Effective Controllable Bias Mitigation for Classification and Retrieval using Gate Adapters

Anonymous ACL submission

Abstract

Bias mitigation of Language Models has been the topic of many studies with a recent focus on learning separate modules like adapters for on-demand debiasing. Besides optimizing for a modularized debiased model, it is often critical in practice to control the degree of bias reduction at inference time, e.g., in order to tune for a desired performance-fairness trade-off in search results or to control the strength of debiasing in classification tasks. In this paper, we introduce *Controllable Gate Adapter (*CONGATER*)*, a novel modular gating mechanism with adjustable sensitivity parameters, which allows for a gradual transition from the biased state of the model to the fully debiased version at inference time. We demonstrate CONGATER performance by (1) conducting adversarial debiasing experiments with three different models on three classification tasks with four protected attributes, and (2) reducing the bias of search results through fairness list-wise regularization to enable adjusting a trade-off between performance and fairness metrics. Our experiments on the classification tasks show that compared to baselines of the same caliber, CONGATER can maintain higher task performance while containing less information regarding the attributes. Our results on the retrieval task show that the fully debiased CONGATER can achieve the same fairness performance while maintaining more than twice as high task performance than recent strong baselines. Overall, besides strong performance CONGATER enables the continuous transitioning between biased and debiased states of models, enhancing personalization of use and interpretability through controllability.

1 Introduction

Pre-trained Language models (LMs) have shown impressive ability in learning effective representations and diverse aspects of language, including harmful biases and stereotypes (Zhao et al., 2019; Sheng et al., 2019; Blodgett et al., 2020; Rekabsaz and Schedl, 2020; Stanovsky et al., 2019). A common bias mitigation category, referred to as representational fairness (Elazar and Goldberg, 2018), aims at minimizing the information regarding a specific attribute in order to make the models' decision blind to the attribute. This is realized in various classification scenarios to make the model invariant to given protected attributes, and also in information retrieval (IR) tasks to opt for the neutrality/balancedness of search results. Common in-processing approaches to mitigate these biases are to extend model optimization with various bias mitigation criteria (e.g., through adversarial optimization or regularization terms) and update the whole model's parameters to a debiased state (Elazar and Goldberg, 2018; Colombo et al., 2021; Zerveas et al., 2022b). New studies focus on modularizing this process by introducing new modules such as adapters (Pfeiffer et al., 2021; Houlsby et al., 2019), and sparse masking networks (Zhang et al., 2021; Hauzenberger et al., 2023; Zhao et al., 2020).

Besides effectiveness in reducing bias, it is often important in practice to be able to control the degree of imposing the debiasing criteria at inference time. This is beneficial particularly to apply possible fairness-performance trade-offs, specific preferences of each user, or the particular needs in processing each given input.[1] Debiasing controllability enables to set the desired degree of a bias constraint's contribution at inference time, while in the current paradigm, one needs to train and deploy multiple parallel models or modules with various mitigation degrees (Kumar et al., 2023; Hauzenberger et al., 2023; Zerveas et al., 2022b), imposing an untenable burden in practice.

[1] Regarding the last point, see for instance Krieg et al. (2023) and Hauzenberger et al. (2023) about the need to control for the gender information in the processing of bias-sensitive inputs (like *how to become CEO?*) versus the "normal" ones (like *earliest pregnancy symptoms*).

In this paper, we address debiasing controllability by introducing *Controllable Gate Adapter* (CONGATER). The proposed module is based on a novel gating mechanism, that learns to reduce protected attribute information from the embedding while allowing information necessary for the task to pass through the model. The CONGATER is equipped with a novel activation function *Trajectory Sigmoid (*`t-sigmoid`*)*, used to form the gate vectors. CONGATER is agnostic to the debiasing optimization, and can be trained with any gradient descent-based signal which removes attributes or increases fairness. During training, `t-sigmoid` has the same shape as a (standard) sigmoid function. At inference time, however, the form of `t-sigmoid` can flatten by decreasing the sensitivity parameter, transitioning from the sigmoid function (full gate intervention) to the constant function (no influence) creating a nonlinear interpolation effect. This transition can be viewed as traversing the trajectory of embeddings from the state of the original (biased) model to its fully debiased version, resulting in adjustable attribute removal qualities in the model's outputs and internal embeddings (§3).

We demonstrate the functionality of CONGATER by doing two sets of experiments: (1) adversarial bias mitigation with three models on three real-world classification datasets namely, occupation prediction from biographies with gender as protected attribute (De-Arteaga et al., 2019), hate speech detection with dialect-based race as protected attribute (Founta et al., 2018a), and mention prediction with two attributes: gender and age of authors (Rangel et al., 2016). In this experiment, we show that CONGATER is able to reduce information about the attribute in the embeddings better than baselines while mostly preserving task performance. We also show that attribute information reduction in the embeddings of the model is continuous. This continuous control results in higher interpretability through controllability about model behavior at inference time. (2) Fairness/Neutrality of search results with gender as a protected attribute. We conduct the experiments on a recent IR benchmark (Rekabsaz et al., 2021; Nguyen et al., 2016), optimizing CONGATER with a recently-introduced list-wise neutrality regularization term (Zerveas et al., 2022b). We demonstrate that the fully debiased CONGATER is able to preserve task performance more than twice as high as the baselines with the same fairness performance, and CONGATER is able to control the trade-off between the biased and debiased model in a continuous and linear fashion (details in §4 and §5).

2 Related Work

Efficient modular training introduces an alternative to fine-tuning, where a (small) network is trained for a specific objective while the core model's parameters remain unchanged (Pfeiffer et al., 2023). Adapters realize modular training with a non-linear feed-forward network added to each layer (Rebuffi et al., 2017; Houlsby et al., 2019; Stickland and Murray, 2019) of transformers. Several works study the various aspects of adapters, such as parameter efficiency (Rücklé et al., 2021; Han et al., 2021a), architectural variations (Mahabadi et al., 2021), and transfer learning capacity (Pfeiffer et al., 2021). Recently, Lian et al. (2022) show that scale and shifting of embedding is sufficient for effectively learning the task.

Bias mitigation. Mitigating societal bias in LMs is explored particularly in the context of *attribute erasure*. The aim of this task is to reduce the encoded information of a specific attribute from the latent embeddings and is particularly utilized in the context of mitigating empirical societal biases in LMs (Mehrabi et al., 2022; Shen et al., 2022). Bias mitigation is approached by methods such as linearly projecting embeddings into the space with minimum correlations to protected attributes (Ravfogel et al., 2020; Kaneko and Bollegala, 2021), to achieve empirical fairness, through using a distribution alignment loss (Guo et al., 2022), or by applying adversarial training to learn representations agnostic to protected attributes (Elazar and Goldberg, 2018; Barrett et al., 2019; Han et al., 2021b; Wang et al., 2021; Rekabsaz et al., 2021; Ganhör et al., 2022).

Controllability. Decoder LMs have been extensively studied for controllability. Researchers mostly cover tasks such as attribute manipulation (like positive/negative sentiment), and imposing predefined syntactic/semantic structure to text generation (Zhang et al., 2022; Ross et al., 2022; Kumar et al., 2022; Qin et al., 2022; Shen et al., 2020). As examples, Subramani et al. (2022) introduces steering vectors as a way of changing the semantics of text generation, while Yu et al. (2021) learn an alignment function to force the text generation in the direction of a specific target concept. More recently, Hallinan et al. (2023) used likelihood between expert and anti-expert models to detoxify text generation.

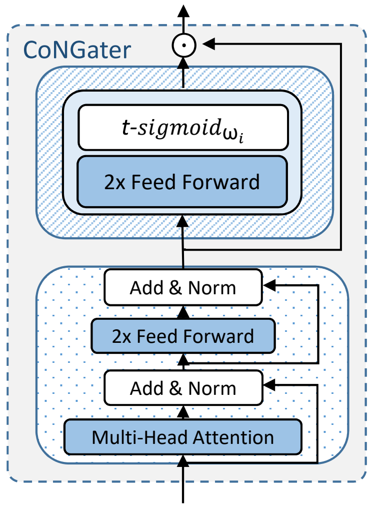
Information Retrieval. In IR tasks, many bias mitigation methods have been proposed. These methods mostly use list-wise optimization (Oosterhuis, 2022; Morik et al., 2021). Rekabsaz et al. (2021) proposed integrating adversarial training in deep-ranking models to improve bias mitigation. Zerveas et al. (2022b) introduced bias-aware optimization method using CODER (Zerveas et al., 2021) with TAS-B (Hofstätter et al., 2021). We will use their method as the training strategy for our IR fairness task 3.2.

Few recent studies explore modularized adversarial bias mitigation of encoder language models. Lauscher et al. (2021) use a stack of adapters, while Kumar et al. (2023) first learns separate adapters for tasks and debiasing attributes and then combine them on-demand using the fusion network. Utilizing masking methods, Zhang et al. (2021) learns binary masks applied to the initial network to erase the concept of interest, and Hauzenberger et al. (2023) train sparse weight-difference subnetworks, one for each attribute, which can be added to the core model on-demand. Our work extends this line of research by introducing a modularized, graded (non-binary), and controllable approach evaluated on continuous concept erasure.

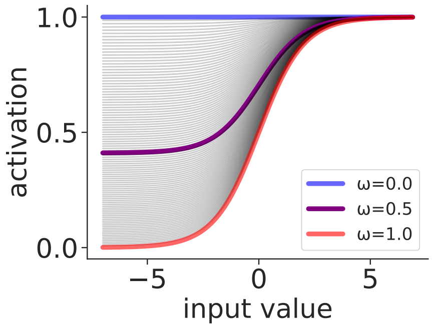
Finally, the *gating mechanism* has been used in various architectures to learn via scaling. As examples, Ramachandran et al. (2017) propose the self-gate activation function with trainable parameters, while Papernot et al. (2021) introduce the tempered *sigmoid* activation with a bias and scaling factor. Hu et al. (2018) introduced Squeeze and Excitation networks which use bottle-neck networks added after each convolutional layer followed by a sigmoid activation function. Compared to Squeeze and excitation which captures global attention to the channels, our model uses a second training signal to isolate and filter out protected attributes. Another difference between our method and Squeeze and Excitation is the usage of a new activation function instead of a sigmoid activation function which gives us the benefit of controllability at inference.

3 Model and Training

In this section, we first introduce the proposed *Controllable Gate Adapter* (CONGATER) and *trajectory-sigmoid* (`t-sigmoid`) activation function. We then explain the parallel and post-hoc training regimes, and how the gating sensitivity parameter of `t-sigmoid` can be adjusted at inference time to control the effectiveness of the gates.



(a) CONGATER



(b) `t-sigmoid`

Figure 1: (a) The overall architecture of CONGATER as an adjustable self-gate adapter network. (b) Effect of ω parameter on `t-sigmoid`. Increasing ω results in a transition from the constant function $y = 1$ (open gate) to the sigmoid function (full functional gate).

3.1 CONGATER Architecture

The CONGATER module follows the principle of adapters (Houlsby et al., 2019) by dedicating a small network, added after each transformer block (Vaswani et al., 2017) of an LM. Figure 1a depicts the architecture of a CONGATER module inside a transformer block, responsible for controlling one attribute. In short, CONGATER applies a gating mechanism for each layer, where the gate vector is defined via bottleneck followed by the `t-sigmoid` activation function. Concretely, for the i^{th} target attribute, we first define the gate vector $\boldsymbol{g}_i$ formulated as:

$$\boldsymbol{g}_i = \texttt{t-sigmoid}_{\omega_i}(\boldsymbol{v}_i) \quad (1)$$

$$\boldsymbol{v}_i = \boldsymbol{W}_i^2 \tanh(\boldsymbol{W}_i^1 \boldsymbol{h} + \boldsymbol{b}_i^1) + \boldsymbol{b}_i^2 \quad (2)$$

where $\boldsymbol{h}$ is the input vector (output of the transformer block), and $\boldsymbol{W}_i$ and $\boldsymbol{b}_i$ are weight and bias parameters, respectively. `t-sigmoid` is a generalized form of the sigmoid function, enhanced with the gating sensitivity variable ω_i. This gating sensitivity parameter can be set to values in the range of $[0, 1]$, which changes the shape of `t-sigmoid`, as illustrated for several values of ω in Figure 1b. In particular when $\omega_i = 1$, `t-sigmoid` is equivalent to the sigmoid function σ and hence: $\boldsymbol{g}_i = \sigma(\boldsymbol{v}_i)$. On the other end, setting $\omega_i = 0$ changes `t-sigmoid` to the constant function $y = 1$ resulting in $\boldsymbol{g}_i = 1$. Regardless of the ω_i's value, the output of `t-sigmoid` and hence each value of gating vector $\boldsymbol{g}_i$ is bounded to $[0, 1]$, indicating the range of the gate mechanism from fully closed to the fully open (more details below).

The transformation of the ConGater module is defined as the self-gate of the input using element-wise multiplication, defined below:

$$output = \boldsymbol{h} \odot \boldsymbol{g}_i \tag{3}$$

This transformation downscales each value of $\boldsymbol{h}$ by its corresponding gate value, except for the cases with a corresponding open gate (gating value of 1), to which no change is made. Overall, a ConGater has the same number of parameters as a standard adapter network (Pfeiffer et al., 2021) with the negligible computation overheads of Eq. 1 and 3.

We now formulate the `t-sigmoid` activation function and discuss how it can be used to control the behavior of the model. The `t-sigmoid` function is formulated below:

$$\texttt{t-sigmoid}_{\omega}(x) = 1 - \frac{\log_2(\omega + 1)}{1 + e^{x}}, \quad \omega \in [0, 1] \tag{4}$$

The gate sensitivity parameter ω is not trainable and can be set manually to change the shape of the activation function. If $\omega = 0$, the `t-sigmoid` becomes the constant function $\texttt{t-sigmoid}(\boldsymbol{v}_i) = 1$, meaning that the whole ConGater module turns into an identity function that simply outputs the given input. By increasing the value of ω, `t-sigmoid` gradually transforms to the shape of a sigmoid function. The gradual transformation of `t-sigmoid` has the following characteristics: (1) For a specific value of ω, the output of `t-sigmoid` monotonically increases with increasing input value x; (2) Given $\omega_2 > \omega_1$, the resulting outputs of the same input value x is $\texttt{t-sigmoid}_{\omega_2}(x) \leq \texttt{t-sigmoid}_{\omega_1}(x)$ (stronger gate); (3) Throughout the spectrum of ω, the shape of `t-sigmoid` gradually changes, avoiding drastic alterations. These characteristics allow a smooth change in the effect of the gating mechanism, non-linear interpolation, and hence continuous controllability of the information flow for the respective attribute.[2]

Algorithm 1 ConGater Training

Input: Task-related parameters Θ, parameters of i^{th} ConGater θ_i
if Parallel-Training **then**
    while training **do**
        Set $\omega_i = 0$
        Update Θ using $\mathcal{L}_{task}$
        Set $\omega_i = 1$ and freeze Θ
        Update θ_i using $\mathcal{L}_{task} + \mathcal{L}_{\rho_i}$
else if Posthoc-Training **then**
    Set $\omega_i = 0$
    while training **do**
        Update Θ using $\mathcal{L}_{task}$
    Set $\omega_i = 1$ and freeze Θ
    while training **do**
        Update θ_i with $\mathcal{L}_{task} + \mathcal{L}_{\rho_i}$

3.2 Training and Inference

Training ConGater an attribute requires two distinct training signals. The first training signal comes from the loss function of the main task, denoted by $\mathcal{L}_{task}$. Th second loss is dedicated to each attribute i, denoted by $\mathcal{L}_{\rho_i}$. ConGater is agnostic to the choice of the training signal for $\mathcal{L}_{task}$ and $\mathcal{L}_{\rho_i}$ as we show by deliberately choosing different training signals for our experiments. Depending on the task and dataset, common approaches to realizing a defined signal are through utilizing the provided labels in a dataset, or by leveraging the indicators specific to an attribute, particularly in the case of lack of reliable supervised data (Lauscher et al., 2021; Romanov et al., 2019). Equation 5 shows the overall loss of each attribute where λ is the scaling factor to influence the strength of attribute loss.

$$\mathcal{L}_{total_i} = \mathcal{L}_{task} + \lambda \mathcal{L}_{\rho_i} \tag{5}$$

Depending on the task and mitigation objective, these loss terms can be defined differently. We explain two realizations of these loss functions later

[2] While we discuss removing only one attribute with ConGater, this definition can be extended to multiple attributes as well. We propose one possible multi-attribute ConGater architecture in Appendix A by element-wise multiplication of gates, and report preliminary results using this two-attribute setting in Appendix C.2.

in this section, which we later utilize in our classification and IR experiments. Regardless of the choice of the loss, we first define training procedures for ConGater as follows. For parallel training, the ConGater modules are trained simultaneously with the task, and in post-hoc training, ConGater is added to a fully-trained model in order to learn attribute-specific information. Algorithm 1 shows the pseudocode of these training strategies. The task-related parameters are denoted with Θ, which can be the whole parameters of an LM, or the ones of an additional task-specific modular network such as a task adapter. In each training cycle, regardless of the loss function, we first deactivate the ConGater by setting $\omega_i = 0$ and use the task loss ($\mathcal{L}_{task}$) to train Θ. We then activate the ConGater module by setting $\omega_i = 1$, and update its parameters using equation 5 to encapsulate the information of the target attribute into the respective ConGater while maintaining task performance. While parallel training enables higher flexibility in optimization and exposes the task head indirectly to bias mitigation loss, post-hoc training offers the practical benefits of adding controllable gates to an existing trained model. As described, the ConGater's parameters are trained only with full engagement ($\omega_i = 1$), and the model is never exposed to the settings with partial engagement ($0 < \omega_i < 1$). This makes the training of ConGater efficient and comparable to the training model for each attribute individually (e.g., using adapters).

At inference time, `t-sigmoid` reshaping characteristics indicate how much the target attribute should affect the embeddings, by setting ω_i to any value in $[0, 1]$. In particular, when $\omega_i = 0$, the model works at its original (initial) state with no effect from the ConGater module. By increasing ω_i and changing the shape of the `t-sigmoid`, the effect of the gate increases, and a stronger transformation is applied to the embeddings. With $\omega_i = 1$, ConGater reaches its full transformation capacity by applying the sigmoid activation function. We examine this continuous controllability in the following sections.

In what follows, we explain the realizations of the $\mathcal{L}_{task}$ and $\mathcal{L}_{\rho_i}$ in the classification and search bias mitigation scenarios.

Classification loss uses cross-entropy loss between task labels y and model's output $f(z)$:

$$\mathcal{L}_{task} = \mathrm{CE}(f(z), y)$$

where z is the encoded output embedding, and f the classification head. The disentanglement loss $\mathcal{L}_{\rho_i}$ can be realized by various methods such as mutual information reduction methods (Colombo et al., 2021), adversarial training (Elazar and Goldberg, 2018) or any other loss related to representational fairness or empirical fairness (Ravfogel et al., 2020). We use adversarial loss (Kumar et al., 2023; Lauscher et al., 2021; Zhang et al., 2021; Hauzenberger et al., 2023) by defining the classification head h_{ρ_i} for the target attribute i to predict the corresponding label y_{ρ_i}. The adversarial loss follows a min-max optimization, aiming to decrease the predictability of the protected attribute while increasing task performance. Following previous studies, we utilize the gradient reversal layer to turn this min-max optimization into a minimization problem, formulated below.

$$\mathcal{L}_{\rho_i} = \mathrm{CE}(h_{\rho_i}(z), y_{\rho_i})$$

Search bias regularization loss Following Zerveas et al. (2022b), we utilize the CODER framework (Zerveas et al., 2022a) to optimize both task and bias mitigation in a list-wise fashion using the ListNet loss (Cao et al., 2007). For the main task, $\mathcal{L}_{task}$ is the KL-divergence between the distributions of ground-truth relevance labels (y) and the predicted scores ($\hat{s}$), defined over N candidate documents. Denoting ground-truth labels and predicted scores as y and $\hat{s}$, accordingly, $\mathcal{L}_{task}$ is formulated as:

$$\mathcal{L}_{task} = \mathrm{D}_{\mathrm{KL}}(\phi(y)||\phi(\hat{s})) = -\sum_{j=1}^{N} \phi(y)_j \log \frac{\phi(\hat{s})_j}{\phi(y)_j}$$

where ϕ refers to the softmax function applied to the values. We enforce the neutrality of retrieved documents, with a list-wise regularization term added to the task loss. The fairness term is similarly formulated with KL-divergence, namely between the distribution of the neutrality scores of the target labels (y_{ρ_i}), and the one of the predicted scores $\hat{s}$, formulated below:

$$\mathcal{L}_{\rho_i} = \mathrm{D}_{\mathrm{KL}}(\phi(\hat{s})||\phi(\rho)) = -\sum_{j=1}^{C} \phi(\hat{s})_j \log \frac{\phi(y_{\rho_i})_j}{\phi(\hat{s})_j}$$

As indicated in Eq. 5, in both classification and IR scenarios the bias mitigation loss is scaled with hyperparameter λ and added to the corresponding task loss.

4 Experiment Setup

Datasets We conduct our classification experiments on three datasets: The **BIOS** (De-Arteaga et al., 2019) dataset which contains short biographies used to predict a person's job. The name and any indication of the person's gender in the biography are omitted. The dataset labels are 28 occupations for the task, and two protected attribute classes (female/male). The second dataset is **FDCL18** (Founta et al., 2018b) for hate speech detection, containing a set of tweets each classified as *hateful*, *abusive*, *spam*, or *none*. Following previous studies (Sap et al., 2019; Ravfogel et al., 2020), we assign race dialect labels of *African American* and *White American* to FDCL18 using the probabilistic model developed by Blodgett et al. (2016). The third dataset is **PAN16** (Rangel et al., 2016) containing a set of tweets accompanied by the labels of gender and age of the authors. The task's objective is to predict whether another user is mentioned in a tweet. PAN16 provides the binary task classes of *mention*, and *no mention*, two gender labels, and five age groups. For the IR task, we use the fairness-sensitive queries dataset **MSMARCO**$_{Fair}$ (Rekabsaz et al., 2021). The queries are from the MSMARCO Passage Retrieval collection (Nguyen et al., 2016), which contains 215 queries and 8,841,822 passages. We use 158 words related to each of the two protected attribute classes (female/male) following (Zerveas et al., 2022b), to calculate the neutrality of each document following (Rekabsaz et al., 2021). The details of the neutrality criteria can be found in appendix D.

LMs and Training. For the classification benchmarks, we conduct the experiments on three LMs namely, BERT-Base (Devlin et al., 2018), BERT-Mini(Turc et al., 2019) and RoBERTa (Liu et al., 2019). In all experiments, $\mathcal{L}_{task}$ is realized by cross-entropy and binary cross-entropy loss for binary classes and adversarial training to remove attributes. We train our models with a parallel strategy. The adversarial head consists of an ensemble of 5 networks for each attribute, and each network consists of two fully connected layers with Tanh activation in between. The overall loss of the adversarial is scaled by $\lambda = 1$. For the IR task, following Zerveas et al. (2021) we conduct the the experiments on DistilBERT (Sanh et al., 2019). As explained in Section 3.2, The loss function ($\mathcal{L}_{task}$) is realized by ListNet (Cao et al., 2007), and fairness is achieved through the fairness regularization loss ($\lambda\mathcal{L}_{\rho}$) (Zerveas et al., 2022b). Each baseline model is trained with several values of the regularization terms λ, while CONGATER is trained only once with $\lambda = 20$.

Models Considering that parameter-wise CONGATER is the same as adapters we choose our baselines as follows: **FT** finetunes all parameters of the LM on the task with no debiasing objective for classification. **FTADV**. **ADP** uses a standard adapter network and trains the adapter only on the task. **ADPADV** uses the same adapter architecture but trains it with task and attribute removal objectives simultaneously. The complete details of our hyperparameters setting and training procedure are explained in Appendix B

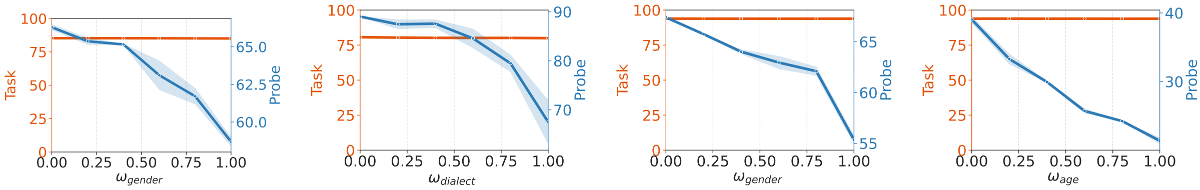
Metrics We evaluate the performance of the classification models on the core task using the accuracy metric. Following the previous works of Kumar et al. (2023) and Hauzenberger et al. (2023) we measure attribute information using strong probing networks. For each model, we train 5 independent classification heads (two-layer feed-forward layer with a Tanh activation) for 30 epochs to extract and predict the target attribute. We report the average performance of the probes in terms of balanced/macro accuracy (average of per-class accuracy scores). This evaluation measures how much information about a given attribute still exist in the model and can be recovered. Balanced accuracy has the benefit of better reflecting the performance of the methods when considering minority groups, particularly given the unbalanced distributions over protected labels in the datasets. For the IR task, we use the Mean Reciprocal Rank (MRR@10) of the top 10 retrieved documents as a metric for the core task evaluation and Normalized Fairness of Retrieval Results (NFaiRR@10) from the top 10 retrieved documents as the fairness metric of the model (Rekabsaz et al., 2021). NFaiRR metric normalizes the pre-query FaiRR score over the ideal FaiRR achieved from a background set of documents allowing comparable results across queries (detail in appendix D). To account for possible variations, we report our results as the average of 3 independently-trained models, and the report mean and standard deviation of the results.

5 Results and Discussion

We start with the results on the classification tasks, and continue with discussing our observations on the IR benchmark.

Model	Type	BIOS		FDCL18		PAN16-Gender		PAN16-Age	
		Task↑	Probe↓	Task↑	Probe↓	Task↑	Probe↓	Task↑	Probe↓
BERT-BASE	FT	$84.6_{0.4}$	$67.3_{0.8}$	$81.0_{1.0}$	$92.9_{1.8}$	$93.6_{1.8}$	$69.6_{0.8}$	$93.6_{1.8}$	$42.3_{0.9}$
	ADP	$84.3_{0.1}$	$67.0_{0.1}$	$80.0_{0.1}$	$93.3_{0.4}$	$92.4_{0.1}$	$70.7_{0.1}$	$92.4_{0.1}$	$42.4_{0.}$
	FTADV	$84.0_{0.3}$	$60.8_{0.2}$	$81.0_{1.0}$	$84.4_{4.0}$	$92.4_{0.8}$	$59.8_{0.7}$	$92.4_{0.8}$	$31.3_{1.1}$
	ADPADV	$84.2_{0.1}$	$61.9_{0.5}$	$79.8_{0.3}$	$75.6_{0.5}$	$92.2_{0.1}$	$\mathbf{54.2}_{0.4}$	$92.1_{0.1}$	$21.7_{0.1}$
	CONGATER	$\mathbf{85.0}_{0.1}$	$\mathbf{58.7}_{0.4}$	$81.0_{0.2}$	$\mathbf{67.5}_{0.6}$	$\mathbf{93.8}_{0.1}$	$55.3_{0.5}$	$\mathbf{93.8}_{0.1}$	$\mathbf{21.3}_{0.6}$
ROBERTA-BASE	FT	$84.5_{0.4}$	$66.2_{0.7}$	$80.6_{0.4}$	$93.2_{1.2}$	$\mathbf{98.5}_{0.1}$	$63.6_{0.4}$	$\mathbf{98.5}_{0.1}$	$22.7_{0.8}$
	ADP	$84.3_{0.1}$	$67.3_{0.7}$	$80.0_{0.6}$	$94.0_{0.6}$	$98.2_{0.1}$	$62.8_{0.4}$	$98.1_{0.1}$	$31.9_{0.1}$
	FTADV	$84.1_{0.3}$	$61.6_{0.3}$	$80.5_{1.0}$	$83.6_{1.9}$	$98.2_{0.1}$	$52.0_{0.9}$	$98.2_{0.1}$	$24.1_{1.4}$
	ADPADV	$84.0_{0.1}$	$62.9_{0.1}$	$80.0_{0.5}$	$79.7_{0.3}$	$98.1_{0.1}$	$\mathbf{53.7}_{0.7}$	$98.0_{0.1}$	$22.3_{1.0}$
	CONGATER	$\mathbf{84.8}_{0.1}$	$\mathbf{61.4}_{1.0}$	$\mathbf{81.4}_{0.2}$	$\mathbf{73.9}_{0.8}$	$98.4_{0.1}$	$55.4_{0.8}$	$98.4_{0.1}$	$\mathbf{22.2}_{0.1}$
BERT-MINI	FT	$\mathbf{82.4}_{0.1}$	$65.7_{0.2}$	$79.9_{1.0}$	$92.4_{0.6}$	$\mathbf{91.5}_{0.1}$	$65.4_{0.8}$	$\mathbf{91.5}_{0.1}$	$40.9_{0.4}$
	ADP	$82.1_{0.1}$	$65.2_{0.4}$	$81.3_{0.1}$	$93.5_{0.8}$	$81.5_{0.1}$	$65.5_{0.4}$	$81.6_{0.1}$	$37.4_{1.4}$
	FTADV	$81.7_{0.1}$	$60.4_{0.4}$	$79.3_{0.8}$	$81.4_{2.1}$	$90.3_{0.4}$	$58.6_{0.8}$	$90.3_{0.4}$	$27.9_{1.8}$
	ADPADV	$82.1_{0.2}$	$61.4_{0.6}$	$80.7_{0.1}$	$75.9_{1.2}$	$81.3_{0.1}$	$\mathbf{53.1}_{0.1}$	$81.1_{0.1}$	$\mathbf{21.7}_{0.3}$
	CONGATER	$81.7_{0.4}$	$\mathbf{59.2}_{0.1}$	$\mathbf{81.9}_{0.1}$	$\mathbf{62.5}_{0.5}$	$90.2_{0.1}$	$56.4_{0.1}$	$89.9_{0.2}$	$21.8_{0.2}$

Table 1: Results of BERT-Base, RoBERTa-Base and BERT-Mini models on three datasets and four attributes. The CONGATER sensitivity parameter is set to fully debiasing ($\omega = 1$) to have comparable results with the baselines



(a) BIOS (Gender) (b) FCDL18 (Dialect-race) (c) PAN16 - Gender (d) PAN16 - Age

Figure 2: Results of the CONGATER models using BERT-Base when increasing the gating sensitivity ω from 0 (no effect) to 1 (full effect). Each trained model is evaluated multiple times on the various ω values adjusted at inference time. The left/right y-axis corresponds to the task performance and attribute probing results, respectively. The results show the continuous reduction in the information presence of the target concept, as ω increases.

5.1 Classification Tasks

Table 1 reports the results of the baselines, as well as CONGATER at full bias mitigation power ($\omega = 1$). As shown, the fully activated CONGATER models contain less information about the attribute in comparison to all fine-tuned baselines (FT*) on all datasets across all models while having lower task performance on Pan16 and Bios for BERT-Mini model and slightly worse on ROBERTA-Base and Pan16 dataset. In comparison to adapter-based models, CONGATER on all models (except for BERT-mini and BIOS dataset) performs on par or better on the main task and removes more information on 3 out of 4 attribute removal tasks. On the BERT-mini and Pan16 dataset we observed that adapter-based models are superior to CONGATER in terms of attribute information removal but (ADP*) are not able to achieve satisfactory task performance which we assume is due to lack of enough learning capacity of adapters for this particular task. Overall our experiments and observations indicate that CONGATER is able to perform the task better than baselines while enhancing information removal from the embeddings of the network.

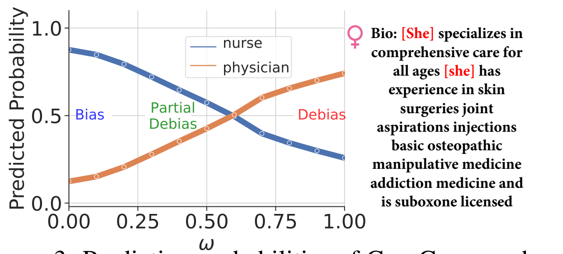


Figure 3: Prediction probabilities of CONGATER when gradually increasing ω, for a female physician's biography, incorrectly classified as a nurse in the initial state. The figure illustrates how changing the strength of gender removal affects the model's decision, providing a higher degree of interpretability through controllability.

We also investigate how much information about each attribute exists in the model when changing ω and its influence on the task performance. We investigate by changing the ω and retraining the probes. Figure 2 shows the task performance and probing results for the CONGATER for the BERT-Base model when increasing ω parameters. The reported result for each probing (each attribute with a specific ω) is the mean and standard deviation of the 5 probes applied to 3 independently trained models. The task performance and probing results are shown in orange (left y-axis) and blue (right y-axis), respectively. The results of the RoBERTa and BERT-Mini are reported in Appendix C.1.

Type	λ	MSMARCO MRR@10↑	MSMARCO NFaiRR@10↑
FT	0.0	$0.234_{0.002}$	$0.904_{0.001}$
	5.0	$0.183_{0.003}$	$0.943_{0.002}$
	10	$0.142_{0.002}$	$0.955_{0.001}$
	20	$0.079_{0.003}$	$0.972_{0.000}$
ADP	0.0	$0.230_{0.002}$	$0.898_{0.001}$
	5.0	$0.147_{0.003}$	$0.933_{0.000}$
	10	$0.082_{0.000}$	$0.949_{0.001}$
	20	$0.023_{0.002}$	$0.965_{0.001}$
CONGATER			
$\omega = 0.0$	0.0	$0.234_{0.003}$	$0.903_{0.001}$
$\omega = 0.4$	-	$0.227_{0.002}$	$0.917_{0.000}$
$\omega = 0.8$	-	$0.208_{0.001}$	$0.942_{0.000}$
$\omega = 1.0$	20.0	$0.168_{0.007}$	$0.970_{0.000}$

Table 2: Results of DistilBERT-Base on **MS-MARCO**$_{Fair}$ benchmark.

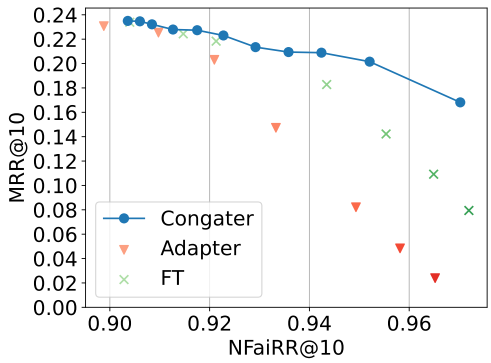


Figure 4: Fairness-performance trade-off between FT, ADP, and CONGATER. For baselines, each point refer to a new model training with color intensities indicating the degree of the regularization coefficient λ. CONGATER is trained only once, and each point indicates the evaluation according to an ω value.

Consistent across all datasets and attributes, we observe that increasing ω leads to a continuous decrease in the presence of the corresponding attribute, until reaching the lowest probing balanced accuracy at $\omega = 1$. This continuous attribute removal is achieved while maintaining task performance. On the whole, our results points to the ability of CONGATER to impose graded control over an attribute at inference time.

As an example of how continuous controllability at inference time can enhance interpretability, and provide a higher level of model transparency to end-users, figure 3 depicts how changes in ω_{gender} can influence models' decision probability about a female physician who is labeled as nurse by the biased model $\omega = 0$. Figures 15-19 in Appendix C provide more examples of such positive/negative effects from the studied datasets.

In Appendix C.2, we investigate the simultaneous bias control of the gender and age attributes of PAN16 using the multi-concept CONGATER and the changes in the behavior of the model by investigating the alterations in uncertainty (C.4) and prediction labels (C.5). We also study the effect of CONGATER adversarial bias mitigation effect on empirical fairness metrics in Appendix C.3.

5.2 Search Result Bias Mitigation Benchmark

Table 2 shows the performance and search results fairness results of the baseline models trained with different regularization coefficients (λ), as well as the CONGATER model trained once with $\lambda = 0$ and $\lambda = 20$ according to the post-hoc strategy. The same results with more data points are illustrated in Figure 4. To achieve a higher degree of fairness in baseline models, we increase λ and retrain the models, while in CONGATER, we simply increase the value of ω at test time.

Consistent with previous studies Zerveas et al. (2022b); Rekabsaz et al. (2021), increasing λ leads to a decrease in task performance (fairness-performance trade-off) in all models. However, interestingly the CONGATER with highest degree of fairness ($\omega = 1$) achieves MRR@10 = 0.168 performance, which is around 2.1 times higher than the FT with the same level of fairness (at $\lambda = 20$). It is noticeable that the performance of ADP– as the modularized approach – radically deteriorates for high values of λ. In addition to higher performance perseverance, CONGATER is able to monotonically navigate between the states (from the one with the least to the most fairness scores), enabling an effective control over the fairness-performance trade-off at inference time. This functionality of CONGATER can be leveraged in practice by designing a personalized control *knob* in the hands of the end-users and practitioners, which empowers them to adjust the level of fairness/neutrality in search results.

6 Conclusion

We introduce CONGATER, a gated module enhanced with a novel controllable activation function, which enables continuous adjustment of the information flow of an attribute at inference time. We conduct two sets of experiments on the classification task to remove attribute information, and on an IR benchmark to increase neutrality of retrieved documents. Our results show that in both experiments CONGATER successfully isolates and filters out attribute information with the least harm to task performance. In addition, we showed that CONGATER is able to continuously traverse between the biased and debiased states, enhancing personalization and interpretability through controllability.

7 Ethical Considerations and Limitations

A limitation of our work concerns the lack of completeness in the definition of the concept and attributes provided by the datasets. In particular, gender in all datasets including BIOS, PAN16, and MSMARCO is limited to the binary female/male, lacking an inclusive and nuanced definition of gender. Similarly in FDCL18, we consider only two dialects of *African American* and *White American*, while clearly this definition is limited and non-inclusive. Furthermore, as in previous work (Sap et al., 2019; Ravfogel et al., 2020; Zhang et al., 2021; Kumar et al., 2023), the labels of this protected attribute are assigned through a probabilistic model, and hence the dataset might not represent the nuances and traits of the real-world.

The second limitation regards the degree of the generalization of our method with respect to various deep learning architectures (such as CNNs), as our definition and experiments are constrained to the use of transformer-based models.

The Third limitation regarding the generalization of the method is the multi-attribute setting for CONGATER over any possible number of concepts or a subset of them. We conduct our experiments with a focus on only one attribute and introduce a possible fusion method and a preliminary result. Our multi-attribute experiment is only conducted on one dataset with two attributes of gender and age, particularly due to the lack of availability of suitable datasets. We note that the conclusion provided in the paper should be viewed to the extent of these experiments, and further studies (as well as more suitable datasets) are required to achieve a more comprehensive picture of this topic.

As a general limitation shared with the other related studies on in this domain, we should note that the aim of representation disentanglement optimizations is to reduce the information about a particular concept inside model embeddings with attributes based on the *observed data*. These data-oriented approaches might lack effective generalization, particularly when the model is evaluated on other domains or out-of-distribution data.

What we are proposing is a single model capable of handling both biased decisions, partially debiased decisions, and unbiased decisions which creates much more flexibility for the end user. Any misuse of the proposed method in tasks where fairness for the users are high priority such as job recommendations is not the intention of the authors and is considered one of the dangers of the proposed model.

References

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6331–6336.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

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

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, ICML ’07, pages 129–136, New York, NY, USA. Association for Computing Machinery.

Pierre Colombo, Pablo Piantanida, and Chloé Clavel. 2021. A novel estimator of mutual information for learning to disentangle textual representations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6539–6550.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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

Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Inf. Process. Manag.*, 57(6):102377.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018a. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018b. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Christian Ganhör, David Penz, Navid Rekabsaz, Oleg Lesota, and Markus Schedl. 2022. Unlearning protected user attributes in recommendations with adversarial training. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2142–2147, New York, NY, USA. Association for Computing Machinery.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023.

Skyler Hallinan, Alisa Liu, Yejin Choi, and Maarten Sap. 2023. Detoxifying text with marco: Controllable revision with experts and anti-experts. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 228–242. Association for Computational Linguistics.

Wenjuan Han, Bo Pang, and Ying Nian Wu. 2021a. Robust transfer learning with pretrained language models through adapters. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 854–861.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021b. 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. Association for Computational Linguistics.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021c. Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2760–2765. Association for Computational Linguistics.

Lukas Hauzenberger, Shahed Masoudian, Deepak Kumar, Markus Schedl, and Navid Rekabsaz. 2023. Modular and On-demand Bias Mitigation with Attribute-Removal Subnetworks. In *Findings of the Association for Computational Linguistics: ACL (Findings of ACL)*.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, pages 113–122. ACM.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning*, volume 97, pages 2790–2799. Proceedings of Machine Learning Research.

Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266.

Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2023. Grep-biasir: A dataset for investigating gender representation-bias in information retrieval results. In *Proceeding of the ACM SIGIR Conference On Human Information Interaction And Retrieval (CHIIR)*.

Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW 2017, Portland, OR, USA, February 25 - March 1, 2017*, pages 417–432. ACM.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. Association for Computational Linguistics.

Sachin Kumar, Biswajit Paria, and Yulia Tsvetkov. 2022. Gradient-based constrained sampling from language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2251–2277, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Oleg Lesota, Navid Rekabsaz, Daniel Cohen, Klaus Antonius Grasserbauer, Carsten Eickhoff, and Markus Schedl. 2021. A modern perspective on query likelihood with deep generative retrieval models. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, pages 185–195.

Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. 2022. Scaling & shifting your features: A new baseline for efficient model tuning. In *NeurIPS*.

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

Rabeeh Karimi Mahabadi, James Henderson, and Sebastian Ruder. 2021. Compacter: Efficient low-rank hypercomplex adapter layers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 1022–1035.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2022. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6):115:1–115:35.

Marco Morik, Ashudeep Singh, Jessica Hong, and Thorsten Joachims. 2021. Controlling fairness and bias in dynamic learning-to-rank (extended abstract). In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4804–4808. ijcai.org.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Harrie Oosterhuis. 2022. Computationally efficient optimization of plackett-luce ranking models for relevance and fairness (extended abstract). In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5319–5323. ijcai.org.

Nicolas Papernot, Abhradeep Thakurta, Shuang Song, Steve Chien, and Úlfar Erlingsson. 2021. Tempered sigmoid activations for deep learning with differential privacy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 9312–9321.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. Adapterfusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503.

Jonas Pfeiffer, Sebastian Ruder, Ivan Vulić, and Edoardo Maria Ponti. 2023. Modular deep learning. *arXiv preprint arXiv:2302.11529*.

Lianhui Qin, Sean Welleck, Daniel Khashabi, and Yejin Choi. 2022. Cold decoding: Energy-based constrained text generation with langevin dynamics. In *Advances in Neural Information Processing Systems*.

Prajit Ramachandran, Barret Zoph, and Quoc V Le. 2017. Searching for activation functions. *arXiv preprint arXiv:1710.05941*.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al.*, pages 750–784.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256.

Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. 2017. Learning multiple visual domains with residual adapters. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30.

Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal biases in retrieved contents: Measurement framework and adversarial mitigation of BERT rankers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 306–316.

Navid Rekabsaz and Markus Schedl. 2020. Do neural ranking models intensify gender bias? In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2065–2068.

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

Alexis Ross, Tongshuang Wu, Hao Peng, Matthew E. Peters, and Matt Gardner. 2022. Tailor: Generating and perturbing text with semantic controls. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 3194–3213. Association for Computational Linguistics.

Andreas Rücklé, Gregor Geigle, Max Glockner, Tilman Beck, Jonas Pfeiffer, Nils Reimers, and Iryna Gurevych. 2021. AdapterDrop: On the efficiency of adapters in transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7930–7946.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95, Online only. Association for Computational Linguistics.

Tianxiao Shen, Jonas Mueller, Regina Barzilay, and Tommi S. Jaakkola. 2020. Educating text autoencoders: Latent representation guidance via denoising. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8719–8729. PMLR.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3398–3403.

Gabriel Stanovsky, Noah A Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected attention layers for efficient adaptation in multi-task learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5986–5995. PMLR.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. 2022. Extracting latent steering vectors from pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 566–581. Association for Computational Linguistics.

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-read students learn better: On the importance of pre-training compact models. *arXiv preprint arXiv:1908.08962v2*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000—6010. Curran Associates, Inc.

Liwen Wang, Yuanmeng Yan, Keqing He, Yanan Wu, and Weiran Xu. 2021. Dynamically disentangling social bias from task-oriented representations with adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3740–3750.

Jiacheng Xu, Shrey Desai, and Greg Durrett. 2020. Understanding neural abstractive summarization models via uncertainty. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6275–6281.

Dian Yu, Zhou Yu, and Kenji Sagae. 2021. Attribute alignment: Controlling text generation from pretrained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2251–2268.

George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2021. Coder: An efficient framework for improving retrieval through contextual document embedding reranking.

George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022a. Coder: An efficient framework for improving retrieval through contextual document embedding reranking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10626–10644.

George Zerveas, Navid Rekabsaz, Daniel Cohen, and Carsten Eickhoff. 2022b. Mitigating bias in search results through contextual document reranking and neutrality regularization. In *Proceedings of the 45th*

International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '22, page 2532–2538, New York, NY, USA. Association for Computing Machinery.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*.

Xiongyi Zhang, Jan-Willem van de Meent, and Byron Wallace. 2021. Disentangling representations of text by masking transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 778–791, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 629–634.

Mengjie Zhao, Tao Lin, Fei Mi, Martin Jaggi, and Hinrich Schütze. 2020. Masking as an efficient alternative to finetuning for pretrained language models. In *Empirical Methods in Natural Language Processing*, pages 2226–2241. Association for Computational Linguistics.

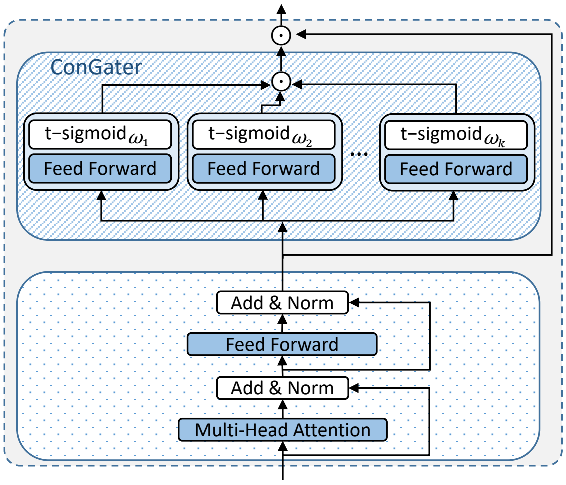


Figure 5: The simple proposed CONGATER for multi-attributes. The fusion gate is defined as the element-wise multiplication of the individual gates

A Multi-Concept CONGATER

In this section, we investigate a version of CONGATER with multi-attribute. Figure 5 depicts this variation, where the individual gates are combined with element-wise multiplication to form a multi-attribute gating vector. This vector is then used for the self-gating mechanism of CONGATER in Eq. 3. The training procedure of the multi-attribute setting is exactly the same as the single-attribute one. During inference, the gating sensitivity of each attribute can be changed independently. Our experiment results on a two-attribute setting are provided in Appendix C.2.

B Additional Experiment Setup

In the FDCL18 dataset, we use the TwitterAAE model (Blodgett et al., 2016) to assign racial dialect classes. The TwitterAAE model predicts four racial classes, *African American*, *White American*, *Hispanic*, and *Others*. We labeled a tweet as *African American* or *White American* if the prediction score was greater than 0.5. For the PAN16 dataset, following (Sap et al., 2019) we balanced the task labels and sampled 200K data. The age groups of this dataset are 18-24, 25-34, 35-49, 50-64, and 65+. Table 3 gives a summary of the whole dataset, training, validation and test data which was used during the training and evaluation of the network.

We select train, validation, and test sets randomly from the pool of data with the proportions 63:12:15 for BIOS, 63:12:15 for FDCL18, and 80:5:15 for PAN16. We use the validation set for hyperparameter tuning, and the best result on the validation set is evaluated on the test set for the final results. The validation and test sets in all datasets follow the same distribution as the whole dataset. To address the unbalanced dataset and the potential problems in adversarial training, we apply upsampling only on the *training sets* of BIOS and FDCL18 datasets, to balance the protected attribute labels within each task label. For instance, genders are balanced in the dentist class by repeating the data items of the minority subgroup.

Models are trained on the task for 15 epochs and for the post-hoc models an additional 15 epochs of adversarial training. Details of hyperparameters are reported in Table 4 and the number of parameters of the models is reported in Table 5.

Each layer that is mentioned in the hyperparameter section has the same width as the original Bert-Base model which in our experiment is (768). In our experiment we trained all of the transformer blocks but in general, any training method on the task completely depends on the designer's will and what CONGATER offers is an extension to the original model with additional training which leads to controllability.

C Additional Results

C.1 Other LMs

Figures 6 and 7 show the results using BERT-Mini and RoBERTa-Base LMs, respectively. We observe the same control capability of attribute as we discussed in the main paper for the other to LMs as well.

C.2 Multi-Attribute Results

To examine the ability of multi-concept CONGATER (introduced in Appendix A), we train the model on the two attributes of PAN16, where each concept has its gating sensitivity parameter (ω_{gender} and ω_{age}). The evaluation results on task performance, gender probing, and age probing are reported in Figure 8 for a specific combination of ω_{gender} and ω_{age}. As shown, the multi-concept CONGATER model can maintain the task performance, as ω values change, while the presence of concept information gradually decreases. We also observe that changing one ω has an influence on the probing results of the other concept, indicating the probable correlations between the concepts. By simultaneously increasing both ω there exist combinations where information about both gender and

Dataset	Classes	Attribute	Train	Validation	Test
FCDL18	4	Dialect	52,352	4,736	5,888
BIOS	28	Gender	294,784	38,016	88,640
PAN16	2	Gender&age	160,000	10,048	30,016

Table 3: Summary of the datasets and their protected attribute(s)

	Dataset			
	FCDL18	PAN16	BIOS	MSMARCO
	Embedding			
batch size	64	64	64	64
number of workers	8	8	8	5
Max document/query length	40	40	120	32
Padding	max length	max length	max length	max length
	Model & Training			
Model Type	Base/Mini/Roberta	Base/Mini/Roberta	Base/Mini/Roberta	DistillBERT
CONGATER Bottleneck factor	8	12	2	4
λ	1	1	1	max = 20
λ warm-up scheduler	3	3	3	-
loss function	Cross Entropy	Cross Entropy	Cross Entropy	ListNet loss
Optimizer	AdamW	AdamW	AdamW	RAdam
task lr	2×10^{-5}	2×10^{-5}	2×10^{-5}	1.7×10^{-6}
weight decay	0.01	0.01	0.01	
adv lr	1×10^{-4}	1×10^{-4}	1×10^{-4}	-
probe lr	1×10^{-4}	1×10^{-4}	1×10^{-4}	-
task/adv dropout	0.1	0.1	0.1	0.0
lr scheduler	Cosine Decay	Cosine Decay	Cosine Decay	-
train epochs	15	15	15	10
adv epochs	15	15	15	10
probe epochs	30	30	30	-
task head layer	1	1	1	-
adv head layer	2	2	2	-
probe head	2	2	2	-
adv/probe activation	Tanh	Tanh	Tanh	-

Table 4: ADP and CONGATER Hyperparameters used to fine tune and remove information from the network on different datasets

Parameter Count	FT	ADP	CONGATER
Total Number of Parameter	109,485,316	116,577,028	116,577,028
Attribute(single) Parameters	109,485,316	7,094,788	7,094,788
Adversarial Training Module (%)	100	6.0	6.0

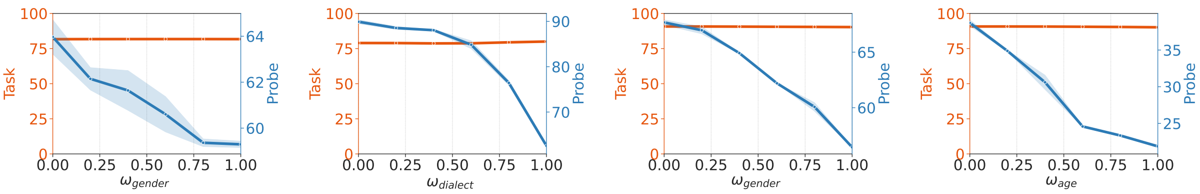
Table 5: BERT-base number of parameters specification for each method

age has the minimum value (e.g., $\omega_{age} = 1$ and $\omega_{gender} = 0.1$). This initial experiment shows the benefits of CONGATER for multi-attribute control, as well as the challenges in this area, suggesting further investigations for future work.

C.3 Study of Fairness

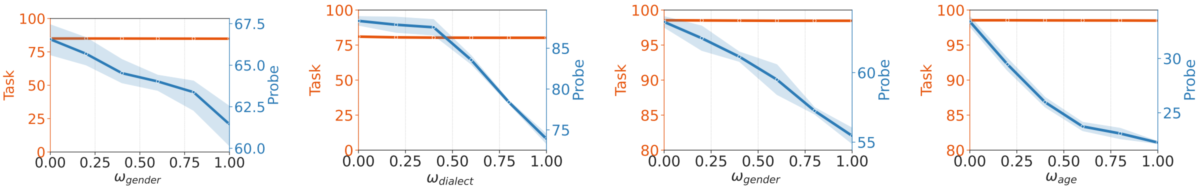
removing attributes has been the focus of studies for several purposes such as bias mitigation, privacy preservation, and fairness improvement. Researchers focused on removing harmful information such as gender or race from the network as the main cause of societal biases to improve fairness with regard to minority groups (Han et al., 2021c; Mehrabi et al., 2022; Shen et al., 2022). In this section, we investigate the effect of CONGATER with regard to empirical fairness metrics. In particular, we utilize GAP (Shen et al., 2022) as the evaluation metric of empirical fairness. The GAP metric for a binary attribute is defined as:

$$GAP = \sqrt{\frac{1}{|Y|} \sum_{y \in Y} (GAP_{a,y}^{TPR})^2} \qquad (6)$$



(a) BIOS (Gender) (b) FCDL18 (Dialect-race) (c) PAN16 - Gender (d) PAN16 - Age

Figure 6: Results of the ConGater models using BERT-Mini. See Figure 2 for full explanation.



(a) BIOS (Gender) (b) FCDL18 (Dialect-race) (c) PAN16 - Gender (d) PAN16 - Age

Figure 7: Results of the ConGater models using RoBERTa-Base. See Figure 2 for full explanation.

where $GAP^{TPR}_{a,y}$ is the difference between the True Positive Rate (TPR) for each class a, and GAP is the normalized difference between binary sub-populations Y.

Figure 9 shows the mean and standard deviation of the GAP results for three independent-trained runs of the ConGater models using the BERT-Base. The results are overall consistent with the core message in previous studies (Shen et al., 2022), indicating that removing concept information (known as representational fairness) is not necessarily correlated with GAP (empirical fairness). We however observe that the average GAP for BIOS and FCDL18 slightly decrease, where similar to the probing results, the changes appear continuously between the initial to the target state. Overall, the continuous controllability of ConGater allows for the choice of the state with the desired representational or empirical fairness, given the context of the task and/or a user's preference.

C.4 Shift in Model Uncertainty

Model uncertainty is another core aspect in models, providing additional information about the model's decision behavior during prediction. We investigate how uncertainty changes during changing ω. Following previous studies (Lesota et al., 2021; Xu et al., 2020), we measure model/prediction uncertainty as the entropy of the predicted probability distribution, namely:

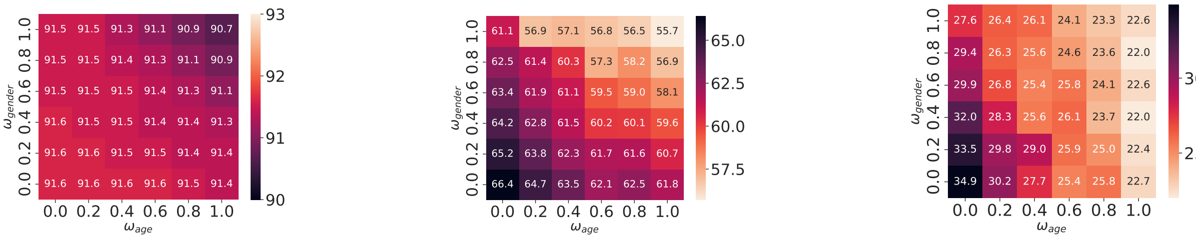
$$\text{Uncertainty}(X) = -\sum_{j=1}^{|X|} p(X_j) \log p(X_j) \quad (7)$$

where X is the predicted probability distribution of a single data point provided by a model, and defined over the categorical space of $|X|$ classes. For each state of the ConGater models, we calculate this measure of uncertainty over the task's predictions for each data point in the respective test set and average over the results.

As depicted in Figure 10, the overall uncertainty of the BERT-Base model constantly changes (increases). We observe the same pattern in the uncertainty values when calculated on sub-populations of each dataset. Looking at the result of BIOS, we can see that the models behave more deterministic (less uncertainty) when it comes to Males and this uncertainty increases as we increase the sensitivity parameter. As for the FCDL18, we can see that for *African American* sub-population the model has much lower uncertainty at $\omega = 0$ compare to $\omega = 1$ where the uncertainty of *African American* is much closer to White. On the other hand in PAN16, we observe that model uncertainty is lower for the Female sub-population and at the beginning, the uncertainty for gender decreases then increases for higher values of ω. As for the age, we observe the same pattern as we increased the ω value of age but the $\omega_{gender} = 0$

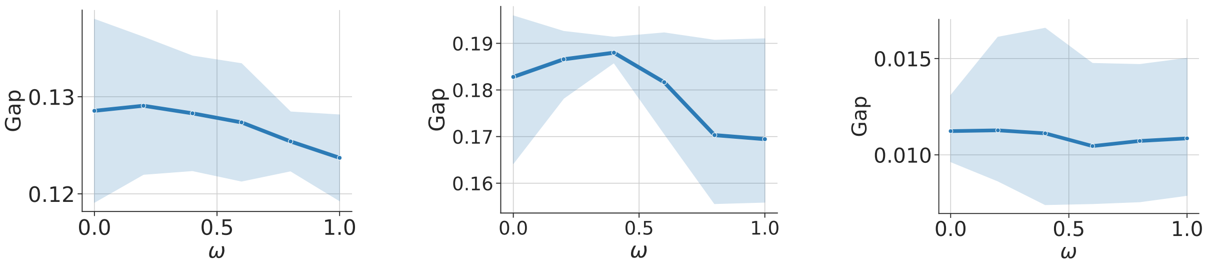
C.5 Prediction Flips

Despite the fact that average classification accuracy – as shown in the previous experiments – only marginally fluctuates during (binary or continuous) concept erasure, we observe that the behavior of the model in terms of the predicted probabilities considerably changes among data points. In some



(a) Task Accuracy (b) Gender Probes (c) Age Probes

Figure 8: Results of the multi-concept CONGATER on the two attributes of PAN16 task using BERT-Mini, as changing the gender and age ω simultaneously. (a) Task performance with accuracy. (b) Balanced accuracy of gender probes. (c) Balanced accuracy of age probes.



(a) BIOS (b) FCDL18 (c) PAN16

Figure 9: Mean and standard deviation of Gap metric for Three independent BERT-base models with parallel training method

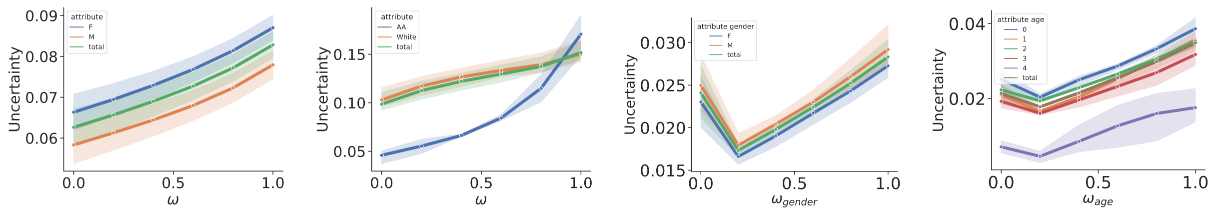
cases, the changes in predicted probabilities become so pronounced that, changing a concept's degree of presence completely alters the model's decision (predicted class). In what follows, we investigate the effect of partial concept erasure on the prediction behavior of the model as we change the gating sensitivity parameter ω.

Figure 11 shows the statistics of the changes in the predicted classes by the CONGATER model when increasing the corresponding ω parameter(s). Figure 11a tracks the percentage of the predicted labels for the data points in the BIOS test set, predicted as *Nurse* by the base model ($\omega = 0$). Figure 11b reports the same on FCDL18 for the *Abusive* predictions. Figures 12, 13 14 depict the results for more labels on these datasets, as well as the ones for PAN16. As shown, the changes in the predicted labels continuously increase as we increase the ω value, indicating that the decision making of the model continuously changes and more predicted labels from the initial state flip. This continuous change is consistent with concept removal results across the three datasets, demonstrating the capability of CONGATER in gradually changing its predictions when moving from the initial state to the target state. To gain a more fine-grained view of this topic, we further investigate the changes in model uncertainty in Appendix C.4.

D NFaiRR

we use Normalized Fairness of Retrieval Results (NFaiRR) to calculate the neutrality score which is fomulated in equation 8. In equation 8 FaiRR is the fairness metric similar to other studies (Kulshrestha et al., 2017; Fabris et al., 2020) is calculated by finding the importance of attribute in relation to its retrieved position. Also Ideal FaiRR is used to normalized the score which is considered as the best possible fairness result one can get by reordering the documents. For more details of the score please refer to (Rekabsaz et al., 2021).

$$NFaiRR(L, \hat{s}) = \sum_{q \in Q} \frac{FaiRR_q(L)}{IFaiRR_q(\hat{s})} \quad (8)$$

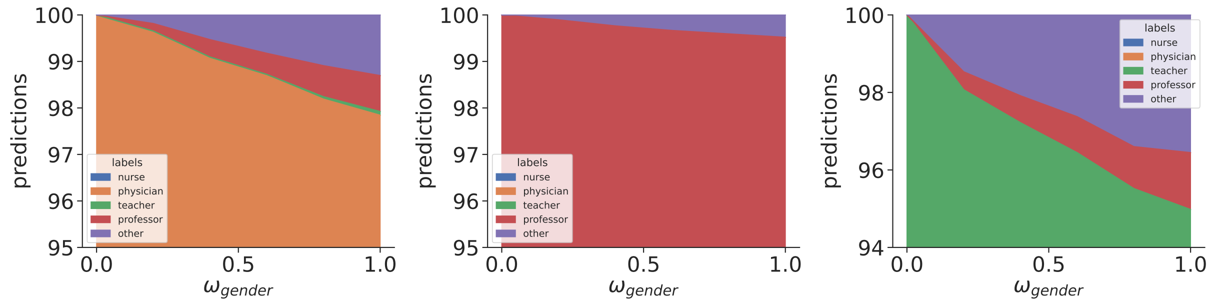


(a) BIOS (b) FCDL18 (c) PAN16-Gender (d) Pan16-Age

Figure 10: Mean of the test data points' uncertainty values, defined as the entropy of the predicted probability distributions provided by the CONGATER models. Increasing sensitivity parameter(s) results in continuous changes in model uncertainties.

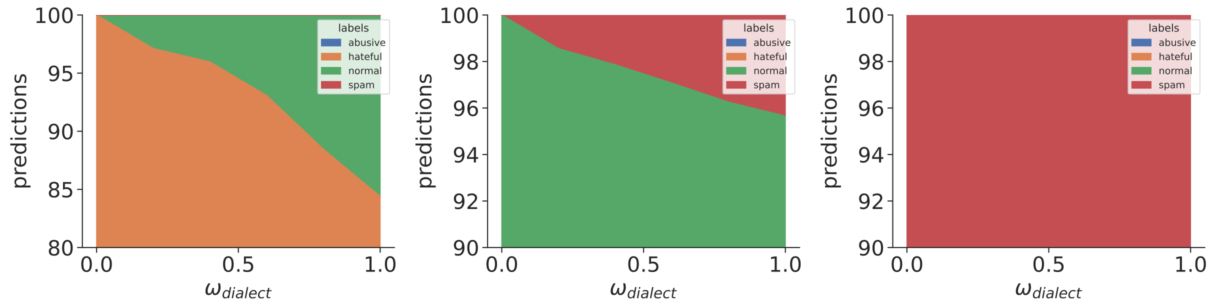
(a) *Nurse* in BIOS (b) *Abusive* in FCDL18

Figure 11: Percentage of the data points with the predicted label in the original ($\omega = 0$) model that remain on the prediction, when increasing ω.



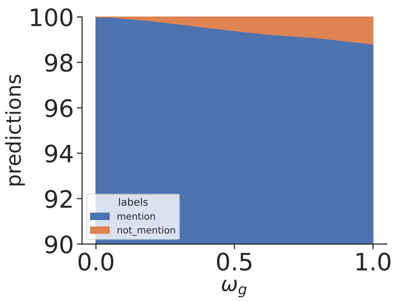
(a) Physician (b) Professor (c) Teacher

Figure 12: percentage of the predicted labels among the data points that are initially predicted as the mentioned label by the initial model during the partial concept removal of the BIOS dataset using CONGATER.

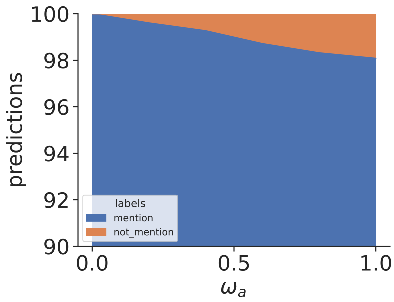


(a) Hateful (b) Normal (c) Spam

Figure 13: percentage of the predicted labels among the data points that are initially predicted as the mentioned label by the initial model during the partial concept removal of the FCDL18 dataset using CONGATER.

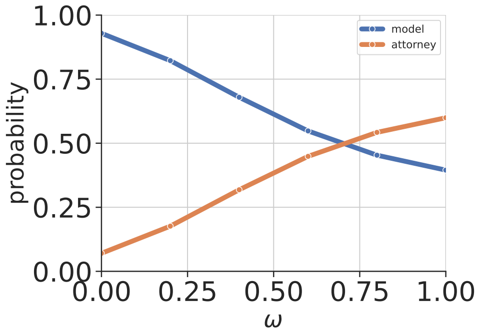


(a) Gender

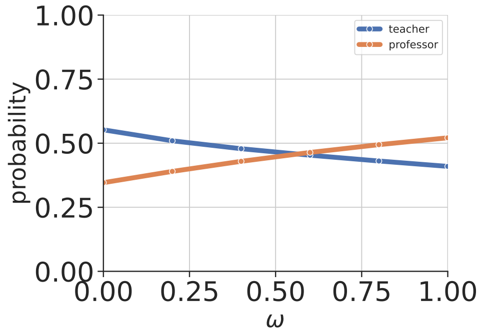


(b) Age

Figure 14: Changes in the Mention prediction as we change ω of the gender or age concept in the respective model.

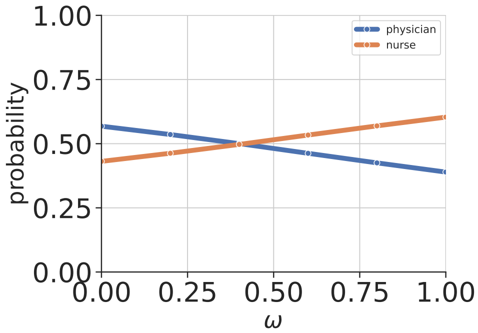


(a) Male attorney

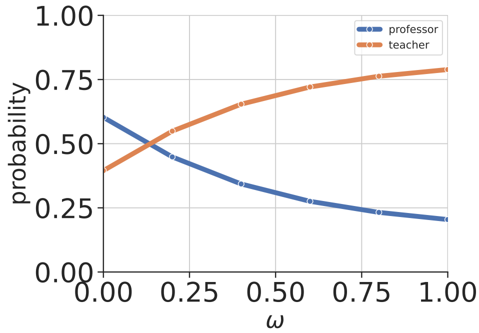


(b) Female professor

Figure 15: Figures show the **positive** effect of CONGATER on the prediction probability for the labels as we increase ω value and remove information about gender in BIOS dataset (a) A male attorney predicted wrongly with the initial model ($\omega = 0$) and switches to the attorney as we increase omega value. **Bio**: "[he] has been stated previously senator conrad has complete confidence in s ability to enforce the law and serve the people of north dakota conrad wrote" (b) A female professor predicted teacher with initial model and switches to correct label as we remove information about gender. **Bio**: [she] has twelve years of teaching experience for the undergraduate students and postgraduate supervised many undergraduate projects and several master theses in aastmt and other universities also supervised several phd students in cooperation with ein shams university egypt

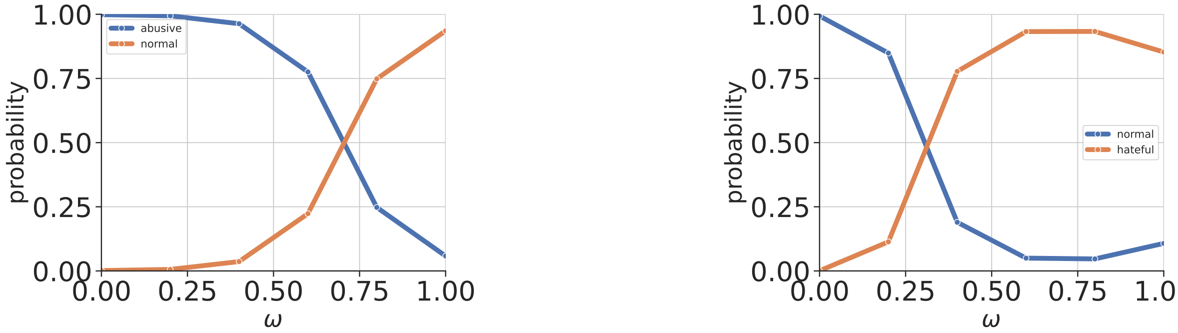


(a) Female Physician



(b) Male Professor

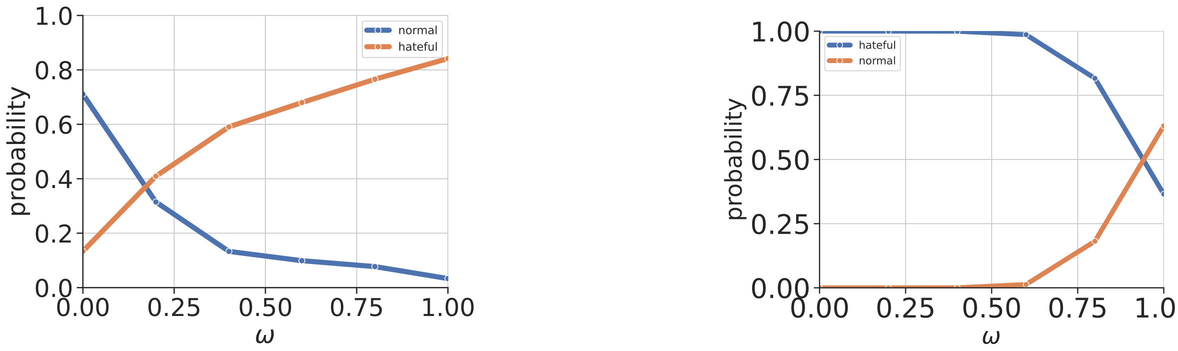
Figure 16: Figures show the **negative** effect of CONGATER on the prediction probability for the labels as we increase ω value and remove information about gender in BIOS dataset (a) A female physician predicted correctly with initial model switches to nurse as we increase omega value. **Bio**: "[she] has worked in both hospital and outpatient clinical settings has experience in internal medicine preventative medicine and urgent care specializes in botox treatments that help people to look their very best without invasive cosmetic surgery" (b) A male professor predicted correctly by the initial model misclassified as we increase ω value. **Bio**: "has been an active member of nacta since serves on the nacta journal editorial review board and has reviewed numerous conference oral and poster presentation abstracts has submitted teaching tips and received the bob gough outstanding teaching tip award and the nacta educator award has made both oral and poster presentations at nacta conferences"



(a) African American Normal tweet

(b) White person Hateful tweet

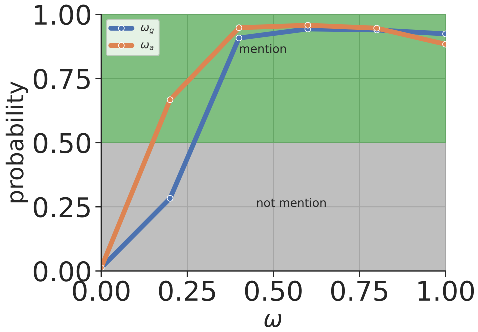
Figure 17: Figures show **positive** effect of CONGATER on the prediction probability for the labels as we increase ω value and remove information about gender in FCDL18 dataset (a) African American tweet labeled abusive at by the initial model, as we increase ω and remove dialect information labels flips to normal. **Tweet**: why is mother nature mad? somebody must have pissed her off! its rainingg hard af! (b) White person hateful tweet predicted normal by the initial model switches to the correct label as we increase ω value. **Tweet**: us is already doing something by backing some of the rebel groups. this is a proxy war afte...



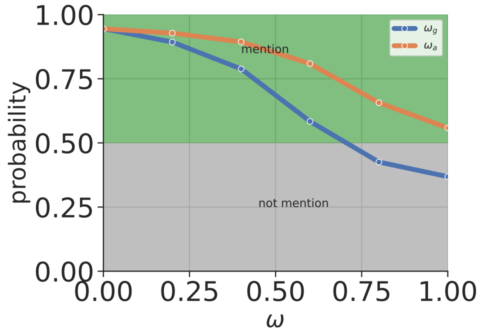
(a) White Normal tweet

(b) African American hateful tweet

Figure 18: Figures show **negative** effect of CONGATER on the prediction probability for the labels as we increase ω value and remove information about gender in the FCDL18 dataset (a) White normal tweet is correctly labels by the initial model. By increasing ω value and removing dialect information model predicts the tweet as hateful. **Tweet**: "yes because the person we voted for is keeping his promises, in spite of the lefts resistance! maga. today and" (b) African American hateful speech negatively switches to normal as we remove dialect information from the model. **Tweet**: "y'all be wanting gifts from y'all [curse word] when y'all mad just get me food ; "



(a) Female Mentioning someone



(b) Male mentioning someone

Figure 19: Figures show the prediction probability for the labels as we increase ω value and remove information about gender and age separately in PAN16 dataset. (a) Female Mentioning someone in the tweet but tagged as-not mentioned by the initial model **positively** switches to mention as we remove gender and age information separately. Removing any of the attributes alone results in the correct class of the prediction of this sample. **Tweet**: "happy charlie. stolen from #cats #tuxedokitty #blackandwhite" (b) Male mentioning someone was correctly labeled by the initial model with gender and age information. increasing ω_{gender} **negatively** influences model until predicts the label wrongly but ω_{age} also influences negatively, but not to the extent of label flip. **Tweet**: "(snort) yes. there it is, and there it is"