{2}(\mathbb{1}_{|\mathcal{U}| \times |\mathcal{U}|} - \mathbb{I}_{|\mathcal{U}| \times |\mathcal{U}|})$ where $\mathbb{I}_{|\mathcal{U}| \times |\mathcal{U}|}$ is the identity matrix of size $|\mathcal{U}| \times |\mathcal{U}|$ and $\mathbb{1}_{|\mathcal{U}| \times |\mathcal{U}|}$ is a matrix of ones. The off diagonal blocks have value $\sqrt[p]{2}\mathbb{I}_{|\mathcal{U}| \times |\mathcal{U}|} + \sqrt[p]{4}(\mathbb{1}_{|\mathcal{U}| \times |\mathcal{U}|} - \mathbb{I}_{|\mathcal{U}| \times |\mathcal{U}|})$.

We have that $\forall w \in \mathcal{W}, \sum_u \mathbb{P}(U=u, W=w) = \mathbb{P}(W=w) = \sum_u \mathbb{P}(U=u) \mathbb{P}(W=w)$ which means that we can consider each diagonal block independently when computing $W_1(p(U, W), p(U)p(W))$, that is compute $\forall w, W_1(p(U|W=w), p(U))$ and then normalize the transport cost by $\mathbb{P}(W=w)$. This will be the optimal cost since the mass that is not transported with a cost of 0 will be transported with a cost of $\sqrt[p]{2}$, which is the smallest

possible cost different from 0. This concludes the proof of the first equality. The proofs of the second and third equality follow using the same arguments. \blacksquare

Appendix B. Connection with Group Fairness

The following lemma shows that minimizing the Wasserstein dependency measure is a sound way to improve either demographic parity or equality of opportunity.

Lemma 1 (Group fairness and Wasserstein dependency measure.) *Let I_W be the Wasserstein dependency measure, and A , Y , \hat{Y} be random variables corresponding to the sensitive attribute, the true label, and the predicted label, respectively. We have that*

$$I_W(\hat{Y}, A) = \frac{\sqrt[p]{2}}{2} \sum_{a \in \mathcal{A}} \mathbb{P}(A = a) \sum_{y \in \mathcal{Y}} \mathbf{DP}_{a,y},$$

$$I_W((\hat{Y} = Y)|Y = y, A|Y = y) = \sqrt[p]{2} \sum_{a \in \mathcal{A}} \mathbb{P}(A = a|Y = y) \mathbf{EO}_{a,y}.$$

Proof Let \hat{Y} and A be the two random variables corresponding to the predicted label and sensitive attribute, respectively. Recall that these random variables are encoded using a one-hot vector, that is $\|y_i - y_j\|_p = \begin{cases} 0 & \text{if } i = j \\ \sqrt[p]{2} & \text{otherwise} \end{cases}$ and $\|a_i - a_j\|_p = \begin{cases} 0 & \text{if } i = j \\ \sqrt[p]{2} & \text{otherwise} \end{cases}$. Then, by successively applying Lemma 7 and Lemma 6, we have that

$$\begin{aligned} I_W(\hat{Y}, A) &:= W_1(p(\hat{Y}, A), p(\hat{Y})p(A)) \\ &= \sum_{a \in \mathcal{A}} W_1(p(\hat{Y}|A = a), p(\hat{Y}))\mathbb{P}(A = a) \\ &= \sum_{a \in \mathcal{A}} \mathbb{P}(A = a) \frac{\sqrt[p]{2}}{2} \sum_{y \in \mathcal{Y}} \left| \mathbb{P}(\hat{Y} = y|A = a) - \mathbb{P}(\hat{Y} = y) \right| \\ &= \frac{\sqrt[p]{2}}{2} \sum_{a \in \mathcal{A}} \mathbb{P}(A = a) \sum_{y \in \mathcal{Y}} \left| \mathbb{P}(\hat{Y} = y|A = a) - \mathbb{P}(\hat{Y} = y) \right| \end{aligned}$$

Noticing that $\left| \mathbb{P}(\hat{Y} = y|A = a) - \mathbb{P}(\hat{Y} = y) \right|$ is the demographic parity for group a and label y concludes the proof of the first statement. Similarly, notice that given a label $y \in \mathcal{Y}$

$$\begin{aligned} I_W((\hat{Y} = Y)|Y = y, A|Y = y) &:= W_1(p((\hat{Y} = Y), A|Y = y), p((\hat{Y} = Y)|Y = y)p(A|Y = y)) \\ &= \sum_{a \in \mathcal{A}} W_1(p((\hat{Y} = Y)|A = a, Y = y), p((\hat{Y} = Y)|Y = y))\mathbb{P}(A = a|Y = y) \\ &= \sum_{a \in \mathcal{A}} \mathbb{P}(A = a|Y = y) \frac{\sqrt[p]{2}}{2} \left(\left| \mathbb{P}(\hat{Y} = Y|A = a, Y = y) - \mathbb{P}(\hat{Y} = Y|Y = y) \right| \right. \\ &\quad \left. + \left| \mathbb{P}(\hat{Y} \neq Y|A = a, Y = y) - \mathbb{P}(\hat{Y} \neq Y|Y = y) \right| \right) \\ &= \sqrt[p]{2} \sum_{a \in \mathcal{A}} \mathbb{P}(A = a|Y = y) \left| \mathbb{P}(\hat{Y} = Y|A = a, Y = y) - \mathbb{P}(\hat{Y} = Y|Y = y) \right| \end{aligned}$$

Noticing that $\left| \mathbb{P}(\hat{Y} = Y | A = a, Y = y) - \mathbb{P}(\hat{Y} = Y | Y = y) \right|$ is the Equality of opportunity for group a and label y concludes the proof of the second statement. \blacksquare

Appendix C. Bounding the $I_W(\hat{Y}, A)$ by the error rate

In this section, we provide the details of the proof of Lemma 2 leading to Theorems 3 and 4.

Lemma 2 *Let \hat{Y}, \hat{A}, A be random variables that correspond to the predicted label, predicted sensitive attribute, and true sensitive attribute, respectively. Then, we have that*

$$I_W(\hat{Y}, A) \leq I_W(\hat{Y}, \hat{A}) + 2\sqrt[2]{2}\mathbb{P}(A \neq \hat{A})$$

Proof Let \hat{Y}, \hat{A} and A be the random variables corresponding to the predicted label, predicted sensitive attribute, and true sensitive attribute, respectively. The Wasserstein Dependency Measure (Ozair et al., 2019) is

$$I_W(\hat{Y}, A) = W_1(p(\hat{Y}, A), p(\hat{Y})p(A)) .$$

The W_1 -metric can be shown to be a proper metric when the compared distributions have the same overall mass (Rubner et al., 2000). Therefore, it satisfies the triangle inequality

$$\begin{aligned} I_W(\hat{Y}, A) &:= W_1(p(\hat{Y}, A), p(\hat{Y})p(A)) \\ &\leq W_1(p(\hat{Y}, A), p(\hat{Y}, \hat{A})) \\ &\quad + W_1(p(\hat{Y}, \hat{A}), p(\hat{Y})p(\hat{A})) \\ &\quad + W_1(p(\hat{Y})p(\hat{A}), p(\hat{Y})p(A)), \end{aligned}$$

with $W_1(p(\hat{Y}, \hat{A}), p(\hat{Y})p(\hat{A})) = I_W(\hat{Y}, \hat{A})$.

Recall that \hat{Y}, \hat{A} and A are encoded using a one hot vector, that is $\|y_i - y_j\|_p = \begin{cases} 0 & \text{if } i = j \\ \sqrt[p]{2} & \text{otherwise} \end{cases}$ and $\|a_i - a_j\|_p = \begin{cases} 0 & \text{if } i = j \\ \sqrt[p]{2} & \text{otherwise} \end{cases}$. Then, by successively applying Lemma 7 and Lemma 6, we have that

$$\begin{aligned} W_1(p(\hat{Y}, A), p(\hat{Y}, \hat{A})) &= \sum_{y \in \mathcal{Y}} W_1(p(A|\hat{Y} = y), p(\hat{A}|\hat{Y} = y))\mathbb{P}(\hat{Y} = y) \\ &= \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{Y} = y) \frac{\sqrt[p]{2}}{2} \sum_{a \in \mathcal{A}} \left| \mathbb{P}(A = a|\hat{Y} = y) - \mathbb{P}(\hat{A} = a|\hat{Y} = y) \right| \quad (7) \end{aligned}$$

By the law of total probability and the union bound for disjoint events, we have that

$$\begin{aligned}
 & \sum_{a \in \mathcal{A}} \left| \mathbb{P}(A = a | \hat{Y} = y) - \mathbb{P}(\hat{A} = a | \hat{Y} = y) \right| \\
 &= \sum_{a \in \mathcal{A}} \left| \mathbb{P}(A = a, \hat{A} = a | \hat{Y} = y) + \mathbb{P}(A = a, \hat{A} \neq a | \hat{Y} = y) \right. \\
 &\quad \left. - \mathbb{P}(\hat{A} = a, A = a | \hat{Y} = y) - \mathbb{P}(\hat{A} = a, A \neq a | \hat{Y} = y) \right| \\
 &= \sum_{a \in \mathcal{A}} \left| \mathbb{P}(A = a, \hat{A} \neq a | \hat{Y} = y) - \mathbb{P}(\hat{A} = a, A \neq a | \hat{Y} = y) \right| \\
 &\leq \sum_{a \in \mathcal{A}} \mathbb{P}(A = a, \hat{A} \neq a | \hat{Y} = y) + \mathbb{P}(\hat{A} = a, A \neq a | \hat{Y} = y) \\
 &= \sum_{a \in \mathcal{A}} \mathbb{P}(A = a, \hat{A} \neq a | \hat{Y} = y) + \mathbb{P}(\hat{A} = a, A \neq a | \hat{Y} = y) \\
 &= \mathbb{P}\left(\bigcup_{a \in \mathcal{A}} A = a, \hat{A} \neq a | \hat{Y} = y\right) + \mathbb{P}\left(\bigcup_{a \in \mathcal{A}} \hat{A} = a, A \neq a | \hat{Y} = y\right) \\
 &= 2\mathbb{P}(A \neq \hat{A} | \hat{Y} = y)
 \end{aligned}$$

Plugging this result in Equation 7, we obtain

$$\begin{aligned}
 W_1(p(\hat{Y}), p(\hat{Y}, \hat{A})) &= \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{Y} = y) \sqrt[p]{2} \mathbb{P}(A \neq \hat{A} | \hat{Y} = y) \\
 &= \sqrt[p]{2} \mathbb{P}(A \neq \hat{A})
 \end{aligned}$$

Using similar arguments, we obtain that

$$\begin{aligned}
 W_1(p(\hat{Y})p(A), p(\hat{Y})p(\hat{A})) &= \sum_{y \in \mathcal{Y}} \mathbb{P}(\hat{Y} = y) \frac{\sqrt[p]{2}}{2} \sum_{a \in \mathcal{A}} \left| \mathbb{P}(A = a) - \mathbb{P}(\hat{A} = a) \right| \\
 &= \sqrt[p]{2} \mathbb{P}(A \neq \hat{A}).
 \end{aligned}$$

This concludes the proof of this lemma. ■

Built on this first result, we consider two scenarios to bound the error rate: either we pre-trained the *demonic* model on the data of the classification task (**in-domain**) or as a DA problem, on different data with shared sensitive attributes (e.g., gender, ethnicity, etc.) (**cross-domain**).

C.1 In-domain bound of the error rate for binary sensitive attributes

Theorem 3 *Let $\hat{A}, A \in \{0, 1\}$, and \mathcal{H} be a hypothesis space of VC-dimension d . Let $\|\cdot\|_p$ be the ground metric for the Wasserstein 1-distance. Assume that we have access to a training set of m i.i.d. examples. Then, with probability at least $1 - \delta$, we have $\forall h \in \mathcal{H}$*

$$I_W(\hat{Y}, A) \leq I_W(\hat{Y}, \hat{A}) + 2\sqrt[p]{2} \left(\hat{\varepsilon} + \sqrt{\frac{4}{m} \left(d \log \frac{2em}{d} + \log \frac{4}{\delta} \right)} \right)$$

with e , the base of the natural logarithm and $\hat{\varepsilon}$ the empirical risk of the demonic model.

Proof From Lemma 2, we derive the following

$$\begin{aligned} I_W(\hat{Y}, A) &\leq I_W(\hat{Y}, \hat{A}) + 2\sqrt[p]{2}\mathbb{P}(A \neq \hat{A}) \\ &= I_W(\hat{Y}, \hat{A}) + 2\sqrt[p]{2}\varepsilon \end{aligned}$$

We apply the Vapnik-Chervonenkis theory (Vapnik, 1998) to bound the true error ε of the demonic model by its empirical risk $\hat{\varepsilon}$. Let h_a be a fixed classification function from Z_a to A and \mathcal{H} be a hypothesis space of VC-dimension d . Therefore, if the training set is of size m i.i.d. samples, with probability at least $1 - \delta$, we have for every $h \in \mathcal{H}$

$$I_W(\hat{Y}, A) \leq I_W(\hat{Y}, \hat{A}) + 2\sqrt[p]{2} \left(\hat{\varepsilon} + \sqrt{\frac{4}{m} \left(d \log \frac{2em}{d} + \log \frac{4}{\delta} \right)} \right)$$

and e is the base of the natural logarithm. ■

Appendix D. Bounding $I_W(\hat{Y}, \hat{A})$ by $I_W(Z_y, Z_a)$

In this section, we present the proof for Theorem 5 recalled below.

Theorem 5 Let \hat{Y}, \hat{A} be random variables that correspond to the predicted label and predicted sensitive attribute, respectively. Assume that $h_y = \sigma_\lambda(f(Z_y))$ and $h_a = \sigma_\lambda(g(Z_a))$ where σ_λ is the softmax function with temperature λ , f and g are both L -lipschitz with respect to the p -norm, and Z_y and Z_a are latent representations of the examples. Let $\|\cdot\|_p$ be the ground metric for the Wasserstein 1-distance. For a given example x with predicted label \hat{y} and predicted sensitive attribute \hat{a} , let $\xi_y(x) = f(Z_y)_{\hat{y}} - \max_{y' \neq \hat{y}} f(Z_y)_{y'}$ and $\xi_a(x) = g(Z_a)_{\hat{a}} - \max_{a' \neq \hat{a}} g(Z_a)_{a'}$ be positive margins. Let $\delta = 1 - \mathbb{P}(\xi_y(X) \geq \xi, \xi_a(X) \geq \xi)$ with $\xi > 0$. Let $\alpha = \sqrt[p]{2} \left\| \binom{|\mathcal{Y}|}{|\mathcal{A}|} - 1 \right\|_p (1 - \delta)$ and $\iota = L(|\mathcal{Y}| + |\mathcal{A}|)^{\frac{1}{2} - \frac{1}{p}}$. Then, setting $\lambda = \frac{1}{\xi} \log \left(\frac{2\xi\alpha}{\iota I_W(Z_y, Z_a)} \right)$, we have that

$$\begin{aligned} I_W(\hat{Y}, \hat{A}) &\leq \min \left(\alpha, 2I_W(Z_y, Z_a) \frac{\iota}{\xi} \left[1 + \log \left(\max \left(4, \frac{2\xi\alpha}{\iota I_W(Z_y, Z_a)} \right) - 1 \right) \right] \right) \\ &\quad + \sqrt[p]{2} \left\| \binom{|\mathcal{Y}|}{|\mathcal{A}|} - 1 \right\|_p \delta. \end{aligned}$$

Proof Since, in our case, the Wasserstein distance is a proper metric, we have that

$$\begin{aligned} I_W(\hat{Y}, \hat{A}) &= W_1(p(\hat{Y}, \hat{A}), p(\hat{Y})p(\hat{A})), \\ &\leq W_1(p(\hat{Y}, \hat{A}), p(\sigma_\lambda(f(Z_y)), \sigma_\lambda(g(Z_a)))) \\ &\quad + W_1(p(\sigma_\lambda(f(Z_y)), \sigma_\lambda(g(Z_a))), p(\sigma_\lambda(f(Z_y)))p(\sigma_\lambda(g(Z_a)))) \\ &\quad + W_1(p(\sigma_\lambda(f(Z_y)))p(\sigma_\lambda(g(Z_a))), p(\hat{Y})p(\hat{A})). \end{aligned} \tag{8}$$

We will first bound each term independently and then show that we can choose the softmax temperature, λ , in order to minimize the right-hand side of the bound.

Bounding $W_1(p(\sigma_\lambda(f(Z_y)), \sigma_\lambda(g(Z_a))), p(\sigma_\lambda(f(Z_y)))p(\sigma_\lambda(g(Z_a))))$. Given $\gamma \in \Gamma$ a coupling between the two distributions, the second term can be bounded as

$$\begin{aligned}
 & W_1(p(\sigma_\lambda(f(Z_y)), \sigma_\lambda(g(Z_a))), p(\sigma_\lambda(f(Z_y)))p(\sigma_\lambda(g(Z_a)))) \\
 &= \inf_{\gamma} \mathbb{E}_{(z_y, z_a, z'_y, z'_a) \sim \gamma} \left\| (\sigma_\lambda(f(z_y)), \sigma_\lambda(g(z_a))) - (\sigma_\lambda(f(z'_y)), \sigma_\lambda(g(z'_a))) \right\|_p, \\
 &\leq \inf_{\gamma} \mathbb{E}_{(z_y, z_a, z'_y, z'_a) \sim \gamma} L\lambda(|\mathcal{Y}| + |\mathcal{A}|)^{\left|\frac{1}{2} - \frac{1}{p}\right|} \left\| (z_y, z_a) - (z'_y, z'_a) \right\|_p, \\
 &\leq L\lambda(|\mathcal{Y}| + |\mathcal{A}|)^{\left|\frac{1}{2} - \frac{1}{p}\right|} W_1(p(Z_y, Z_a), p(Z_y)p(Z_a)), \\
 &= L\lambda(|\mathcal{Y}| + |\mathcal{A}|)^{\left|\frac{1}{2} - \frac{1}{p}\right|} I_W(Z_y, Z_a).
 \end{aligned}$$

where the first inequality comes from the λ -lipschitzness of the softmax function ℓ_2 -norm (Gao and Pavel, 2017) and equivalence of norms properties.

Bounding $W_1(p(\hat{Y}, \hat{A}), p(\sigma_\lambda(f(Z_y)), \sigma_\lambda(g(Z_a))))$ and $W_1(p(\sigma_\lambda(f(Z_y)))p(\sigma_\lambda(g(Z_a))), p(\hat{Y})p(\hat{A}))$.

Let the softmax function $\sigma_\lambda(f(z)) = \frac{e^{\lambda f(z)}}{\|e^{\lambda f(z)}\|_1}$ for z a vector representation of an example x and $\lambda \geq 0$ the temperature. Then, we have that

$$\begin{aligned}
 W_1(p(\hat{Y}, \hat{A}), p(\sigma_\lambda(f(Z_y)), \sigma_\lambda(g(Z_a)))) &= W_1(p(\hat{Y})p(\hat{A}), p(\sigma_\lambda(f(Z_y)))p(\sigma_\lambda(g(Z_a)))) \\
 &= \mathbb{E} c(X, X).
 \end{aligned}$$

Indeed, for an example x represented as z_y, z_a and with predictions \hat{y} and \hat{a} and an example x' represented as z'_y, z'_a and with predictions \hat{y}' and \hat{a}' the cost matrix c is such that:

$$c(x, x') = \left\| (\hat{y}, \hat{a})^\top - \left(\frac{e^{\lambda f(z'_y)}}{\|e^{\lambda f(z'_y)}\|_1}, \frac{e^{\lambda f(z'_a)}}{\|e^{\lambda f(z'_a)}\|_1} \right)^\top \right\|_p. \text{ Thus, the minimal cost is achieved when}$$

each example is mapped onto itself since the predictions are obtained by taking the labels and sensitive attributes predicted as being most likely. We then have that

$$\begin{aligned}
 c(x, x) &= \left[\left(1 - \frac{e^{\lambda f(z_y)_{\hat{y}}}}{\|e^{\lambda f(z_y)}\|_1} \right)^p + \left(\frac{\sum_{y' \neq \hat{y}} e^{\lambda f(z_y)_{y'}}}{\|e^{\lambda f(z_y)}\|_1} \right)^p + \left(1 - \frac{e^{\lambda g(z_a)_{\hat{a}}}}{\|e^{\lambda g(z_a)}\|_1} \right)^p + \left(\frac{\sum_{a' \neq \hat{a}} e^{\lambda g(z_a)_{a'}}}{\|e^{\lambda g(z_a)}\|_1} \right)^p \right]^{\frac{1}{p}}, \\
 &= \left[2 \left(\frac{\sum_{y' \neq \hat{y}} e^{\lambda f(z_y)_{y'}}}{\|e^{\lambda f(z_y)}\|_1} \right)^p + 2 \left(\frac{\sum_{a' \neq \hat{a}} e^{\lambda g(z_a)_{a'}}}{\|e^{\lambda g(z_a)}\|_1} \right)^p \right]^{\frac{1}{p}}.
 \end{aligned}$$

For a given example x with predicted label \hat{y} and predicted sensitive attribute \hat{a} , let $\xi_y(x) = f(z_y)_{\hat{y}} - \max_{y' \neq \hat{y}} f(z_y)_{y'}$ and $\xi_a(x) = g(z_a)_{\hat{a}} - \max_{a' \neq \hat{a}} g(z_a)_{a'}$ be positive margins.

We note $m_y = f(z_y)_{\hat{y}}$ and $m_a = f(z_a)_{\hat{a}}$. Then, we have that

$$\begin{aligned} c(x, x) &\leq \left[2 \left(\frac{(|\mathcal{Y}| - 1)e^{\lambda(m_y - \xi_y)}}{e^{\lambda m_y} + e^{\lambda(m_y - \xi_y)}} \right)^p + 2 \left(\frac{(|\mathcal{A}| - 1)e^{\lambda(m_a - \xi_a)}}{e^{\lambda m_a} + e^{\lambda(m_a - \xi_a)}} \right)^p \right]^{\frac{1}{p}}, \\ &\leq \left[2 \left(\frac{(|\mathcal{Y}| - 1)e^{\lambda m_y}}{e^{\lambda \xi_y} e^{\lambda m_y} + e^{\lambda m_y}} \right)^p + 2 \left(\frac{(|\mathcal{A}| - 1)e^{\lambda m_a}}{e^{\lambda \xi_a} e^{\lambda m_a} + e^{\lambda m_a}} \right)^p \right]^{\frac{1}{p}}, \\ &\leq \left[2 \left(\frac{|\mathcal{Y}| - 1}{e^{\lambda \xi_y} + 1} \right)^p + 2 \left(\frac{|\mathcal{A}| - 1}{e^{\lambda \xi_a} + 1} \right)^p \right]^{\frac{1}{p}}. \end{aligned}$$

Let $\delta = 1 - \mathbb{P}(\xi_y(X) \geq \xi, \xi_a(X) \geq \xi)$ with $\xi > 0$, then we have that

$$\begin{aligned} \mathbb{E} c(X, X) &= \mathbb{E} [c(X, X) | \xi_y(x) \geq \xi, \xi_a(x) \geq \xi] (1 - \delta) + \mathbb{E} [c(X, X) | \overline{\xi_y(x) \geq \xi, \xi_a(x) \geq \xi}] \delta, \\ &\leq \left[2 \frac{(|\mathcal{Y}| - 1)^p + (|\mathcal{A}| - 1)^p}{(e^{\lambda \xi} + 1)^p} \right]^{\frac{1}{p}} (1 - \delta) + \left[2 \frac{(|\mathcal{Y}| - 1)^p + (|\mathcal{A}| - 1)^p}{(2)^p} \right]^{\frac{1}{p}} \delta, \\ &\leq \sqrt[p]{2} \frac{\left\| \begin{pmatrix} |\mathcal{Y}| \\ |\mathcal{A}| \end{pmatrix} - 1 \right\|_p}{e^{\lambda \xi} + 1} (1 - \delta) + \frac{\sqrt[p]{2}}{2} \left\| \begin{pmatrix} |\mathcal{Y}| \\ |\mathcal{A}| \end{pmatrix} - 1 \right\|_p \delta. \end{aligned}$$

Optimizing the softmax temperature. Our goal is to minimize the right hand side of Equation (8), we need to solve

$$\arg \inf_{\lambda} \frac{2 \sqrt[p]{2} \left\| \begin{pmatrix} |\mathcal{Y}| \\ |\mathcal{A}| \end{pmatrix} - 1 \right\|_p}{e^{\lambda \xi} + 1} (1 - \delta) + \lambda L(|\mathcal{Y}| + |\mathcal{A}|)^{\left| \frac{1}{2} - \frac{1}{p} \right|} I_W(Z_y, Z_a).$$

Let $\alpha = \sqrt[p]{2} \left\| \begin{pmatrix} |\mathcal{Y}| \\ |\mathcal{A}| \end{pmatrix} - 1 \right\|_p (1 - \delta)$ and $\beta = L(|\mathcal{Y}| + |\mathcal{A}|)^{\left| \frac{1}{2} - \frac{1}{p} \right|} I_W(Z_y, Z_a)$ which are both positive values, then we consider

$$\arg \inf_{\lambda} \frac{2\alpha}{e^{\lambda \xi} + 1} + \lambda \beta.$$

Let $\gamma = \lambda \xi$, since $\xi > 0$ and $\alpha > 0$ then,

$$\arg \inf_{\lambda} \frac{2\alpha}{e^{\lambda \xi} + 1} + \lambda \beta = \frac{1}{\xi} \arg \inf_{\gamma} \frac{1}{e^{\gamma} + 1} + \gamma \frac{\beta}{2\xi\alpha}.$$

Let $c = \frac{\beta}{2\xi\alpha} \geq 0$ by definition, then we solve

$$\arg \inf_{\gamma} \frac{1}{e^{\gamma} + 1} + c\gamma.$$

We can study this function by looking at the sign of its derivative. Considering the derivative equal to 0, we have

$$\begin{aligned}
 c - \frac{e^\gamma}{(e^\gamma + 1)^2} &= 0 \\
 \Leftrightarrow c(e^\gamma + 1)^2 - e^\gamma &= 0 \\
 \Leftrightarrow ce^{2\gamma} + 2ce^\gamma + c - e^\gamma &= 0 \\
 \Leftrightarrow ce^{2\gamma} + (2c - 1)e^\gamma + c &= 0.
 \end{aligned}$$

With a change of variables $x = e^\gamma$, we solve

$$cx^2 + (2c - 1)x + c = 0,$$

and obtain the following root $\Delta = (2c - 1)^2 - 4c^2 = 1 - 4c$. In the following, we consider two cases:

- Let $c \geq \frac{1}{4}$, then $\Delta \leq 0$ and there no or a single root. Since $c \geq 0$, the gradient is always positive which implies that the minimum is reached at $\gamma = 0$ which is $\lambda = 0$. Therefore, in this case, the bound is equal to $\alpha = \sqrt[p]{2} \|(|\mathcal{Y}|) - 1\|_p$.
- Let $c < \frac{1}{4}$, then $\Delta > 0$ and we have $x = \frac{1-2c \pm \sqrt{1-4c}}{2c}$. Since, $x = e^\gamma$ and $\gamma \geq 0$, then $x \geq 1$.

If $x = \frac{1-2c-\sqrt{1-4c}}{2c}$ and $x \geq 1$, then $1 - 4c \geq \sqrt{1 - 4c}$ which is impossible since $c < \frac{1}{4}$.

It implies that $x = \frac{1-2c+\sqrt{1-4c}}{2c}$ which is $\lambda = \frac{1}{\xi} \log\left(\frac{1-2c+\sqrt{1-4c}}{2c}\right)$. Then, we have

$$\lambda = \frac{1}{\xi} \log\left(\frac{1 - 2c + \sqrt{1 - 4c}}{2c}\right) = \frac{1}{\xi} \log\left(\frac{1}{2c} (1 + \sqrt{1 - 4c}) - 1\right).$$

Since we have an increasing function for $\lambda' \geq \lambda$ and $\sqrt{1 - 4c} \leq 1$, we can consider

$$\lambda \leq \lambda' = \frac{1}{\xi} \log\left(\frac{1}{c} - 1\right).$$

In this case, the bound becomes

$$\begin{aligned}
 \frac{2\alpha}{e^{\frac{1}{\xi} \log(\frac{1}{c}-1)} \xi} + \frac{1}{\xi} \log\left(\frac{1}{c} - 1\right) \beta &= 2\alpha \frac{1}{\frac{2\xi\alpha}{\beta}} + \frac{1}{\xi} \log\left(\frac{2\xi\alpha}{\beta} - 1\right) \beta, \\
 &= \frac{\beta}{\xi} \left[1 + \log\left(\frac{2\xi\alpha}{\beta} - 1\right)\right].
 \end{aligned}$$

Thus, we obtain the following bound

$$\min\left(\alpha, \frac{\beta}{\xi} \left[1 + \log\left(\max\left(4, \frac{2\xi\alpha}{\beta}\right) - 1\right)\right]\right),$$

where the left term of the minimization corresponds to the bound when $\frac{2\xi\alpha}{\beta} \leq 4$, otherwise the bound is equal to the right term.



Appendix E. Details of the bounds handling under different sensitive attribute scenarios.

We summarize the different scenarios of sensitive attributes and how adaptable our theoretical framework is in Table 10.

Type of SA	Lemma 1 & 2	Th. 3	Th. 4	Empirically
$A \in \{0, 1\}$	✓	✓	✓	✓
$A \in \{0, \dots, K\}$	✓	✓ (with Natarajan dimension)	✓ (with generalization of the $\mathcal{H}\Delta\mathcal{H}$ -divergence (Sicila et al., 2022))	✓
Multiple SA	✓ (considering the intersection of Y and A so that $A \in \{0, \dots, K\}$)			
$A \in \mathbb{R}$	✓ (with binning)	✓ (with binning)	✓ (with binning)	✓ (with regression)

Table 10: Summary of the different possible scenarios for the sensitive attributes (SA)

Appendix F. Experimental details

F.1 WFC algorithm

In this section, we describe the full algorithm of WFC. Algorithm 1 provides the detailed algorithm for WFC used in our experiments.

F.2 Details when using BERT-encoder

In this section, we provide additional experimental details, notably, we detail the architectures of the MLPs and give the optimal hyperparameters when BERT model is used to obtain the initial representations.

F.2.1 MLP ARCHITECTURE

In Table 11a, we present the architectural details of the classifier MLP. We grid searched over the learning rate ($lr \in \{1e^{-5}, 1e^{-4}, 1e^{-3}, 5e^{-5}, 5e^{-4}, 5e^{-3}\}$, the number of training batches for classification per epoch $n_d \in \{5, 10, 20\}$, the value used to clip the weights to enforce the Lipschitz constraint *clipping value* $\in \{0.001, 0.01, 0.1\}$, the parameter $\beta \in \{0.1, 0.5, 1, 2, 5, 10, 100\}$, the layer used between the *first hidden*, *last hidden*, or *last layer*.

F.2.2 CRITIC ARCHITECTURE

In Table 11b, we present the architectural details of the Critic, which is a simple multi-layer perceptron. We grid searched over the learning rate $lr \in \{5e^{-5}, 5e^{-4}, 5e^{-3}\}$.

Data: $D = \{(x_i, y_i, a_i)\}_{i=1}^n$ the training set, n_e the number of epochs, n_c and n_d the number of training iterations per epoch for the critic and the classifier respectively, a batch size n_b , two neural networks $h_a(Enc(x))$ and $h_y(Enc(x); \theta)$, a Critic C_ω , a weight on the regularization β

```

for  $e = 1, \dots, n_e$  do
    for  $t = 1, \dots, n_c$  do
        Sample  $\{x_i, y_i, a_i\}_{i=1}^{n_b}$ 
        Encode :  $z_a \leftarrow \{h_a(Enc(x_i))\}_{i=1}^{n_b}$ ,  $z_y \leftarrow \{h_y(Enc(x_i))\}_{i=1}^{n_b}$ 
        Concatenate vectors to get  $Z_{dep} \leftarrow [z_{a,i}, z_{y,i}]_{i=1}^{n_b}$ 
        Shuffle the  $z_{a,i}$  vectors.
        Concatenate vectors to get  $Z_{ind} \leftarrow [z_{s,i}, z_{y,i}]_{i=1}^{n_b}$ 
         $grad(w) \leftarrow \nabla_w \frac{1}{n_b} (\sum_{i=1}^{n_b} C_\omega(Z_{dep,i}) - \sum_{i=1}^{n_b} C_\omega(Z_{ind,i}))$ 
         $\omega \leftarrow Adam(\omega; grad(w))$ 
    end
    for  $t = 1, \dots, n_d$  do
        Sample  $\{x_i, y_i, a_i\}_{i=1}^{n_b}$ 
        Encode :  $z_s \leftarrow \{h_a(x_i)\}_{i=1}^{n_b}$ ,  $z_y \leftarrow \{h_y(x_i)\}_{i=1}^{n_b}$ 
        Concatenate vectors to get  $Z_{dep} = [z_{a,i}, z_{y,i}]_{i=1}^{n_b}$ 
        Shuffle the  $z_{a,i}$  vectors.
        Concatenate vectors to get  $Z_{ind} = [z_{a,i}, z_{y,i}]_{i=1}^{n_b}$ 
         $\mathcal{L} \leftarrow \sum_{i=1}^{n_b} \mathcal{L}(y_i, h_y(Enc(x_{y,i})))$ 
         $\mathcal{L} \leftarrow \mathcal{L} + \beta (\sum_{i=1}^{n_b} C_\omega(Z_{dep,i}) - \sum_{i=1}^{n_b} C_\omega(Z_{ind,i}))$ 
         $\theta \leftarrow Adam(\theta; \nabla_\theta \frac{1}{n_b} \mathcal{L})$ 
    end
end
    
```

Algorithm 1: WFC Algorithm

F.3 Details when using SFR-Embeddings-2_R

F.3.1 MLP ARCHITECTURE

In Table 12a, we present the architectural details of the classifier MLP when the embeddings are produced by the SFR-Embeddings-2_R. We grid searched over the learning rate ($lr \in \{3e^{-7}, 3e^{-6}, 3e^{-5}, 3e^{-3}, 3e^{-1}\}$), the number of training batches for classification per epoch and for the Critic training $n_d, n_c \in \{5, 10, 20\}$, and the hidden layer dimension (100, 300, 900).

F.3.2 CRITIC ARCHITECTURE

In Table 12b, we present the architectural details of the Critic for the task using SFR-Embeddings-2_R. We grid searched over the learning rate $lr \in \{3e^{-7}, 3e^{-6}, 3e^{-5}, 3e^{-3}, 3e^{-1}\}$.

F.3.3 BASELINES HYPERPARAMETERS

We select the best hyperparameters for the baselines for the classification of the representations generated by the SFR-Embedding-2_R model. Following Shen et al. (2022b), we

Data set	Bios	Moji	Hyperparameter	Value
input dimension	768	2304	number hidden layer	1
hidden layers	2	2	hidden dimension	512
hidden dimension	300	300	activation	ReLU
learning rate	1^{-4}	1^{-5}	optimizer	Root Mean Square Propagation
batch size	128	128	learning rate	$5e^{-5}$
epochs max	10000	10000	(b) Details on hyperparameters used for the Critic MLP.	
activation	TanH	TanH		
β	1	1		
n_c	20	5		
n_d	5	5		
clipping value	0.01	0.01		
layer used	last	last		

(a) Details on hyperparameters used for the classifying MLP.

Table 11: Hyperparameter details when using BERT-encoder.

first determine the optimal hyperparameters of the classification models and keep those hyperparameters fixed when searching for the method-specific best hyperparameters. We tune the learning rate ($lr \in \{3e^{-1}, 3e^{-2}, 3e^{-3}, \mathbf{3e^{-4}}, 3e^{-5}\}$) and the hidden dimension ($\in \{100, \mathbf{300}, 900\}$). For the ADV baseline, we take 3 adversaries and consider several values for the following hyperparameters $adv_diverse_lambda \in \{1e^{-1}, 1e^{-2}, \mathbf{1e^{-3}}, 1e^{-4}\}$ and $adv_lambda \in \{0.3, \mathbf{0.5}, 1, 2\}$. Values in bold are the selected ones. When BTEO is used the hyperparameters are set to 'EO' for *BTObj*, 'Resampling' for *BT* as in (Shen et al., 2022b). Finally, the embedding size is 4096, the batch size is 1024, and we set a patience of 10 for the early stopping.

F.3.4 DETAILS FOR CROSS-DOMAIN WFC

In this section, we explain how we build the data set used for the cross-domain experiment to increase the divergence with the Bios data set. To do so, we remove a set of words from the MP data set with regard to the sensitive attribute: gender. The words included in the set are the following: 'he', 'him', 'his', 'himself', 'Mr.', 'Sir', 'Lord', 'King', 'Prince', 'man', 'boy', 'gentleman', 'father', 'son', 'husband', 'brother', 'uncle', 'nephew', 'king', 'prince', 'she', 'her', 'hers', 'herself', 'Mrs.', 'Ms.', 'Miss', 'Lady', 'Dame', 'Queen', 'Princess', 'woman', 'girl', 'lady', 'mother', 'daughter', 'wife', 'sister', 'aunt', 'niece', 'queen', 'princess'.

Hyperparameter	Value
input dimension	4096
hidden layers	1
hidden dimension	300
learning rate	$3e^{-5}$
batch size	128
epochs max	10000
activation	TanH
β	1
n_c	20
n_d	10
clipping value	0.01
layer used	last

(a) Details on hyperparameters used for the classifying MLP.

Hyperparameter	Value
number hidden layer	1
hidden dimension	512
activation	ReLU
optimizer	Root Mean Square Propagation
learning rate	$3e^{-6}$

(b) Details on hyperparameters used for the Critic MLP.

Table 12: Hyperparameter details for SFR-Embeddings-2-R.

References

- Civil Rights Act. Civil rights act of 1964. *Title VII, Equal Employment Opportunities*, 1964.
- Martin Arjovsky and Leon Bottou. Towards principled methods for training generative adversarial networks. In *International Conference on Learning Representations*, 2017.
- Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, pages 214–223. PMLR, 2017.
- James Atwood, Nino Scherrer, Preethi Lahoti, Ananth Balashankar, Flavien Prost, and Ahmad Beirami. Inducing group fairness in prompt-based language model decisions. 2024.
- Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds post-processing under imperfect group information. In *International Conference on Artificial Intelligence and Statistics*, pages 1770–1780. PMLR, 2020.
- Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.
- Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine learning*, 79:151–175, 2010.
- Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *FAccT*, pages 610–623, 2021.

- Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.
- Su Lin Blodgett, Lisa Green, and Brendan O’Connor. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*, 2016.
- Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’23, page 370–378, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594004. URL <https://doi.org/10.1145/3593013.3594004>.
- Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.
- Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7), April 2024. ISSN 0360-0300. doi: 10.1145/3616865. URL <https://doi.org/10.1145/3616865>.
- L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* ’19, page 319–328, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287586. URL <https://doi.org/10.1145/3287560.3287586>.
- Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35:19152–19164, 2022.
- Jiahao Chen and Bing Su. Transfer knowledge from head to tail: Uncertainty calibration under long-tailed distribution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19978–19987, 2023.
- Myra Cheng, Esin Durmus, and Dan Jurafsky. Marked personas: Using natural language prompts to measure stereotypes in language models. *arXiv preprint arXiv:2305.18189*, 2023.
- Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. Fairfil: Contrastive neural debiasing method for pretrained text encoders, 2021.
- Zhibo Chu, Zichong Wang, and Wenbin Zhang. Fairness in large language models: A taxonomic survey. *arXiv preprint arXiv:2404.01349*, 2024.
- Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression via plug-in estimator and recalibration with statistical guarantees. *Advances in Neural Information Processing Systems*, 33:19137–19148, 2020.

- Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 91–98, 2019.
- Nicolas Courty, Rémi Flamary, and Devis Tuia. Domain adaptation with regularized optimal transport. In *ECML PKDD*, pages 274–289. Springer, 2014.
- Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *FaccT*, pages 120–128, 2019.
- Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772, 2009.
- Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. Fairness guarantees in multi-class classification with demographic parity. *Journal of Machine Learning Research*, 25(130):1–46, 2024. URL <http://jmlr.org/papers/v25/23-0322.html>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019.
- Kush Dubey. Evaluating the fairness of task-adaptive pretraining on unlabeled test data before few-shot text classification. In Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Amirhossein Kazemnejad, Christos Christodoulopoulos, Mario Giulianelli, and Ryan Cotterell, editors, *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*, pages 1–26, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.genbench-1.1. URL <https://aclanthology.org/2024.genbench-1.1>.
- Jannik Dunkelau. Fairness-aware machine learning an extensive overview. 2020. URL <https://api.semanticscholar.org/CorpusID:237483522>.
- Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of EMNLP*. Association for Computational Linguistics, 2018.
- European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016. URL <https://data.europa.eu/eli/reg/2016/679/oj>.
- European Parliament and Council of the European Union. Regulation (eu) 2024/1689 of the european parliament and of the council. 2024.
- Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, 2017.

- Susan T Fiske. Prejudices in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on psychological science*, 12(5):791–799, 2017.
- Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boissunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. Pot: Python optimal transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021. URL <http://jmlr.org/papers/v22/20-451.html>.
- Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A Poggio. Learning with a wasserstein loss. *NeurIPS*, 28, 2015.
- Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *arXiv preprint arXiv:1704.00805*, 2017.
- Alison L. Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International Statistical Review / Revue Internationale de Statistique*, 70(3):419–435, 2002. ISSN 03067734, 17515823. URL <http://www.jstor.org/stable/1403865>.
- Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *ICML*, pages 2357–2365. PMLR, 2019.
- Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the International Conference on Neural Information Processing Systems*, NIPS’17, page 5769–5779. Curran Associates Inc., 2017. ISBN 9781510860964.
- Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017.
- Maya Gupta, Andrew Cotter, Mahdi Milani Fard, and Serena Wang. Proxy fairness. *arXiv preprint arXiv:1806.11212*, 2018.
- Umang Gupta, Aaron M Ferber, Bistra Dilkina, and Greg Ver Steeg. Controllable guarantees for fair outcomes via contrastive information estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):7610–7619, May 2021. doi: 10.1609/aaai.v35i9.16931. URL <https://ojs.aaai.org/index.php/AAAI/article/view/16931>.
- Xudong Han, Timothy Baldwin, and Trevor Cohn. Decoupling adversarial training for fair NLP. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 471–477, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.41. URL <https://aclanthology.org/2021.findings-acl.41>.

- Xudong Han, Timothy Baldwin, and Trevor Cohn. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, Online, April 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.239. URL <https://aclanthology.org/2021.eacl-main.239>.
- Xudong Han, Timothy Baldwin, and Trevor Cohn. Balancing out bias: Achieving fairness through balanced training. *arXiv preprint arXiv:2109.08253*, 2021c.
- Xudong Han, Aili Shen, Yitong Li, Lea Frermann, Timothy Baldwin, and Trevor Cohn. Fairlib: A unified framework for assessing and improving fairness. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 60–71, 2022.
- Xudong Han, Timothy Baldwin, and Trevor Cohn. Fair enough: Standardizing evaluation and model selection for fairness research in NLP. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 297–312, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.23. URL <https://aclanthology.org/2023.eacl-main.23>.
- Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *NeurIPS*, 29, 2016.
- Max Hort, Zhenpeng Chen, Jie M. Zhang, Mark Harman, and Federica Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM J. Responsib. Comput.*, 1(2), June 2024. doi: 10.1145/3631326. URL <https://doi.org/10.1145/3631326>.
- Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in NLP models as barriers for persons with disabilities. In *ACL*, pages 5491–5501, 2020.
- Shadi Iskander, Kira Radinsky, and Yonatan Belinkov. Leveraging prototypical representations for mitigating social bias without demographic information. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 379–390, Mexico City, Mexico, June 2024. Association for Computational Linguistics. URL <https://aclanthology.org/2024.naacl-short.33>.
- Taeuk Jang and Xiaoqian Wang. Fades: Fair disentanglement with sensitive relevance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12067–12076, June 2024.
- Mariana Jatobá, Juliana Santos, Ives Gutierrez, Daniela Moscon, Paula Odete Fernandes, and João Paulo Teixeira. Evolution of artificial intelligence research in human resources. *Procedia Computer Science*, 164:137–142, 2019.

- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
- Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In *Uncertainty in artificial intelligence*, pages 862–872. PMLR, 2020.
- Zhijing Jin, Julius von Kügelgen, Jingwei Ni, Tejas Vaidhya, Ayush Kaushal, Mrinmaya Sachan, and Bernhard Schölkopf. Causal direction of data collection matters: Implications of causal and anticausal learning for NLP. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP*, pages 9499–9513. Association for Computational Linguistics, 2021.
- Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012. ISBN 978-3-642-33486-3.
- Leonid Kantorovich. On the translocation of masses. In *C.R. (Doklady) Acad. Sci. URSS(N.S.)*, volume 37(10), pages 199–201, 1942.
- Patrik Joslin Kenfack, Samira Ebrahimi Kahou, and Ulrich Aïvodji. Fairness under demographic scarce regime. *arXiv preprint arXiv:2307.13081*, 2023.
- Svetlana Kiritchenko and Saif M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems, 2018.
- Charlotte Laclau, Ievgen Redko, Manvi Choudhary, and Christine Largeron. All of the fairness for edge prediction with optimal transport. In *AISTATS*, pages 1774–1782. PMLR, 2021.
- Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*, 2024.
- Thibaud Leteno, Antoine Gourru, Charlotte Laclau, Rémi Emonet, and Christophe Gravier. Fair text classification with Wasserstein independence. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15790–15803, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.978. URL <https://aclanthology.org/2023.emnlp-main.978/>.
- Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*, 2023.

- Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *ICML*, pages 6565–6576. PMLR, 2021.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.
- Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/1b486d7a5189ebe8d8c46afc64b0d1b4-Paper.pdf.
- Michael Lohaus, Matthäus Kleindessner, Krishnaram Kenthapadi, Francesco Locatello, and Chris Russell. Are two heads the same as one? identifying disparate treatment in fair neural networks. *Advances in Neural Information Processing Systems*, 35:16548–16562, 2022.
- David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.
- David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 875–884. PMLR, 26–28 Aug 2020. URL <https://proceedings.mlr.press/v108/mcallester20a.html>.
- Daniel McNamara, Cheng Soon Ong, and Robert C Williamson. Provably fair representations. *arXiv preprint arXiv:1710.04394*, 2017.
- Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. Sfr-embedding-2: Advanced text embedding with multi-stage training, 2024. URL https://huggingface.co/Salesforce/SFR-Embedding-2_R.
- Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences*, pages 666–704, 1781.
- Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. doi: 10.48550/ARXIV.2210.07316. URL <https://arxiv.org/abs/2210.07316>.
- Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. *Advances in Neural Information Processing Systems*, 33:20673–20684, 2020.
- Balas K Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

- Sherjil Ozair, Corey Lynch, Yoshua Bengio, Aaron Van den Oord, Sergey Levine, and Pierre Sermanet. Wasserstein dependency measure for representation learning. *Advances in Neural Information Processing Systems*, 32, 2019.
- Sangdon Park, Osbert Bastani, James Weimer, and Insup Lee. Calibrated prediction with covariate shift via unsupervised domain adaptation. In *International Conference on Artificial Intelligence and Statistics*, pages 3219–3229. PMLR, 2020.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the Annual Meeting of ACL*, pages 7237–7256. ACL, 2020.
- Laurent Risser, Alberto González Sanz, Quentin Vincenot, and Jean-Michel Loubes. Tackling algorithmic bias in neural-network classifiers using wasserstein-2 regularization. *Journal of Mathematical Imaging and Vision*, 64(6):672–689, 2022.
- Gabriel Roccabruna, Massimo Rizzoli, and Giuseppe Riccardi. Will LLMs replace the encoder-only models in temporal relation classification? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20402–20415, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1136. URL <https://aclanthology.org/2024.emnlp-main.1136/>.
- Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. In *International Conference on Learning Representations*, 2021.
- Qian Ruan, Ilya Kuznetsov, and Iryna Gurevych. Are large language models good classifiers? a study on edit intent classification in scientific document revisions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15049–15067, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.839. URL <https://aclanthology.org/2024.emnlp-main.839>.
- Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover’s distance as a metric for image retrieval. *Int. J. Comput. Vision*, 40(2):99–121, 2000. ISSN 0920-5691. doi: 10.1023/A:1026543900054.
- Bernhard Schölkopf, Dominik Janzing, Jonas Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In *Proceedings of the International Conference on Machine Learning, ICML*, 2012.
- Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688*, 2019.
- Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Optimising equal opportunity fairness in model training. In *Proceedings of the Conference of the North*

- American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4073–4084. Association for Computational Linguistics, 2022a.
- Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2022*, pages 81–95, 2022b.
- Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.
- Anthony Sicilia, Katherine Atwell, Malihe Alikhani, and Seong Jae Hwang. Pac-bayesian domain adaptation bounds for multiclass learners. In *Uncertainty in artificial intelligence*, pages 1824–1834. PMLR, 2022.
- Chiappa Silvia, Jiang Ray, Stepleton Tom, Pacchiano Aldo, Jiang Heinrich, and Aslanides John. A general approach to fairness with optimal transport. In *Proceedings of AAAI*, volume 34, pages 3633–3640, 2020.
- Olivia Sturman, Aparna R Joshi, Bhaktipriya Radharapu, Piyush Kumar, and Renee Shelby. Debiasing text safety classifiers through a fairness-aware ensemble. In Franck Dernoncourt, Daniel Preoŧiuc-Pietro, and Anastasia Shimorina, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 199–214, Miami, Florida, US, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-industry.16. URL <https://aclanthology.org/2024.emnlp-industry.16>.
- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, 2019.
- Yi Chern Tan and L Elisa Celis. Assessing social and intersectional biases in contextualized word representations. *NeurIPS*, 32, 2019.
- Luis Caicedo Torres, Luiz Manella Pereira, and M Hadi Amini. A survey on optimal transport for machine learning: Theory and applications. *arXiv preprint arXiv:2106.01963*, 2021.
- Nicolás Torres. Contrastive adversarial gender debiasing. *Natural Language Processing Journal*, 8:100092, 2024. ISSN 2949-7191. doi: <https://doi.org/10.1016/j.nlp.2024.100092>. URL <https://www.sciencedirect.com/science/article/pii/S2949719124000402>.
- VN Vapnik. Statistical learning theory. *Adaptive and Learning Systems for Signal Processing Communications and control*, 1998.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017.

- Ximei Wang, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable calibration with lower bias and variance in domain adaptation. *Advances in Neural Information Processing Systems*, 33:19212–19223, 2020.
- Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. Improving the improved training of wasserstein gans: A consistency term and its dual effect. In *International Conference on Learning Representation (ICLR)*, 2018.
- Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pages 1920–1953. PMLR, 2017.
- Xilin Yang. Diagnosing hate speech classification: Where do humans and machines disagree, and why? *arXiv e-prints*, pages arXiv–2410, 2024.
- Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333. PMLR, 2013.
- Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.
- Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, WSDM ’22*, page 1433–1442, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450391320. doi: 10.1145/3488560.3498493. URL <https://doi.org/10.1145/3488560.3498493>.