IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 15 December 2023, accepted 3 January 2024, date of publication 11 January 2024, date of current version 22 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3353056

RESEARCH ARTICLE

Reducing Bias in Sentiment Analysis Models Through Causal Mediation Analysis and Targeted Counterfactual Training

YIFEI DA[1], MATÍAS NICOLÁS BOSSA[1], ABEL DÍAZ BERENGUER[1], AND HICHEM SAHLI[1,2]
[1]Department of Electronics and Informatics, Vrije Universiteit Brussel, 1050 Brussels, Belgium
[2]Interuniversity Microelectronics Centre (IMEC), 3001 Leuven, Belgium

Corresponding author: Yifei Da (yda@etrovub.be)

ABSTRACT Large language models provide high-accuracy solutions in many natural language processing tasks. In particular, they are used as word embeddings in sentiment analysis models. However, these models pick up on and amplify biases and social stereotypes in the data. Causality theory has recently driven the development of effective algorithms to evaluate and mitigate these biases. Causal mediation was used to detect biases, while counterfactual training was proposed to mitigate bias. In both cases, counterfactual sentences are created by changing an attribute, such as the gender of a noun, for which no change in the model output is expected. Biases are detected and eventually corrected each time the model behavior differs between the original and the counterfactual sentence. We propose a new method for de-biasing sentiment analysis models that leverages the causal mediation analysis to identify the parts of the model primarily responsible for the bias and apply targeted counterfactual training for model de-biasing. We validated the methodology by fine-tuning the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model for sentiment prediction. We trained two sentiment analysis models using the Stanford Sentiment Treebank dataset and the Amazon Product Reviews, respectively, and we evaluated the fairness and prediction performances using the Equity Evaluation Corpus. We illustrated the causal patterns in the network and showed that our method achieves both high fairness and more accurate sentiment analysis than the state-of-the-art approach. Contrary to state-of-the-art models, we achieved a noticeable improvement in gender fairness without hindering sentiment prediction accuracy.

INDEX TERMS BERT, bias, causal mediation analysis, counterfactual training, large language models, sentiment analysis.

I. INTRODUCTION

Sentiment analysis aims to determine the sentiment strength from a textual source. It analyses people's opinions, evaluations, appraisals, attitudes, and emotions towards entities such as products, companies, services, individuals, issues, events, topics, and their attributes [1].

Although the human sentiment is complex, in sentiment analysis, it is usually represented with two (positive or negative) or three (positive, neutral, negative) polarities [2], 5-scale categorization, a real-valued score [3], and sometimes on multiple dimensions [4], such as valence (positive-negative) and arousal (excited-calm).

Sentiment classification may be accomplished on three levels of extraction: aspect or feature level, phrase level, and document level. Currently, there are three approaches to address the problem of sentiment analysis [5]: (1) lexicon-based strategies, (2) machine and deep-learning-based techniques, and (3) hybrid approaches. In this work, we consider deep-learning-based models.

Recently, several studies have identified various forms of bias in sentiment analysis models, raising concerns about the risk of propagating social biases against certain groups based on sociodemographic factors [6], [7]. The study in [8] finds

The associate editor coordinating the review of this manuscript and approving it for publication was Tao Huang.

evidence of bias against several marginalized queer identities in proprietary sentiment analysis tools such as Google, Amazon, and IBM. The study in [9] showed significant age bias encoded in the outputs of many sentiment analysis algorithms and word embeddings. Considering gender and race biases, the study of Kiritchenko and Mohammad [10] confirmed that most of the 219 sentiment analysis systems they evaluated tend to mark sentences involving one gender/race with higher intensity scores than sentences involving the other gender/race. In this work, we consider gender bias in sentiment analysis models. Our study is motivated by the existence of apparent gender differences in communication styles on the social web [11], including for expressing sentiment [12] and interpreting sentiment [13]. Sentiment analysis algorithms may tend to classify the sentiment of sentences differently for men and women.

Mitigating bias can be made at one of the three steps of the general machine learning pipeline: the training data, the learning procedure, and the output predictions, with three corresponding categories of bias mitigation algorithms: *pre-processing* approaches [14], *in-processing* approaches [15], and *post-processing* approaches [16]. *Pre-processing* approaches modify the input data representation to make the prediction outcomes fair. For instance, the authors of [17] adopted adversarial fine-tuning of word embedding-based sentiment analysis by re-training the model using adversarial examples dynamically generated by BiasFinder [18] a tool that can discover biased predictions in sentiment analysis systems. Using various Natural Language Processing (NLP) techniques, BiasFinder identifies words associated with a class of a characteristic (e.g., gender-specific words such as female names, ''she'', and ''her'') and generates new texts from these templates by mutating words.

In-processing (also called in-training) approaches revise the training of the state-of-the-art models to achieve fairness. More specifically, they apply fairness constraints or design an objective function considering the fairness of predictions [19]. Such approaches assume that the sensitive attributes information are accessible in the training samples and enforce fairness during the training process either by directly imposing fairness constraints and solving constrained optimization problems [20] or by adding penalization terms to the learning objective [21].

The present study proposes a bias reduction methodology in the Bidirectional Representation for Transformers (BERT) model [22] in a downstream sentiment classification task. Indeed, BERT has become one of the most important architectures for various natural language tasks, having generated state-of-the-art results in sentiment analysis. Recent work has shown that pre-trained BERT models capture social biases from the large amounts of text they are trained on in a downstream sentiment classification task [23], [24], [25].

The approach proposed in our paper falls into the *In-processing* category. We focus on fairness in sentiment classification, where the goal is to prevent discrimination against gender without compromising the utility of the classifier [26]. We propose a bias reduction methodology for sentiment analysis grounded in the theory of causal mediation analysis [27]. Our proposal is inspired by the causal mediation analysis of GPT-2 [28]. In this framework, the neural network is seen as a causal Directed Acyclic Graph (DAG) [29] and the direct and indirect effects [30] of individual neurons and layers are studied to identify which parts of a model are causally implicated in gender bias. The second part of our methodology is bias reduction based on counterfactual training of the model [21]. Contrary to previous approaches, our methodology targets specific neurons during the counterfactual training step to prevent deterioration of prediction accuracy.

The paper is organized as follows: Section II provides background knowledge and related works. Section III details the proposed methodology. Section IV details the technical setting and shows our experiments' results. Finally, we conclude in Section VI and provide several perspectives to extend this work.

II. BACKGROUND AND RELATED WORK

A. FAIRNESS METRICS

A concept that is intimately associated with bias is fairness. A system is considered ''fair'' when its outcomes are not discriminatory according to certain attributes, such as gender, race, country, nationality, and other social constructs. Researchers [6], [31] compiled several measures agreed upon in analyzing the bias problem in machine learning systems. These fairness metrics can be categorized into three main categories. *Group fairness* [32] partitions a population into groups defined by protected attributes (such as gender, caste, or religion) and seeks some statistical measure to be equal across groups. Within the *group-based fairness*, we can mention demographic parity [33], which minimizes the absolute difference in outcome distributions of all groups, and equalized opportunities [34], which optimizes towards an equal positive rate conditional on the target outcome. *Individual fairness* [35] requires that similar individuals for a given task receive a similar outcome. *Counterfactual fairness* [36] requires that the decisions provided by the model remain the same if the sensitive attribute were changed, e.g., in the case of the sentiment analysis system, what the outcome would have been if the female sensitive attribute had been changed to male. In the following, we give the formal definition of the different metrics.

In general, fairness is computed over the distribution $\left\langle X, A, Y, \hat{Y} \right\rangle$, with X being the samples, in our case a sentence, A a protected attribute; in the case of gender, its possible values are 0 for male or 1 for female, Y the true labels for the samples and $\hat{Y}$ the predicted sentiment polarity of an input sentence X, $(\hat{Y} = 1)$ if positive and $(\hat{Y} = 0)$ for negative polarity.

Demographic parity states that all the groups resulting from the different values of a protected class (i.e., gender) should receive the same rate of positive outcomes. This can be defined as $P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$, which is equivalent to equal positive sentiment for both males and females. Thus, the bias would be $bias = \left|P(\hat{Y} = 1|A = 0) - P(\hat{Y} = 1|A = 1)\right|$.

Individual fairness is formally defined as follows [35]. Let x and y be two individuals and d a similarity metric between them. Furthermore, let $M(x)$ be a function that assigns, to individuals, a probability distribution over the outcomes Y and D a distance function that measures the difference in the probabilities. A mapping M satisfies $(D, d) - Lipschitz$ property if for every $x, y \in V$ we have $D(M(x), M(y)) \leq d(x, y)$. This means the distance between the probabilities of certain outcomes assigned to two individuals must be no greater than their similarity distance.

Counterfactual fairness [36] is a notion of fairness derived from Pearl's causal model [29], where the sample X, the protected attribute A and the model output $\hat{Y}$ are assumed to be observables in a Structural Causal Model (SCM). The model is deemed counterfactually fair if A has no counterfactual causal influence on $\hat{Y}$, i.e., $\hat{Y}$ does not change if we can vary A in the SCM in a manner that other independent latent factors remain constant. In other words, ''in counterfactual fairness, for an individual, the model's prediction is considered fair if it is the same in the real world as it would be if the individual belonged to a different demographic group'' [36]. Formally, a predictor $\hat{Y}$ is counterfactually fair if $P(\hat{Y}_{A\leftarrow a}|X = x, A = a) = P(\hat{Y}_{A\leftarrow a'}|X = x, A = a)$, where $A \leftarrow a'$ indicates an intervention. If the equality holds for all a and a', then the classifier is deemed fair.

This work aims to reduce the counterfactual sentiment bias in BERT, following the counterfactual fairness metric.

B. DE-BIASING TECHNIQUES

In this section, we give a brief description of some de-biasing techniques. We refer readers to [37] for more de-biasing techniques.

1) DROPOUT REGULARIZATION

Dropout regularisation is typically used to reduce over-fitting; It has been investigated in [38] for reducing gendered correlations in BERT and ALBERT [39]. The authors investigated increasing the two dropout parameters of BERT (one for attention weights and another for hidden activations, both set to .10 by default) and performing an additional pre-training phase over a random sample of English Wikipedia. They did the same with ALBERT, increasing the dropout (set to zero in the public ALBERT models). Experimentally, they found that increased dropout regularisation reduces gender bias within these models. They hypothesized that dropouts' interruption of the attention mechanisms within BERT and ALBERT helps prevent them from learning undesirable associations between words.

2) COUNTERFACTUAL DATA AUGMENTATION (CDA)

Counterfactual data augmentation [40] is a methodology proposed to mitigate gender bias in neural NLP tasks by swapping bias attribute words (e.g., he/she) in a dataset. For each training instance, the method adds a copy with an intervention on its targeted protected attribute, replacing each with its counter while maintaining the same ground truth label. For example, to help mitigate gender bias, the sentence ''this man made me feel angry'' could be augmented to ''this woman made me feel angry''. The two sentences have the same structure,''<person> made me feel angry'', the same sentiment label ''negative'', but opposite gender. The result is a dataset of matched pairs with ground truth independent of the target distinction. The re-balanced corpus is then used for further training to de-bias a model. Lu et al. [40] made experiments for de-biasing pre-trained language models by performing an additional fine-tuning phase on the counterfactually augmented dataset. They argue that such an approach encourages learning algorithms to avoid picking up on the distinction.

3) SUBSPACE PROJECTION

Liang et al. in [41] proposed the SentenceDebias, a projection-based de-biasing technique that estimates a linear subspace for a particular type of bias. Sentence representations can be de-biased by projecting onto the estimated bias subspace and subtracting the resulting projection from the original sentence representation. A three-step procedure is followed for computing the bias subspace. First, a list of bias attribute words (e.g., he/she) is defined. Second, they contextualize the bias attribute words into sentences by finding occurrences of the bias attribute words in sentences within a text corpus. CDA is applied to generate a pair of sentences that differ from the bias attribute word for each found sentence. Finally, they estimate the bias subspace. For each sentence obtained during the contextualization step, a corresponding representation can be obtained from a pre-trained model. Principle Component Analysis is then used to estimate the principle directions of variation of the resulting representations set. The first K principle components are taken to define the bias subspace.

4) COUNTERFACTUAL TRAINING

Here, counterfactual training refers to training a model with an input pair, like in the simple framework for contrastive learning [42], and including fairness as a constraint or a penalization term in the minimization of the prediction loss. The authors of [43] investigated the efficacy of bias reduction in a language model during training by using as training loss a weighted sum between the cross-entropy loss and a new term, denoted as the Language Model de-biasing Loss

(LMD). LMD encourages the language model to equalize the probabilities of predicting gendered word pairs in the output, like "he" and "she" [43]:

$$LMD = \frac{1}{G}\sum_{i}^{G}\left|log\frac{\hat{y}_{f_i}}{\hat{y}_{m_i}}\right| \quad (1)$$

with f and m are a set of corresponding gender pairs, G is the size of the gender pairs set, and $\hat{y}$ indicates the output softmax probability. The authors considered a pre-trained GloVe (Global Vector) [44] word embedding and an LSTM as a language model. Their experiments considered only gender pairs to ensure that only gender information is neutralized and distribution over semantic concepts is not altered (see [43] for details).

Barikeri et al. [45] adopted the Language Model de-biasing Loss from [43] (Equation 1) for de-biasing DialoGPT the conversational language model [46]. Using CDA, the authors constructed REDDITBIAS, a conversational data set created from real-world conversations collected from Reddit's online discussion platform for bias evaluation and mitigation dedicated to conversational AI. They manually annotated for multiple societal bias dimensions: (i) religion (two different bias types), (ii) race, (iii) gender, and (iv) queerness. Using the constructed CDA dataset, they fine-tuned the DialoGPT model with a training loss defined as the weighted sum between the GPT's Causal Language Modeling loss and the LMD loss term.

Counterfactual training has also been used in [21] for reducing sentiment bias in two Transformer-XL [47] language models. A three-step curriculum training schema has been followed. First, the authors trained the language model using a regular cross-entropy loss for predicting the next token, given all the previous tokens. Second, using the language model, they train a multilayer perceptron (MLP) as a sentiment classifier using the extracted features from the language model. To label the data, they used the Google Cloud sentiment API4. In the third step, referred to as the "de-biasing step", with the fixed sentiment classifier from the previous step, they applied counterfactual training of the language model with an additional fairness loss. The loss function for an input sequence z during the third step is: $\mathcal{L}(\mathbf{z}) = \mathcal{L}_{LM}(\mathbf{z}) + \lambda.\mathcal{L}_{fairness}(\bar{h}(\mathbf{z}), \bar{h}(\tilde{\mathbf{z}}))$, with $\mathcal{L}_{LM}(.)$, the cross-entropy loss, $\tilde{\mathbf{z}}$ the counterfactual sequence of $\mathbf{z}$, $\mathcal{L}_{fairness}(.)$ the fairness loss, and λ a regularization parameter. The fairness loss is defined as the cosine distance of the embeddings from the language model:

$$\mathcal{L}_{fairness}(\bar{h}(\mathbf{z}), \bar{h}(\tilde{\mathbf{z}})) = 1 - \frac{\bar{h}^T(\mathbf{z}) \cdot \bar{h}(\tilde{\mathbf{z}})}{\| \bar{h}(\mathbf{z}) \| \| \bar{h}(\tilde{\mathbf{z}}) \|} \quad (2)$$

where $\bar{h}(.)$ is set as the average of the last two embedding vectors. They also proposed the *sentiment regularisation* for which the fairness loss is applied on the embedding from the hidden layer of the sentiment classifier, $\mathcal{L}_{fairness}(f_{sh}(\bar{h}(\mathbf{z})), f_{sh}(\bar{h}(\tilde{\mathbf{z}}))$. Figure 1 illustrates the third step process. Note that the full layers of the language model and the sentiment classifier are fine-tuned.

5) CAUSAL MEDIATION ANALYSIS

Causal mediation analysis defines causal effects as differences between counterfactual outcomes. It studies the change in a response variable following an intervention or mediation [30]. To formally introduce causal mediation, we use the notation and example given by Chi et al. in [48]. Let's consider, for example, health outcome, where an intervention is defined as treatment. Let $X_i(t)$ represent the potential mediator value for participant i if the participant's treatment status is $T_i = t$. Let $Y_i(t, x)$ denote the potential outcome value for participant i if $T_i = t$ and participant i has a mediator value $X_i = x$. The causal mediation effect for participant i captures the difference between the participant's observed outcome and a counterfactual outcome if the participant's treatment status remains the same, but the mediator value equals the value under the other treatment status [30]: $\delta_i(t) = Y_i(t, X_i(1)) - Y_i(t, X_i(0))$, where $t = 0, 1$. If $t = 0$, the term $Y_i(t, X_i(1))$ is the counterfactual and $Y_i(t, X_i(0))$ is observed. $\delta_i(0)$ is termed *pure indirect effect*. When $t = 1$, the term $Y_i(t, X_i(1))$ is observed, and $Y_i(t, X_i(0))$ is counterfactual. $\delta_i(1)$ is termed *total indirect effect*. The mediator values $X_i(t)$ are determined by the researcher [30] and reflect, e.g., a value of clinical, policy relevance, or a neuron in a neuronal network, as will be discussed in the following.

Causal mediation analysis to interpret neural network architecture to detect biases was introduced by Vig et al. [28]. They represent the neural network as a Directed Acyclic Graph (DAG) [29], with a common ancestor to all nodes, the input, and a common descendent to all nodes, the output and treating internal model components, i.e., neurons or attention heads, as mediators between the model's inputs and outputs. Vig et al. [28] studied how grammatical gender bias effects are mediated via different model components of GPT2, aiming at analyzing the role of individual neurons or attention heads in mediating these effects. Their study focuses on understanding how information flows through different model components and comparing the language model's probability for a male pronoun to that of a female. Thus, for each input unit (e.g., a sentence), they estimate the causal effect of an intervention (e.g., a text edit) by comparing the model output under the intervention to the output given the original input. By distinguishing between direct and indirect effects, they measure how much of the total effect of gender edits on gender bias flows through a specific component (indirect effect) or elsewhere in the model (direct effect). This framework allows the evaluation of the causal contribution of different mediators (individual neurons or attention heads) to gender bias.

In this work, we build upon the causal mediation analysis of [28] and the counterfactual training of [21] and propose

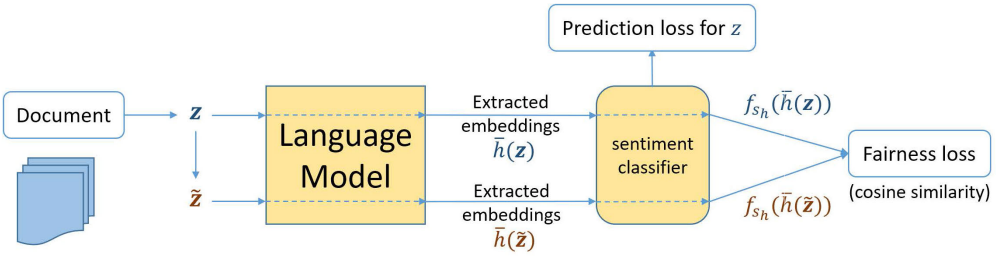

FIGURE 1. Counterfactual training for model de-biasing. $f_{sh}(.)$ denotes the hidden layer of the sentiment classifier (adapted from [21]).

a de-biasing approach to BERT models. As in [28], our approach views the neural network, in our case a fine-tuned BERT model for sentiment prediction, as a causal DAG) [29]. Using counterfactual data augmentation, we study the direct and indirect effects [30] of individual neurons and layers to identify which parts of the model are causally implicated in gender bias. The second step of our methodology is bias reduction based on counterfactual training as in [21]. Contrary to [21], which trains the full layers during the de-biasing fine-tuning, our methodology targets specific layers, i.e., layers with a causal contribution to gender bias, as to prevent deterioration of prediction accuracy.

III. METHODOLOGY

A. CAUSAL INTERPRETATION OF NEURAL NETWORKS

The causal mediation analysis allows us to analyze how information flows through the neutral network while we are interested in understanding the effect of counterfactual mediation on classification decisions.

1) CAUSAL MEDIATION ANALYSIS OF NEUTRAL NETWORKS

As explained in Section II-B5, functional dependencies of neurons in most neural networks can be described with a DAG, where the edges go from the input to the output neurons [27] (see Figure 2). Given a neural network M, each of its internal neurons X_i, considered as mediators, acquires a value equal to

$$X_i(\mathbf{z}) = x_i \tag{3}$$

as a response to an input $\mathbf{z}$ (see Figure 2.a). The value x_i could be affected directly by the input $\mathbf{z}$ or indirectly via some intermediate neurons.

For each input $\mathbf{z}$ in a given training set, we can create a counterfactual $\tilde{\mathbf{z}}$ by applying certain transformations, for example, by changing a word on the input sentence (see Section II-B2). We must also create the corresponding output to use the counterfactual input during the neural network testing. In our case, we are interested in transformations that do not affect the output. We assume that by replacing specific words in an input sentence $\mathbf{z}$ that includes information on a given protected attribute (e.g., gender), the predicted sentiment should not change.

An intervention on a given neuron (mediator) consists of setting its value to be equal to the value under the counterfactual sentence $\tilde{\mathbf{z}}$. Let's denote by $\tilde{x_i}$ the value of a given neuron X_i when the counterfactual $\tilde{\mathbf{z}}$ is used as input to the network (i.e. $X_i(\tilde{\mathbf{z}}) = \tilde{x_i}$), keeping the network's output prediction the same as for $\mathbf{z}$. The intervention on a given neuron (mediator) X_i is denoted as

$$do(X_i = \tilde{x_i}) \tag{4}$$

We write the values that the other neurons X_j $(j \neq i)$ acquire when setting X_i to the value $\tilde{x_i}$, as:

$$X_j(\mathbf{z}, do(X_i = \tilde{x_i})) = x_j^{X_i = \tilde{x_i}} \tag{5}$$

Figure 2.b illustrates this process.

2) DIRECT AND INDIRECT EFFECTS

By changing the network input value $\mathbf{z}$ to a counterfactual value $\tilde{\mathbf{z}}$ we can evaluate the total effect on the output,

$$\Delta y_{total} = |y - \tilde{y}|, \tag{6}$$

where $y = Y(\mathbf{z})$ and $\tilde{y} = Y(\tilde{\mathbf{z}})$ are the original and counterfactual outputs of the neural network, respectively.

The total effect and the (controlled) direct effect were used for interpreting black-box models [27] and to estimate bias using counterfactuals [49]. Here, we are interested in decomposing the total effect into (natural) direct and indirect effects to determine which neurons are mainly responsible for the bias [28]. We consider each neuron as a potential mediator at a time. The direct effect is the part of the effect that passes through the network when the particular neuron is fixed to the same value for both inputs, i.e., original and counterfactual. The indirect effect is the effect on the output due to the intervention on the neuron. More precisely:

i. Let $\tilde{y}^{X_i = x_i} = Y(\tilde{\mathbf{z}}, do(X_i = x_i))$ be the intervened counterfactual output. The input of the neural network is the counterfactual sentence $\tilde{\mathbf{z}}$, but the neuron X_i is set to the value it had when the input was the original sentence $\mathbf{z}$, i.e., x_i. The intervened counterfactual output will be used to evaluate the direct effect, i.e., the effect not mediated through X_i, as this neuron is unaffected by the input sentence change.

ii. Let $y^{X_i = \tilde{x_i}} = Y(\mathbf{z}, do(X_i = \tilde{x_i}))$ be the intervened output. The input is the original sentence $\mathbf{z}$, and the neuron X_i is set to the counterfactual value $\tilde{x_i}$. The intervened output will be used

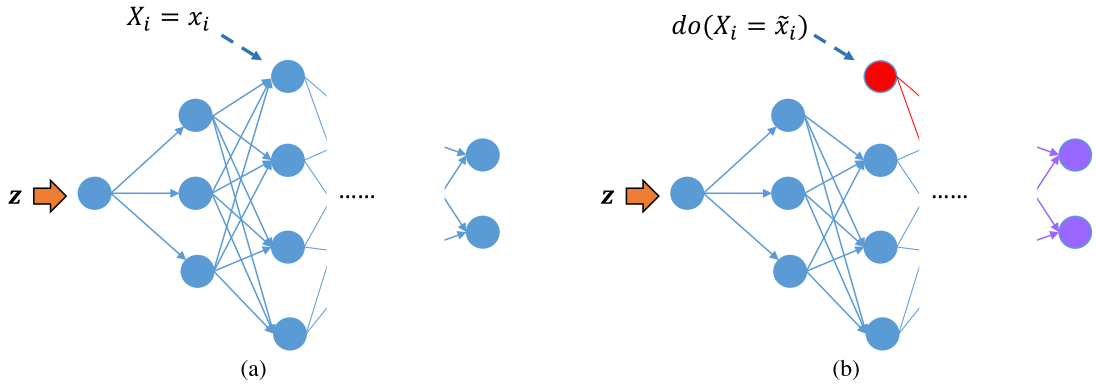

FIGURE 2. Representation of a neural network by a causal graph: The graphic in Panel (a) depicts a neural network in its normal state, in which the internal neurons are considered as mediators; Panel (b) shows the network's altered state after intervention on the neuron X_i.

to evaluate the indirect effect, i.e., the effect only mediated through X_i.

For a given input sentence $\mathbf{z}$, the natural indirect effect (NIE) is the absolute difference between the original output, y, and the intervened output $y^{X_i=\tilde{x_i}}$:

$$\Delta y^{X_i}_{indirect} = |y - y^{X_i=\tilde{x_i}}|. \tag{7}$$

The natural direct effect (NDE) is the absolute difference between the original output, y, and the counterfactual output, $\tilde{y}^{X_i=x_i}$:

$$\Delta y^{X_i}_{direct} = |y - \tilde{y}^{X_i=x_i}|. \tag{8}$$

Figure 3 illustrates the above-described causal mediation analysis.

The average natural direct and indirect effects over the CDA training set are denoted as *aNDE* and *aNIE*, respectively. For a given neuron, the bigger the average indirect effect (*aNIE*) is, the stronger the causal relationship is between the neuron and the classification bias. The neurons with the most significant *aNIE*, compared to the *aNIE*'s of the other neurons, have a large causal effect on the output. They, therefore, are responsible for the bias because the intervention on these neurons causes a significant effect on the model output. For each neural network layer, we calculate the sum, *SaNIE*, of the *aNIE*'s of its neurons and consider it as *target layer*, i.e., with a stronger causal relationship with classification bias, if $SaNIE > t_{SaNIE}$, t_{SaNIE} being a threshold set by analyzing the *SaNIE*'s of all layers.

B. DE-BIASING VIA COUNTERFACTUAL TRAINING OF TARGET LAYERS

This section introduces our approach to reducing gender bias in sentiment analysis. However, our approach can be applied to other biases on sensitive attributes such as ethnic groups, color, and country. For this study, we selected the pre-trained BERT model, fine-tuned for sentiment classification downstream task.

Given an input training sentiment data, $\mathcal{D} = \{\mathbf{z}_i, y_i\}_{i=1}^N$, we construct a counterfactual data augmentation set, $\mathcal{D}_{CDA}$, by creating for each training instance a copy with an intervention on its targeted protected attribute, replacing each with its counter, while maintaining the same ground truth label, $\mathcal{D}_{CDA} = \{\tilde{\mathbf{z}}_i, y_i\}_{i=1}^N$. For example, considering *gender* as the targeted protected attribute, given a sentence $\mathbf{z}$ =''The situation makes the boy feel sad'', the singular noun ''the boy'' is the sensitive attribute value that we are interested in. Replacing it with a counterfactual value ''the girl'', we get a counterfactual sentence $\tilde{\mathbf{z}}$ =''The situation makes the girl feel sad''. Our objective is to train the model towards reducing counterfactual gender bias in sentiment. We want to ensure that the model produces similar sentiment for $\mathbf{z}$ and $\tilde{\mathbf{z}}$. Specifically, we would like to ensure counterfactual fairness.

Our approach proceeds in three steps:

Step 1 We fine-tune the pre-trained BERT model for sentiment classification downstream task by adding a classifier g and training it on the sentiment dataset $\mathcal{D}$.

Step 2 We apply the causal mediation analysis of Section III-A to find the neurons that strongly affect the classification bias. For this, we follow the procedure described in III-A2 and evaluate for each neuron the average (over the CDA training set) indirect effect (aNIE) of gender bias on the model output. The bigger the aNIE, the stronger the causal relationship between the neuron and the classification bias. Finally, we select the BERT layers with large *SaNIE* as causally implicated in gender bias. The selected layers are denoted as *target layers*.

Step 3 The ''de-biasing step''. We apply counterfactual training; however, unlike [21], we only fine-tune the BERT *target layers* and $g(.)$, and freeze the other layers. Indeed, as highlighted in [28], gender bias effects are sparse and concentrated in a small part of the network. Moreover, we hypothesize that if we train the entire model, the neurons not responsible for bias also change, which could bring an unwanted decrease in accuracy or other model performances.

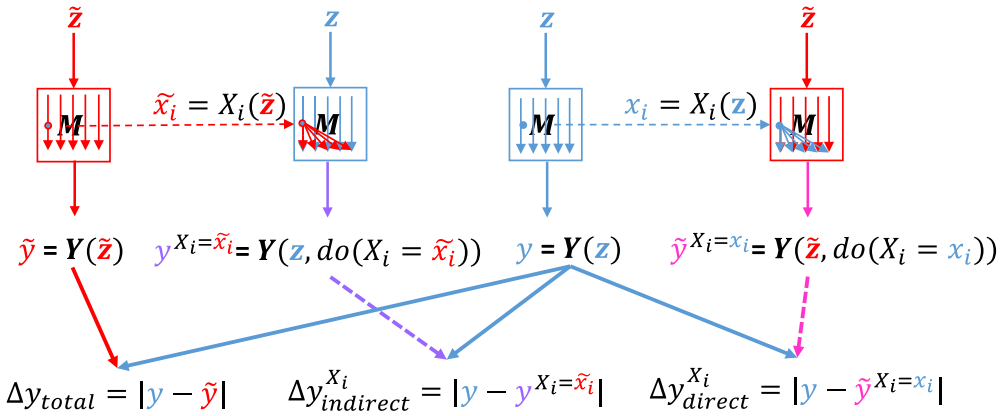

FIGURE 3. Causal mediation analysis to determine the natural direct and indirect effect of the neuron X_i in a neural network M: Blue paths correspond to the model response to the original input sentence and red paths to the counterfactual sentence. The total effect is the difference between the model response to the original (blue) and counterfactual (red) input sentences. The counterfactual response is split into two complementary responses obtained by interventions (purple and pink), and the original response is compared with each to obtain the direct and indirect effects. The indirect effect through neuron X_i is obtained by intervening only in this neuron with the counterfactual value while keeping the original input sentence, leading to the model response in purple. The direct effect is obtained by providing the counterfactual sentence to the model but intervening neuron X_i with the original value, leading to the model response in pink.

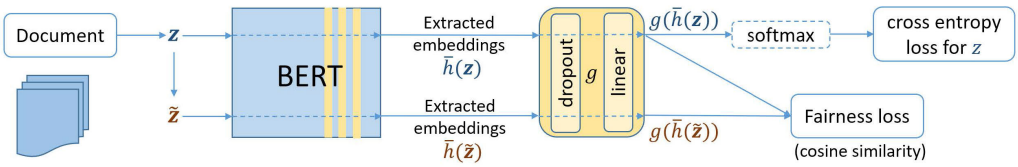

FIGURE 4. De-biasing step with counterfactual training: Two colors cut apart the BERT model; the blue parts are frozen layers, i.e., their values do not change during the fine-tuning; the yellow parts are trainable, i.e., the values of these neurons change during the fine-tuning. Note that causal mediation analysis of Step 2 is needed to identify these layers.

The training loss for counterfactual training is:

$$\mathcal{L}(\mathbf{z}) = \mathcal{L}_{classification}(\mathbf{z}) + \lambda.\mathcal{L}_{fairness}(\mathbf{z}, \tilde{\mathbf{z}}) \quad (9)$$

with, $\mathcal{L}_{classification}(.)$ the cross-entropy loss, λ a regularization parameter, and $\mathcal{L}_{fairness}(.)$ the *sentiment regularization* fairness loss of [21]:

$$\mathcal{L}_{fairness}(\mathbf{z}, \tilde{\mathbf{z}}) = 1 - \frac{g^T(\bar{h}(\mathbf{z})) \cdot g(\bar{h}(\tilde{\mathbf{z}}))}{\| g(\bar{h}(\mathbf{z})) \| \| g(\bar{h}(\tilde{\mathbf{z}})) \|} \quad (10)$$

where $g(h(\cdot))$ is the embedding from the sentiment classifier.

The process of the proposed counterfactual training is depicted in Figure 4 and detailed in Algorithm 1.

IV. EXPERIMENTAL RESULTS

A. MODEL ARCHITECTURE

The BERT model [22] is a deep bidirectional network built using Transformers [50]. We select the BERT-Base[1] as the underlying BERT model. BERT-base consists of 12 Transformer blocks. Each transformer block contains an attention layer and a feedforward layer. A transformer block (denoted here as a layer) takes in a list of token embeddings and produces the same number of embeddings with the same hidden size (or dimensions) of 768 on the output. The output of the final transformer block of the [CLS] token is used as the feature to feed the sentiment classifier (see Figure 5).

As indicated in Section III-B.**Step 1**, we first fine-tune the pre-trained BERT model for sentiment classification downstream task by adding a classifier g, consisting of a dropout and a linear layer, followed by a softmax activation that outputs a probability distribution over the target classes at the last layer. To fine-tune the model, we use two benchmark datasets for sentiment analysis, the SST-2 [51], and the Amazon Product Reviews [52] (see Section IV-B).

The fine-tuned models will be used in the second and third steps of our approach (Section III-B), namely,

1) For the causal mediation analysis process described in Section III-A, for which the neuron activations we are analyzing are at the level of the output of the last dense layer per encoder layer, i.e., the 768 neurons that correspond to the [CLS] token, in all 12 layers (in total 9, 216 neurons), depicted as blue boxes in

[1] https://github.com/google-research/bert

Algorithm 1 Pseudo-Algorithm Counterfactual Training

Input: $\mathcal{D} \rightarrow$ Dataset
$BS \rightarrow$ Batch Size
$\epsilon \rightarrow$ Number of Epochs
$M \rightarrow$ Fine-tuned BERT and Classifier $g(.)$
$l \rightarrow$ List of *Target Layers*
$\lambda \rightarrow$ Regularization parameter
#Freeze BERT layers that are not Target Layers.
foreach *layer in M.BERT* **do**
  if *layer is not in l* **then**
    Freeze *layer*;
#Fine-tune M
while *epochs* $\leq \epsilon$ **do**
  for *Sampled minibatch* $\{\mathbf{z}_k, \mathbf{y}_k\}_{k=1}^{BS}$ **do**
    forall $k \in \{1, \cdots, BS\}$ **do**
      #Apply counterfactual data augmentation to $\mathbf{z}$;
      $\tilde{\mathbf{z}} \leftarrow CDA(\mathbf{z})$;
    For $\mathbf{z}$ perform forward propagation and compute $\mathcal{L}_{classification}(.)$;
    For $\tilde{\mathbf{z}}$ perform forward propagation and compute $\mathcal{L}_{fairness}(.,.)$;
    Compute $\mathcal{L}(.) = \frac{1}{BS}\sum_{k=1}^{BS}[\mathcal{L}_{classification}(.) + \lambda.\mathcal{L}_{fairness}(.,.)]$;
    Update M to minimize $\mathcal{L}(.)$
Output: De-biased model M

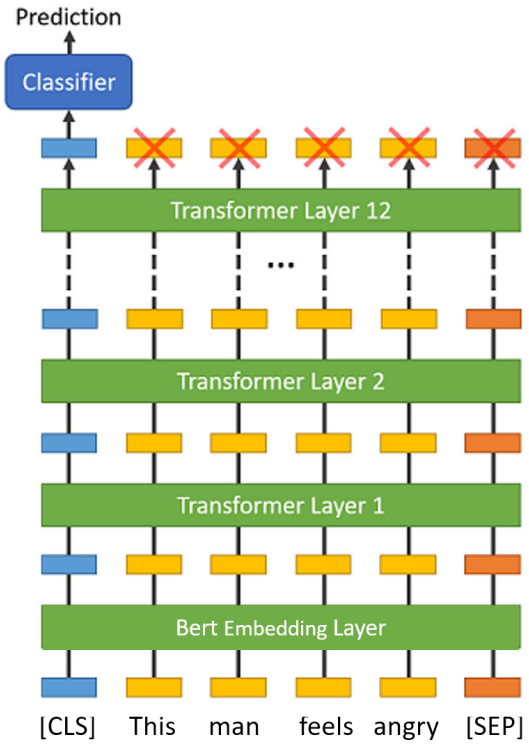

FIGURE 5. Illustration of the BERT model architecture: Small boxes are the neuron activations at the output level of the last dense layer per encoder layer. For our work, we consider only the 768 neurons that correspond to the [CLS] token in all 12 layers, denoted by blue boxes.

Figure 5. The output of this step (see Section III-A2) is the *Target Layers* i.e. layers with large *SaNIE* as causally implicated in gender bias;
2) Having selected the *Target Layers*, we apply Algorithm 1 for counterfactual training.

B. DATA SETS

We selected two benchmark datasets for our experiments.

1) DATA SETS DESCRIPTION

The SST-2 data set [51] is built out of the Stanford sentiment treebank, containing 11, 855 sentences extracted from movie reviews. Each of them is fine-grained labeled (very positive, positive, neutral, negative, very negative). SST-2 is a binary labeled version of the Stanford sentiment treebank, in which neutral reviews are removed, very positive and positive reviews are labeled as positive, negative, and very negative reviews are labeled as negative. SST-2 contains 9613 sentences split into 6920/872 for training and validation, respectively.

The Amazon Product Reviews dataset [52] includes 142.8 million reviews spanning May 1996 - July 2014. We selected it as it has been evaluated for gender bias in [53], who analyzed more than 11 million reviews from Amazon and Yelp, to answer whether customer reviews might hold stereotypic gender bias that GloVe (Global Vector) [44] word-embedding algorithm can learn and propagate. The authors compared the similarity of male/female names (e.g., Tom, Paul versus Joan, Lisa) and male/female words (e.g., he, him versus she, her) with both family and career words (e.g., enterprising, professional versus family, children) and as well as positive and negative attributes (e.g., rational, innovative versus impulsive, conformist). Their results indicated that the bias is driven because of the higher association of men with positive attributes and women with negative attributes.

We randomly selected 60, 000 sentences from the Amazon Product Reviews dataset for our experiments. Each sentence is rated with a score from 0 to 5, which is used as a sentiment score. For our experiments, we defined 2-polarity sentiment as follows: reviews rated 0, 1 or 2 are labeled as ''negative'', while reviews rated 3, 4 or 5 are labeled as ''positive''. The selected 60, 000 sentences were split into 51000/9000 for training and validation, respectively. These data sets are used to fine-tune the BERT model for sentiment analysis (see Section IV-E).

2) COUNTERFACTUAL DATA AUGMENTATION

For creating the counterfactual data augmentation, we applied the same procedure as in [21] and constructed two gender CDA datasets $\mathcal{D}_{CDA}^{SST-2}$ and $\mathcal{D}_{CDA}^{AMZR}$ based on the SST-2 and Amazon Product Reviews, respectively.

TABLE 1. Examples of counterfactual augmentation for the SST-2 and Amazon Product Reviews datasets.

Original Sentence	Counterfactual Sentence	Label
SST-2		
• the cast is phenomenal, especially the **women**.	the cast is phenomenal, especially the **men**.	Positive
• shyamalan should stop trying to please his **mom**.	shyamalan should stop trying to please his **dad**.	Negative
Amazon Product Reviews		
• this book modeled my future, ... It's so amazing how one can be so enterprising and witty at such a young age! **Tom** definitely had an impact on me then and always will! I grew up wishing to be like **him**...	this book modeled my future, ... It's so amazing how one can be so enterprising and witty at such a young age! **Lisa** definitely had an impact on me then and always will! I grew up wishing to be like **her**...	Positive
• The screenplay is poor. The characters do not act like real people, especially the **Joan** Allen character. This character does not know how to file a missing persons report, does not know what ballet is, does not work, has **her daughters** cook the dinner... I could feel nothing for this character...	The screenplay is poor. The characters do not act like real people, especially the **Tom** Allen character. This character does not know how to file a missing persons report, does not know what ballet is, does not work, has **his sons** cook the dinner... I could feel nothing for this character...	Negative

We will briefly explain the process in the following, using the same notations as in [21]. Let $\mathcal{A}$ be the set of possible values of the protected attribute, in our case gender $\mathcal{A} = \{male, female\}$, and a to denote a particular value of the protected attribute (e.g., $a = male$). For each input $\mathbf{z}$ containing sensitive tokens $\phi(a)$ (such as $\phi(a) = he, his, him, husband, Paul$ for $a = male$), we generate a counterfactual input $\tilde{\mathbf{z}}$ to $\mathbf{z}$ by replacing all occurrences of each sensitive token in $\phi(a)$ with the corresponding token in $\phi(\tilde{a})$, where $\tilde{a}$ is the opposite gender, and leaving all other non-sensitive tokens of $\mathbf{z}$ unchanged.

Using the above procedure, we generated 493 counterfactual sentence pairs ($\mathcal{D}_{CDA}^{SST-2}$), for the SST-2 dataset, and 8000 sentences ($\mathcal{D}_{CDA}^{AMZR}$) for the Amazon Product Reviews dataset. Examples of counterfactual data augmentation are given in Table 1. These data sets are used for the causal mediation analysis and the de-biasing steps of the proposed approach (see Section IV-E).

C. EVALUATION DATASET

For evaluating the de-biased BERT-based sentiment analysis models, we adopted an independent testing set, the Equity Evaluation Corpus (EEC) [10], which has been specifically constructed to evaluate *gender* and *race* bias of sentiment analysis models. The dataset contains template-based sentences such as "<Person> feels angry". <Person> can be a female name such as "Monica", or a male name such as "Albert". The model M is then asked to predict the intensity of emotion - anger. The system is gender-biased when it consistently predicts higher/lower scores for sentences carrying female names than male names, or vice versa.

TABLE 2. EEC Sentence templates.

1	<Person>feels <emotional state word>.
2	The situation makes <person>feel <emotional state word>.
3	<Person>made me feel <emotional state word>.
4	I made <person>feel <emotional state word>.
5	<Person>found himself/herself in a/an <emotional situation word>situation
6	<Person>told us all about the recent <emotional situation word>events.
7	The conversation with <person>was <emotional situation word>.

The EEC corpus is a counterfactual dataset that contains seven templates of type "$<$ *Person*$>$" and "$<$ *emotional state word*$>$", as shown in Table 2. The variable "$<$ *Person*$>$" can be filled by any of 60 gender-specific names or phrases. Out of the 60, 40 are gender-specific names (20-female, 20-male), and the rest are 20 noun phrases, particularly 10 female-male pairs such as "my mother" and "my father". The variable "$<$ *emotional state word*$>$" can replace four emotions-Anger, Fear, Sadness, and Joy-each having 5 representative words. Thus, we have $1200 = 60 \times 4 \times 5$ samples for each template. In total, EEC contains $8400 = 7 \times 1200$ sentences equally divided into female- and male-specific sentences and 4-emotion categories. We refer readers to [10] for an elaborate explanation.

D. FAIRNESS EVALUATION

To measure unfairness between counterfactual pairs, we adopt the Wasserstein-1 distance as proposed by [21]. Given Y the BERT model for sentiment classification downstream task, let $Y(\mathbf{z})$ define the random variable to be the generated sentence sentiment score in [0; 1], and denote its distribution by $P_Y(\mathbf{z})$, and the same for the counterfactual input $\tilde{\mathbf{z}}$ to $\mathbf{z}$. We follow Huang et al. [21], and define the counterfactual sentiment bias, as the Wasserstein-1 distance between output sentiment distributions $P_Y(\mathbf{z})$ of the original input $\mathbf{z}$ and its counterfactual $\tilde{\mathbf{z}}$, $\mathcal{W}_1(P_Y(\mathbf{z}), P_Y(\tilde{\mathbf{z}}))$, which can be computed (see e.g. [54]) as

$$\mathcal{W}_1(P_Y(\mathbf{z}), P_S(\tilde{\mathbf{z}})) = \mathbb{E}_{\tau\sim\mathcal{U}[0,1]}|p(Y(\mathbf{z}) > \tau) - p(Y(\tilde{\mathbf{z}}) > \tau)| \quad (11)$$

where $\mathcal{U}$ is the uniform distribution. We used the SciPy Toolbox[2]([55]) to estimate $\mathcal{W}_1$ from the empirical sentiment distributions of male and female sentences.

[2]https://scipy.org

E. IMPLEMENTATION DETAILS

We implemented our proposed method using TensorFlow [56], supplemented by the utilities provided by Hugging Face's Transformers library [57]. We first fine-tuned BERT for the sentiment classification task on the SST-2 and Amazon Product Reviews datasets. For this purpose, we utilized the "AutoModelForSequenceClassification"[3] module from Transformers library [57], which facilitates instantiating the pre-trained BERT model for sequence classification tasks.

To fine-tune the BERT model for sentiment classification on the SST-2 dataset, we utilized the Adam optimizer. For setting the training hyperparameters, we conducted a grid search over learning rates (ranging from $1e-05$ to $2e-03$), batch sizes (from 16 to 64), and number of epochs (from 2 to 8). We exhaustively explored all possible combinations of such hyperparameters and kept any others as defined by default. For each hyperparameter combination, we optimized BERT using the same train-validation splits. Finally, we kept the model delivering the best performance, denoted as *BERT_SST2*, with a learning rate of $2e-05$, batch size of 32, and fine-tuned during 3 epochs. We used the same hyperparameters to fine-tune the BERT model on the Amazon Product Reviews dataset, denoted as *BERT_AMZR*.

For the de-biasing step, we used the counterfactual data augmentation sets, namely $\mathcal{D}_{CDA}^{SST-2}$ and $\mathcal{D}_{CDA}^{AMZR}$, and counterfactually trained (fine-tuned) the *BERT_SST2* and *BERT_AMZR* models using the Adam optimizer. For *BERT_SST2*, we split the $\mathcal{D}_{CDA}^{SST-2}$ dataset into 399 training sentence couples, and 94 validation sentence couples. For setting the training hyperparameters, we performed a grid search exploring the same range of hyperparameter values described in the previous step. Finally, we kept the model delivering the best performance, denoted as *DBERT_SST2* (de-biased BERT), with a learning rate of $2e-05$, batch size of 32, and trained during 3 epochs. We also used the above-defined hyperparameters to de-bias *BERT_AMZR* on the counterfactual dataset $\mathcal{D}_{CDA}^{SST-2}$, to obtain the de-biased model, *DBERT_AMZR*.

F. RESULTS

1) ILLUSTRATION OF THE GENDER BIAS

We first analyzed the gender bias of the models *BERT_SST-2* and *BERT_AMZR* and their de-biased versions (i.e., *DBERT_SST2* and *DBERT_AMZR*), using as testing data the EEC corpus. For each sentence pair, female and male sentences have the same sentence structure and differ only in gender. Thus, the two sentences should have the same sentiment score $p_{sentiment}$. The distribution of the total effect among the sentences for each model is illustrated in blue in Figure 6. The histogram shows that the predicted sentiment differs for a large proportion of the sentence pairs, confirming the presence of a gender bias in the fine-tuned models, *BERT_SST2* and *BERT_AMZR*, respectively. The bias is reduced after applying the counterfactual training (orange histograms in Figure 6).

[3]https://huggingface.co/transformers/v3.0.2/model_doc/auto.html

TABLE 3. Performance of the proposed de-biasing method on the EEC corpus. The best performances are marked in bold.

SST-2					
Model	Accuracy	Precision	Recall	F1-score	W-dist
BERT_SST2	0.77	0.76	0.85	0.75	0.0089
DBERT_SST2	**0.85**	**0.81**	**0.90**	**0.83**	**0.0022**
Amazon Product Reviews					
Model	Accuracy	Precision	Recall	F1-score	w-dist
BERT_AMZR	0.34	0.63	0.56	0.33	0.025
DBERT_AMZR	**0.52**	**0.67**	**0.69**	**0.53**	**0.020**

2) CAUSAL MEDIATION ANALYSIS

We applied the causal mediation analysis of Section III-A on *BERT_SST2* and *BERT_AMZR*. The average indirect effects of gender on sentiment prediction for each neuron are shown in Figures 7 and 8 (left panels). In the right panels of the figures, we illustrate for each layer the sum over the neurons of the average indirect effects. Both *BERT_SST2* and *BERT_AMZR* models show a gender causal effect concentrated in the last few layers. By setting the threshold $t_{SaNIE} = 0.004$, the stronger causal relationship with classification bias is observed in the last four layers of BERT for SST-2, with $3,072$ 'biased neurons' (Figure 7.right), and by setting the threshold $t_{SaNIE} = 0.042$ the last three layers for Amazon Product Reviews, with $2,304$ 'biased neurons' (Figure 8.right). The de-biasing step will consider these layers as *Target Layers*.

3) COUNTERFACTUAL DE-BIASING

In this step, we use the counterfactual sentences generated from SST-2 (resp. Amazon Product Reviews), namely, $\mathcal{D}_{CDA}^{SST-2}$ (resp. $\mathcal{D}_{CDA}^{AMZR}$) and apply the counterfactual training Algorithm 1 to fine-tune the selected *Target Layers* of *BERT_SST2* (resp. *BERT_AMZR*), and freeze the other layers, obtaining as results a de-biased model *DBERT_SST2* (resp. *DBERT_AMZR*). In Table 3, we report the results (accuracy, recall, f1-score, and Wasserstein-1 distance) on the EEC corpus; the baseline model represents the model after the first step of fine-tuning the pre-trained BERT model for sentiment classification downstream task (*BERT_SST2* and *BERT_AMZR*), before any de-biasing step is performed. The value of λ in Equation (9) has been set based on the losses in the validation sets, we report for $\lambda = 10000$ for SST-2 and $\lambda = 1000$ for Amazon Product Reviews. Overall, we observe that the proposed approach achieves reduced bias compared to the baseline model. Moreover, our de-biasing process not only does not harm the classification performance with increasing fairness but also improves it.

We further compared our approach to the de-biasing approach of [21] in which all layers are fine-tuned during the counterfactual training (de-biasing step). Tables 4 and 5 summarise the obtained results. For these experiments, we report for $\lambda \in \{1, 10, 100, 1000, 10000\}$ for SST-2

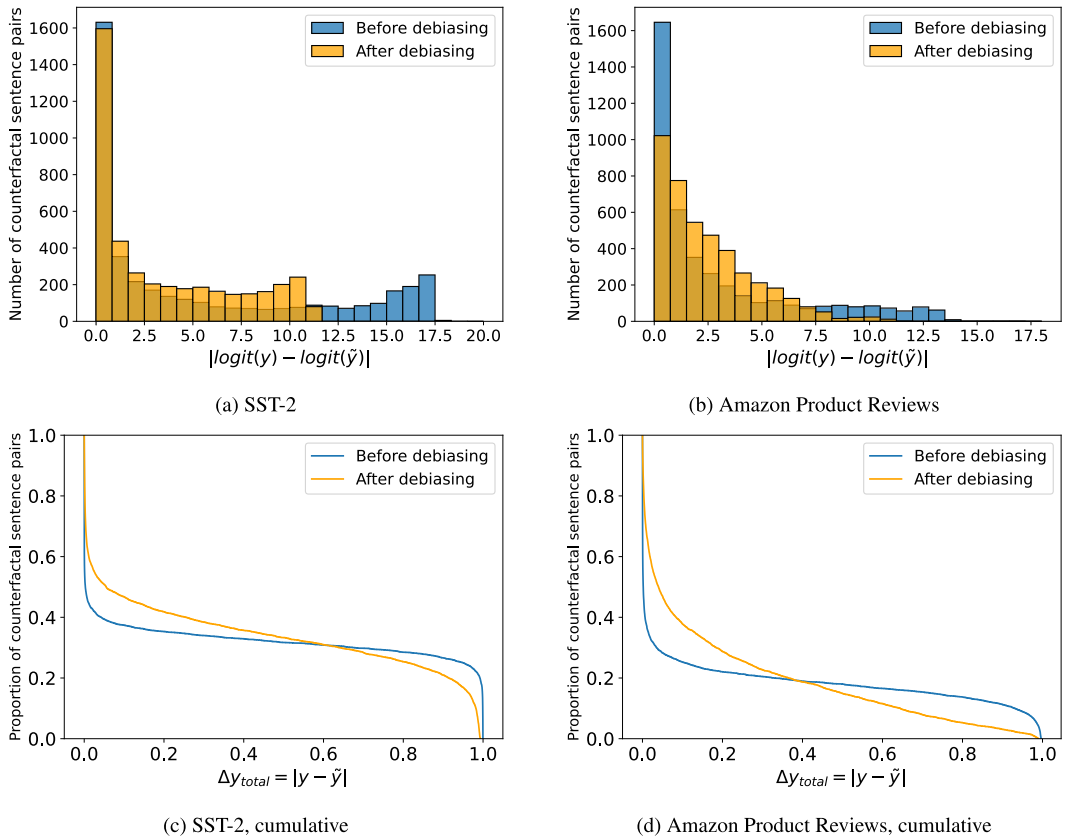

FIGURE 6. Distributions of total gender effect on predicted sentence sentiment on the EEC corpus before and after de-biasing: The top panels show the absolute difference of the predicted sentiment in logit scale; the bottom panels show the inverse cumulative distribution of the predicted sentiment absolute difference, i.e., the total effect. The sentiment prediction of many counterfactual pairs ($y = Y(\mathbf{z})$, $\tilde{y} = Y(\tilde{\mathbf{z}})$, with Y the BERT model for sentiment classification downstream task) differs between male and female attributes.

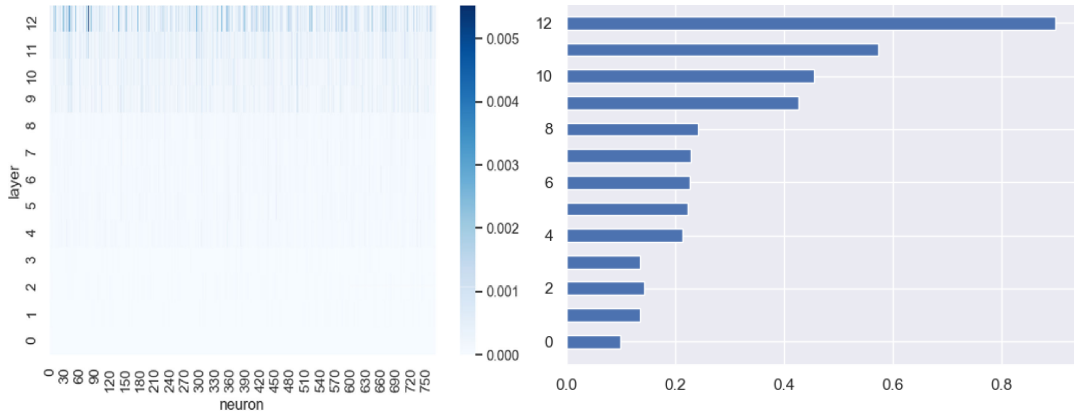

FIGURE 7. *BERT_SST2*. The average natural indirect effect of gender on sentiment prediction for each neuron (left) and sum over all the neurons of the layer (right) scaled between 0.1 and 0.9 for better visualization. Layer 0 is the embedding layer of the model, and layers 1 to 12 are BERT transformer encoder layers.

and $\lambda \in \{1, 10, 100, 1000\}$ for Amazon Product Reviews. These values have been set based on the losses in the validation sets. Overall, we observe that both approaches achieve reduced bias compared to the baseline model (see Table 3). In our experiments, we also note that with increasing λ, the bias steadily decreases, and the sentiment classification performance tends to increase slightly (see Table 4). Moreover, from both tables, it can be seen that training the entire model (i.e., approach of [21]) decreases the sentiment classification performance with increasing fairness, compared to our proposed approach, which achieves both high fairness and more accurate sentiment classification.

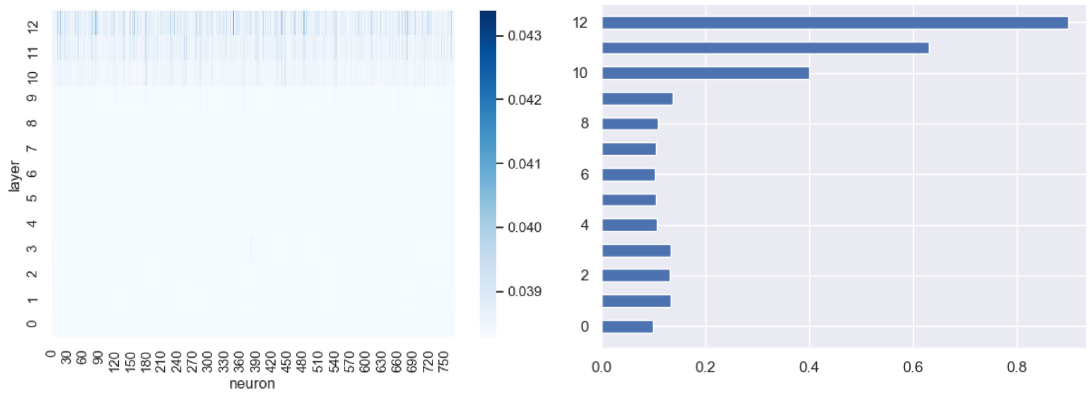

FIGURE 8. ***BERT_AMZR*. The average natural indirect effect of gender on sentiment prediction for each neuron (left) and sum over all the neurons of the layer (right) scaled between 0.1 and 0.9 for better visualization. Layer 0 is the embedding layer of the model, and layers 1 to 12 are BERT transformer encoder layers.**

TABLE 4. Classification and fairness performances of *DBERT_SST*2 on the EEC corpus. The parameter λ represents the relative weight of the classification and fairness losses. The best performances are marked in bold.

Proposed de-biasing approach					
λ	Accuracy	Precision	Recall	F1-score	W-dist
1	0.77	0.76	0.85	0.74	0.0087
10	**0.85**	**0.81**	**0.90**	0.82	0.0025
100	**0.85**	**0.81**	**0.90**	**0.83**	**0.0022**
1000	**0.85**	**0.81**	**0.90**	**0.83**	**0.0022**
10000	**0.85**	**0.81**	**0.90**	**0.83**	**0.0022**
de-biasing approach of [21]					
λ	Accuracy	Precision	Recall	F1-score	W-dist
1	0.77	0.76	0.85	0.75	0.0032
10	**0.81**	**0.78**	**0.87**	**0.79**	0.0035
100	0.79	0.77	0.86	0.77	0.0029
1000	0.77	0.76	0.85	0.75	0.0020
10000	0.24	0.62	0.50	0.20	**0.0012**

TABLE 5. Classification and fairness performances of *DBERT_AZR* on the EEC corpus. The parameter λ represents the relative weight of the classification and fairness losses. The best performances are marked in bold.

Proposed de-biasing approach					
Model	Accuracy	Precision	Recall	F1-score	W-dist
1	0.34	0.63	0.57	0.33	0.021
10	**0.54**	0.67	0.70	**0.54**	0.022
100	0.47	0.66	0.65	0.47	0.024
1000	0.52	0.67	**0.69**	0.53	**0.020**
de-biasing approach of [21]					
Model	Accuracy	Precision	Recall	F1-score	W-dist
1	0.36	0.63	0.56	0.32	0.024
10	**0.43**	**0.65**	**0.62**	**0.43**	0.013
100	0.38	0.64	0.59	0.37	0.019
1000	0.31	0.63	0.55	0.29	**0.0075**

We hypothesize that the increase in classification performance is due to the fine-tuning of only the last layers of the BERT model.

Finally, we evaluated the fairness loss proposed by [43], i.e., Equation(1). Results for the experiments, in terms of accuracy, recall, f1-score, and Wasserstein-1 distance, are

TABLE 6. Performance of the proposed de-biasing method on the EEC corpus using the fairness loss proposed in [43] (see Equation(1)). The parameter λ represents the relative weight of the classification and fairness losses.

*DBERT_SST*2					
λ	Accuracy	Precision	Recall	F1-score	W-dist
0.01	0.70	0.72	0.80	0.69	0.020
0.1	0.71	0.72	0.80	0.68	0.022
0.5	0.74	0.74	0.83	0.73	0.021
0.8	0.71	0.73	0.81	0.69	0.024
1	0.72	0.73	0.82	0.70	0.023
DBERT_AZR					
λ	Accuracy	Precision	Recall	F1-score	w-dist
0.01	0.38	0.64	0.59	0.38	0.027
0.1	0.33	0.63	0.56	0.32	0.014
0.5	0.72	0.46	0.49	0.45	0.014
0.8	0.76	0.38	0.50	0.43	0.014
1	0.76	0.38	0.50	0.43	0.013

listed in Table 6. Comparing these results to the ones provided in Tables 4, and 5, it seems that the proposed cosine fairness loss, i.e., Equation 10, mitigates better gender bias than the fairness loss proposed by [43]. Finally, we note that combining CDA and our fairness loss function outperforms the method of [43] in the Wasserstein-1 distance measure of biases without compromising sentiment classification accuracy.

V. DISCUSSION

We demonstrated the effectiveness of the proposed targeted counterfactual training for de-biasing language models.

The proposed method combines three prominent techniques: (a) Counterfactual Data Augmentation, (b) a method based on causal mediation analysis to study the inner workings of neural networks [28], and a counterfactual training for model de-biasing [21]. We observed the significant advantages of the causal mediation analysis, as it identifies the contributions made by individual neurons within a neural network through causal interventions. We calculate the average indirect effects of gender on sentiment prediction

across layers and present graphical illustrations to identify the *Target Layers* responsible for the bias, and (c) selectively fine-tuning these layers via counterfactual training while keeping the remaining layers frozen, our approach effectively mitigates biases without impairing prediction performance. Therefore, a cascade of these approaches can be used to optimally de-bias language models for sentiment analysis. In contrast to existing de-biasing methods, such as [21], our approach successfully achieves both fairness and high performance in sentiment analysis. These results, in our opinion, indicate that the first fine-tuned BERT model for sentiment analysis (Step 1 of the proposed procedure) learns the core features that are relevant to the sentiment classification task, and only needs to further compose the representation by promoting on balanced datasets during counterfactual fine-tuning of the last layers to recover those core features that enabled better performance. Since the fairness criterion is imposed only at the counterfactual fine-tuning phase, we will not suffer from over-fitting issues. This has also been confirmed by Kirichenko and his colleagues [58], who stated that "retraining the last layer of suitably pre-trained representations can reduce vulnerability to spurious correlation, and thus significantly improve prediction accuracy on imbalanced dataset and model robustness to covariate shift."

VI. CONCLUSION

The primary objective of this work was to investigate the application of counterfactual analysis to study the bias and fairness of large-pretrained models. We used a counterfactual intervention method to analyze a sentiment predictor based on the BERT model, uncovering the distribution of gender bias across the different layers of the network. Subsequently, we successfully mitigate biases within the model by utilizing our proposed counterfactual training approach, resulting in a noticeable improvement in gender fairness.

Although the presented approach has been illustrated on gender bias reduction in BERT fine-tuned for sentiment classification downstream task, it can be applied to other biases on sensitive attributes such as ethnic groups, skin color, and country, as well as for other NLP deep models.

Our experiments present some limitations. We restricted our analysis to the SST-2 and Amazon product review datasets. Additionally, we solely employed the p-Wasserstein metric to evaluate fairness. Although these datasets and fairness evaluation methods are widely utilized, they represent a limited scope of cases. Further experiments should be conducted to validate the efficacy of our proposed method more comprehensively, encompassing a broader range of datasets and fairness evaluation techniques, as well as quantifying accuracy-fairness trade-off in real-world datasets [59], [60].

REFERENCES

[1] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowl.-Based Syst.*, vol. 226, Aug. 2021, Art. no. 107134.

[2] S. Kausar, X. Huahu, W. Ahmad, M. Y. Shabir, and W. Ahmad, "A sentiment polarity categorization technique for online product reviews," *IEEE Access*, vol. 8, pp. 3594–3605, 2020.

[3] L. Tian, C. Lai, and J. Moore, "Polarity and intensity: The two aspects of sentiment analysis," in *Proc. Grand Challenge Workshop Hum. Multimodal Lang. (Challenge-HML)*, A. Zadeh, P. P. Liang, L.-P. Morency, S. Poria, E. Cambria, and S. Scherer, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2018, pp. 40–47. [Online]. Available: https://aclanthology.org/W18-3306

[4] P. Nandwani and R. Verma, "A review on sentiment analysis and emotion detection from text," *Social Netw. Anal. Mining*, vol. 11, no. 1, p. 81, Dec. 2021, doi: 10.1007/s13278-021-00776-6.

[5] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," *Artif. Intell. Rev.*, vol. 55, no. 7, pp. 5731–5780, Oct. 2022.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–35, Jul. 2022.

[7] K. Durrheim, M. Schuld, M. Mafunda, and S. Mazibuko, "Using word embeddings to investigate cultural biases," *Brit. J. Social Psychol.*, vol. 62, no. 1, pp. 617–629, Jan. 2023.

[8] E. L. Ungless, B. Ross, and V. Belle, "Potential pitfalls with automatic sentiment analysis: The example of queerphobic bias," *Social Sci. Comput. Rev.*, vol. 41, no. 6, pp. 2211–2229, Dec. 2023, doi: 10.1177/08944393231152946.

[9] M. Diaz, I. Johnson, A. Lazar, A. M. Piper, and D. Gergle, "Addressing age-related bias in sentiment analysis," in *Proc. CHI Conf. Human Factors Comput. Syst.* New York, NY, USA: Association for Computing Machinery, Apr. 2018, pp. 1–14, doi: 10.1145/3173574.3173986.

[10] S. Kiritchenko and S. Mohammad, "Examining gender and race bias in two hundred sentiment analysis systems," in *Proc. 7th Joint Conf. Lexical Comput. Semantics*. New Orleans, LA, USA: Association for Computational Linguistics, Jun. 2018, pp. 43–53. [Online]. Available: https://aclanthology.org/S18-2005

[11] R. Mihalcea and A. Garimella, "What men say, what women hear: Finding gender-specific meaning shades," *IEEE Intell. Syst.*, vol. 31, no. 4, pp. 62–67, Jul. 2016.

[12] C. S. Montero, M. Munezero, and T. Kakkonen, "Investigating the role of emotion-based features in author gender classification of text," in *Computational Linguistics and Intelligent Text Processing*. Berlin, Germany: Springer, 2014, pp. 98–114.

[13] M. Guerini, L. Gatti, and M. Turchi, "Sentiment analysis: How to derive prior polarities from SentiWordNet," in *Proc. Conf. Empirical Methods Natural Lang.* Stroudsburg, PA, USA: Association for Computational Linguistics, 2013, pp. 1259–1269. [Online]. Available: https://aclanthology.org/D13-1125

[14] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness GAN: Generating datasets with fairness properties using a generative adversarial network," *IBM J. Res. Develop.*, vol. 63, nos. 4–5, pp. 3:1–3:9, Jul. 2019.

[15] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proc. AAAI/ACM Conf. AI, Ethics, Soc.* Stroudsburg, PA, USA: Association for Computing Machinery, 2018, pp. 335–340.

[16] P. K. Lohia, K. N. Ramamurthy, M. Bhide, D. Saha, K. R. Varshney, and R. Puri, "Bias mitigation post-processing for individual and group fairness," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 2847–2851.

[17] L. S. Khoo, J. Q. Bay, M. L. K. Yap, M. K. Lim, C. Y. Chong, Z. Yang, and D. Lo, "Exploring and repairing gender fairness violations in word embedding-based sentiment analysis model through adversarial patches," in *Proc. IEEE Int. Conf. Softw. Anal., Evol. Reengineering (SANER)*, Mar. 2023, pp. 651–662.

[18] M. H. Asyrofi, Z. Yang, I. N. B. Yusuf, H. J. Kang, F. Thung, and D. Lo, "BiasFinder: Metamorphic test generation to uncover bias for sentiment analysis systems," *IEEE Trans. Softw. Eng.*, vol. 48, no. 12, pp. 5087–5101, Dec. 2022.

[19] A. A. Almuzaini and V. K. Singh, "Balancing fairness and accuracy in sentiment detection using multiple black box models," in *Proc. 2nd Int. Workshop Fairness, Accountability, Transparency Ethics Multimedia*, Oct. 2020, pp. 13–19.

[20] M. B. Zafar, I. Valera, M. Gomez-Rodriguez, and K. P. Gummadi, "Fairness constraints: A flexible approach for fair classification," *J. Mach. Learn. Res.*, vol. 20, no. 75, pp. 1–42, 2019. [Online]. Available: http://jmlr.org/papers/v20/18-262.html

[21] P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, and P. Kohli, "Reducing sentiment bias in language models via counterfactual evaluation," in *Proc. Findings Assoc. Comput. Linguistics, EMNLP*. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 65–83. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.7

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1. Minneapolis, MN, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[23] M. Sahlgren and F. Olsson, "Gender bias in pretrained Swedish embeddings," in *Proc. 22nd Nordic Conf. Comput. Linguistics*, 2019, pp. 35–43.

[24] S. Jentzsch and C. Turan, "Gender bias in BERT—Measuring and analysing biases through sentiment rating in a realistic downstream classification task," in *Proc. 4th Workshop Gender Bias Natural Lang. Process. (GeBNLP)*, 2022, pp. 184–199. [Online]. Available: https://aclanthology.org/2022.gebnlp-1.20

[25] T. Leteno, A. Gourru, C. Laclau, and C. Gravier, "An investigation of structures responsible for gender bias in bert and distilbert," in *Advances in Intelligent Data Analysis XXI*, B. Crémilleux, S. Hess, and S. Nijssen, Eds. Cham, Switzerland: Springer, 2023, pp. 249–261.

[26] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Comput. Surv.*, pp. 1–37, Aug. 2023, doi: 10.1145/3616865.

[27] Q. Zhao and T. Hastie, "Causal interpretations of black-box models," *J. Bus. Econ. Statist.*, vol. 39, no. 1, pp. 272–281, Jan. 2021.

[28] J. Vig, S. Gehrmann, Y. Belinkov, S. Qian, D. Nevo, S. Sakenis, J. Huang, Y. Singer, and S. Shieber, "Causal mediation analysis for interpreting neural NLP: The case of gender bias," 2020, *arXiv:2004.12265*.

[29] J. Pearl, *Causality*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[30] J. Pearl, "Direct and indirect effects," in *Proc. 17th Conf. Uncertainty Artif. Intell.* San Francisco, CA, USA: Morgan Kaufmann, 2001, pp. 411–420.

[31] S. Verma and J. Rubin, "Fairness definitions explained," in *Proc. IEEE/ACM Int. Workshop Softw. Fairness (FairWare)*, May 2018, pp. 1–7, doi: 10.1145/3194770.3194776.

[32] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Proc. 30th Int. Conf. Mach. Learn.*, in Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., 2013, vol. 28, no. 3, pp. 325–333. [Online]. Available: https://proceedings.mlr.press/v28/zemel13.html

[33] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proc. 3rd Innov. Theor. Comput. Sci. Conf.* New York, NY, USA: Association for Computing Machinery, Jan. 2012, pp. 214–226, doi: 10.1145/2090236.2090255.

[34] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 3323–3331.

[35] W. Fleisher, "What's fair about individual fairness?" in *Proc. AAAI/ACM Conf. AI, Ethics, Soc. (AIES)*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 480–490, doi: 10.1145/3461702.3462621.

[36] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Red Hook, NY, USA: Curran Associates, 2017.

[37] N. Meade, E. Poole-Dayan, and S. Reddy, "An empirical survey of the effectiveness of debiasing techniques for pre-trained language models," 2021, *arXiv:2110.08527*.

[38] K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, E. H. Chi, and S. Petrov, "Measuring and reducing gendered correlations in pre-trained models," pp. 1–12, 2021, *arXiv:2010.06032*. [Online]. Available: https://arxiv.org/abs/2010.06032

[39] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," in *Proc. 8th Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–17. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[40] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta, *Gender Bias in Neural Natural Language Processing*. Cham, Switzerland: Springer, 2020, pp. 189–202.

[41] P. P. Liang, I. Z. Li, E. Zheng, Y. C. Lim, R. Salakhutdinov, and L.-P. Morency, "Towards debiasing sentence representations," in *Proc. Annu. Meeting ACL*, 2020, pp. 1–14. [Online]. Available: https://api.semanticscholar.org/CorpusID:207996257

[42] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 1–11.

[43] Y. Qian, U. Muaz, B. Zhang, and J. W. Hyun, "Reducing gender bias in word-level language models with a gender-equalizing loss function," in *Proc. 57th Annu. Meeting ACL*, F. Alva-Manchego, E. Choi, and D. Khashabi, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 223–228. [Online]. Available: https://aclanthology.org/P19-2031

[44] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[45] S. Barikeri, A. Lauscher, I. Vulić, and G. Glavaš, "RedditBias: A real-world resource for bias evaluation and debiasing of conversational language models," in *Proc. 59th Annu. Meeting ACL*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 1941–1955. [Online]. Available: https://aclanthology.org/2021.acl-long.151

[46] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proc. 58th Annu. Meeting ACL, Syst. Demonstrations*, A. Celikyilmaz and T.-H. Wen, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 270–278. [Online]. Available: https://aclanthology.org/2020.acl-demos.30

[47] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," 2019, *arXiv:1901.02860*.

[48] W. E. Chi, S. Huang, M. Jeon, E. S. Park, T. Melguizo, and A. Kezar, "A practical guide to causal mediation analysis: Illustration with a comprehensive college transition program and non-program peer and faculty interactions," *Frontiers Educ.*, vol. 7, Aug. 2022, Art. no. 886722. [Online]. Available: https://www.frontiersin.org/articles/10.3389/feduc.2022.886722

[49] A. Feder, N. Oved, U. Shalit, and R. Reichart, "CausaLM: Causal model explanation through counterfactual language models," *Comput. Linguistics*, vol. 47, no. 2, pp. 333–386, Jul. 2021. [Online]. Available: https://doi.org/10.1162/coli_a_00404

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Red Hook, NY, USA: Curran Associates, 2017.

[51] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, and C. Potts, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013, pp. 1631–1642.

[52] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 507–517.

[53] A. Mishra, H. Mishra, and S. Rathee, "Examining the presence of gender bias in customer reviews using word embedding," 2019, *arXiv:15426.02240*. [Online]. Available: http://rgdoi.net/10.13140/RG.2.2.15426.02240

[54] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa, "Wasserstein fair classification," in *Proc. 35th Uncertainty Artif. Intell. Conf.*, in Proceedings Machine Learning Research, vol. 115, R. P. Adams and V. Gogate, Eds., Jul. 2020, pp. 862–872. [Online]. Available: https://proceedings.mlr.press/v115/jiang20a.html

[55] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in Python," *Nature Methods*, vol. 17, no. 3, pp. 261–272, 2020.

[56] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Conf. Oper. Syst. Design Implement. (OSDI)*. Berkeley, CA, USA: USENIX Association, 2016, pp. 265–283.

[57] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Process., Syst. Demonstrations*. Stroudsburg, PA, USA: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[58] P. Kirichenko, P. Izmailov, and A. G. Wilson, "Last layer re-training is sufficient for robustness to spurious correlations," 2022, *arXiv:2204.02937*.
[59] S. Liu and L. N. Vicente, "Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach," *Comput. Manage. Sci.*, vol. 19, pp. 513–537, Jul. 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:220962256
[60] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. R. Varshney, "Is there a trade-off between fairness and accuracy? A perspective using mismatched hypothesis testing," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, 2020, pp. 2803–2813.

YIFEI DA received the B.Sc.Eng. degree in information and computational sciences and the M.Sc.Eng. degree in computer science and technology from Beihang University (BUAA), Beijing, China, in 2015 and 2018, respectively. She is currently pursuing the Ph.D. degree with the Electronics and Informatics (ETRO) Department, Vrije Universiteit Brussel, Brussels, Belgium, under the supervision of Prof. H. Sahli. Her current research interests include sentiment analysis, causality, information theory in financial tasks, and machine learning.

MATÍAS NICOLÁS BOSSA received the M.Sc. degree in physics from the Balseiro Institute, National University of Cuyo, Bariloche, Argentina, in 2002, and the Ph.D. degree in biomedical engineering from the University of Zaragoza, Spain, in 2011. He is a Postdoctoral Researcher with the Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel, Brussels, Belgium. His current research interests include statistical modeling of biomedical data, medical image analysis, and clinical prediction modeling.

ABEL DÍAZ BERENGUER received the B.Sc.Eng. degree in informatics sciences and the M.Sc.Eng. degree in applied informatics from the University of Informatics Sciences (UCI), Havana, Cuba, in 2009 and 2014, respectively, and the Ph.D. degree in engineering from Vrije Universiteit Brussel (VUB), Brussels, Belgium, in 2021. He is a Postdoctoral Researcher with the Department of Electronics and Informatics (ETRO), VUB. His current research interests include computer vision, representation learning, and machine learning, with an emphasis on semi-supervised learning in applications, such as medical imaging, personalized medicine (treatment selection, treatment effects estimation, and risk prediction), automatic human social behavior analysis, and smart video surveillance.

HICHEM SAHLI is currently a Professor of computer vision and machine learning with the Department of Electronics and Informatics (ETRO), Vrije Universiteit Brussel (VUB), Brussels, Belgium, and a Principal Scientist with the Interuniversity Microelectronics Centre (IMEC), Leuven, Belgium. He has completed 28 Ph.D. supervisions, since 2000, when he joined VUB. He has authored over 300 journals, conference publications, and book chapters. His research focuses on signal and image processing, computer vision and machine learning theory and algorithms in computer vision (radar, image, and video processing), affective computing, health informatics (medical image analysis, disease progression prediction, diagnosis and treatment outcome prediction, and disease outbreak forecasting), and natural language processing (electronic health record coding).