ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240576

Explaining Text Classifiers with Counterfactual Representations

Pirmin Lemberger [a,*] **and Antoine Saillenfest** [a,**]

[a]onepoint, 29 rue des Sablons, 75116 Paris (France)

Abstract. One well motivated explanation method for classifiers leverages counterfactuals which are hypothetical events identical to real observations in all aspects except for one feature. Constructing such counterfactual poses specific challenges for texts, however, as some attribute values may not necessarily align with plausible real-world events. In this paper we propose a simple method for generating counterfactuals by intervening in the space of text representations which bypasses this limitation. We argue that our interventions are minimally disruptive and that they are theoretically sound as they align with counterfactuals as defined in Pearl's causal inference framework. To validate our method, we conducted experiments first on a synthetic dataset and then on a realistic dataset of counterfactuals. This allows for a direct comparison between classifier predictions based on ground truth counterfactuals—obtained through explicit text interventions—and our counterfactuals, derived through interventions in the representation space. Eventually, we study a real world scenario where our counterfactuals can be leveraged both for explaining a classifier and for bias mitigation.

1 Introduction

Providing an explanation for the predictions made by a text classifier for a particular document is essential in situations where social bias could have detrimental consequences, for example when documents refer to individuals belonging to different social groups. One well motivated explanation method for classifiers leverages counterfactuals which are hypothetical individuals identical to real ones in all aspects except for one feature that is being intervened on [18]. Understanding how a classifier reacts to such fictitious individuals will indeed furnish an explanation for how it uses different pieces of information for its predictions. This approach to explanations has been well investigated in the context of social fairness [15, 12] but it obviously has a wider scope.

Creating counterfactuals (CF) for text documents poses specific challenges when compared to producing CF for tabular data. One of these is related to the fact that it is often by no means obvious how to define a CF text for which the value of some text attribute is modified while everything else is kept unchanged. A number of recent works do, however, propose methods for constructing explicit counterfactuals at the text level in restricted contexts. Zeng et al. [27] for instance propose to intervene on entities, separated from their context, to provide CF text samples that can be used for improving the generalization of a NER classifier under limited observational samples. Calderon et al. [6] also intervene on the text by replacing some domain-specific terms to create coherent counterfactuals also used for data augmentation purposes. Madaan et al. [17] perform controlled text generation to enforce some user provided label.

In this paper, leveraging insights from recent research on concept erasure [10, 4, 24], we propose a simple method for producing counterfactual representations (CFR) defined as interventions on text representations produced by a generic neural encoder like BERT. A CFR thus implements the alteration of the value of a single protected text attribute. Although the corresponding information is spread over all components of a high-dimensional representation we ensure that our interventions are minimally disruptive in a precise sense. More importantly, CFRs can be instantiated even in cases where a corresponding intervention on the text proves impossible. Finally, these CFRs moreover turn out to be easy to compute.

Use cases of counterfactuals representations

Our method for creating CFRs for texts can be applied to various use cases. It will however prove especially valuable in scenarios where direct interventions on the texts would either lack meaning or incur excessive costs, whether it is human labour or using a generative AI service. Explaining why some texts have been classified in an unexpected category is one important use case. Our method can indeed isolate the role of specific values of a concept in a classifier's prediction whereas traditional erasure methods only provide an evaluation of the global impact of a concept, typically as an average treatment effect.

The value-by-value analysis we propose will be of particular interest when the fairness of a classifier is at stake because it will reveal precisely which demographic groups are discriminated against. Beyond this explanation use case our CFRs can also serve for counterfactual data augmentation which consists in adding CF to an existing train set. This task is generally performed for OOD generalizability improvement [13, 14] or model robustness and fairness [12].

Contributions

Our contributions are as follows:

1. We propose a simple method for generating textual counterfactual representations which corresponds to replacing one concept value with another (sections 3.1 and 3.2).
2. Beyond intuitive arguments, we show that our model aligns with the definition of counterfactuals in Pearl's causal inference framework and is thus theoretically sound (section 3.3).

* Corresponding Author. Email: p.lemberger@groupeonepoint.com.
** Corresponding Author. Email: a.saillenfest@groupeonepoint.com

3. We introduce the EEEC+ synthetic dataset, which enables the generation of genuine counterfactuals by performing explicit interventions on texts (section 4.1) that will serve as a ground truth when we compare the response of a classifier to these with the response to our counterfactuals (sections 5.1 and 5.2).
4. We exhibit practical use cases for our counterfactuals in realistic contexts where counterfactuals are generally not available. We use them to evaluate the causal effects in a sentiment prediction task (section 5.3) and to explain a classifier's prediction (section 5.4).

The aim of this work is not to achieve any kind of SOTA performance. Instead, we aim to demonstrate the usefulness of a simple and theoretically sound regression-based approach to generate counterfactual representations, which can serve as a strong baseline for a variety of tasks. Code and data are available on github[1] and a version of this paper including the Supplementary Material on arxiv [16].

2 Related Work

Using Pearl's causal inference framework for defining counterfactual fairness was pioneered in Kusner et al. [15]. This work motivates and formalizes the intuition that a classifier which is fair towards individuals belonging to different social groups should produce the same predictions for an actual individual and for a counterfactual individual belonging to a different group, other things being equal. It also stresses the importance of taking into account causal relationships between the variables that describe an individual when constructing fair classifiers. These relations are typically expressed with a DAG associated to the structural causal model (SCM) [19, 20] which describes the data generation process. One central observation is that fair predictors should only rely on variables that are non-descendant of protected variables in the causal graph. Finally, the authors describe an algorithm for training fair classifiers that uses a deconvolution approach. Our method for producing CFR for text representations could be used as a practical way of identifying which values of a sensitive text attribute imply a violation of the counterfactual fairness of a possibly biased classifier.

A slightly stronger notion of counterfactual invariance (CFI) is introduced in Veitch et al. [25] in order to formalize what it means for a classifier $\widehat{Y}$ to successfully pass stress tests which involve intervening on a protected attribute Z. Intuitively, a CFI classifier $\widehat{Y}$ does not rely on that part of the information in X that can be causally affected by the value of Z. The main result of this work is that, depending on the underlying causal structure of the data generating process, a CFI predictor $\widehat{Y}$ obeys different independence relations that form a testable signature of the desired invariance. Our method for creating CFR also relies on a part $X^\perp$ that is unrelated to a protected attribute Z, although in a weaker sense. But, unlike Veitch et al. [25] and following Shao et al. [24], we exhibit $X^\perp$ explicitly under some linearity assumptions on how the sensitive information Z is hidden in $X^\perp$.

Another line of work [21, 26, 10, 4, 11], focuses on defining methods for concept erasure. The aim is still to build fair predictors that use data from which information on protected attributes has been "scrubbed". However, achieving such erasure can be tricky due to the presence in the text of numerous factors correlated with the concept to be erased [7]. To circumvent this difficulty, one possibility is to intervene on text representations rather than on texts themselves [2]. Intervention methods on representations generally fall into two categories: adversarial methods and linear methods. The former rely on a gradient-reversal layer during training to produce representations that do not encode information about the protected attribute [11], but have been proven to fail at fully removing this information [9]. Focusing on linear methods, [21, 10, 4, 24] use projections that remove unwanted information from the representation space. In our work we also opt to intervene on representations using a closed form projector acting on the representation space as in Belrose et al. [4]. Two aspects of this approach are worth mentioning. First, there is no need to train a machine learning model, and therefore it requires minimal computational resources. Second, except for the erased information, this method preserves as much information as possible in a precise sense.

Finally, several papers propose counterfactual benchmarks. In De-Arteaga et al. [7], approximate CFs are generated to assess the impact of gender information on the occupation classification in HR systems. Abraham et al. [1] proposes a benchmark for explanation methods consisting of a large set of interventions on short restaurant reviews. Feder et al. [11] evaluates the causal effect of a concept on a classification task using synthetic counterfactuals. In most cases the interventions are defined for binary attributes only. As our method can be applied beyond scenarios with binary attributes, we introduce in this study a counterfactual benchmark dataset with a non-binary attribute.

3 Creating Counterfactual Representations

3.1 Background

Our aim is to define CFRs that can be used as reliable substitutes for genuine CF. Let us thus start by enumerating what we intuitively expect from a "good" CFR. For the sake of clarity, let's assume that a sentence s describes the emotional state of an individual and that it is represented by an embedding $X(s) \in \mathbb{R}^d$, obtained from a standard encoder like BERT. Suppose that $\widehat{Y}$ is a classifier for some discrete Y like the emotional content conveyed by s, and that $Z(s)$ is a discrete protected attribute like the gender or the race of the individual referred to in s. We thus assume the causal graph is $Z \to X \to \widehat{Y}$.[2]

Now suppose that starting from a text s conveying a race $Z(s)$ we can explicitly exhibit a CF text $s_{Z\leftarrow z}$ referring to a hypothetical individual whose race is $Z(s_{Z\leftarrow z}) = z \neq Z(s)$, all other things being equal. Let $X(s_{Z\leftarrow z})$ be the representation of this CF sentence and let $X(s)_{Z\leftarrow z}$ be a tentative CFR obtained by intervening directly on $X(s)$. From a "good" CFR $X_{Z\leftarrow z}$ we expect that:

- it should fool any classifier $\widehat{Y}$ most of the time, namely we expect that $P[\widehat{Y}(X(s_{Z\leftarrow z})) \neq \widehat{Y}(X(s)_{Z\leftarrow z})]$ is small in a sense to be made precise,
- similarly, it should fool any classifier $\widehat{Z}$ most of the time,
- the CFR $X_{Z\leftarrow z}$ should preserve as much information in X as possible, except for that part on which we intervene to change the value of Z,
- finally, calculating $X_{Z\leftarrow z}$ from X should be computationally inexpensive.

3.2 Making Minimal Interventions

To define a CFR as a minimally disruptive intervention we follow Ravfogel et al. [23] which introduced the concept of linear guardedness that we now briefly review. It formalizes the intuition that only

[1] github.com/ToineSayan/counterfactual-representations-for-explanation

[2] The causal relationships with Y do not concern us because we focus on explaining predictions $\widehat{Y}$.

part of the information in X is useful for predicting Z with a linear predictor. Let thus $\eta(X)$ be a predictor for Z and assume that the loss function $\ell(\eta, Z)$ is convex in its first argument, which is the case for the usual cross-entropy for instance. A representation $X^\perp$ is then said to linearly guard Z (as a one-hot encoded variable for k categories in $\{0,1\}^k$) if no linear predictor $\eta(X^\perp) = \mathbf{W}X^\perp + \mathbf{b}$ is able to predict Z better than a constant predictor $\eta(X^\perp) \equiv \mathbf{b}$. More formally, $X^\perp$ as a function of the text s should maximize the minimum expected loss over linear predictors:

$$X^\perp \in \max_X \min_{\mathbf{W},\mathbf{b}} \mathbb{E}[\ell(\eta(X), Z)]. \tag{1}$$

The linearity assumption is a strong one[3] but this is the price to pay for having the useful equivalence in [4] that $X^\perp$ linearly guards Z iff

$$\mathbf{\Sigma}_{X^\perp Z} := \mathrm{Cov}[X^\perp, Z] = 0. \tag{2}$$

Moreover [4] show that such a protected $X^\perp$ can be obtained from an arbitrary representation X by a simple projection

$$X^\perp = \mathbf{P}X \tag{3}$$

provided $\mathbf{P}$ satisfies $\ker(\mathbf{P}) \supseteq \mathrm{im}(\mathbf{\Sigma}_{XZ})$. In words, the projector $\mathbf{P}$ should nullify the column space of the covariance matrix $\mathbf{\Sigma}_{XZ} := \mathrm{Cov}[X, Z]$. In general $\mathbf{P} \neq \mathbf{P}^\top$ and thus the projection is oblique and is not unique.[4] We simply use the orthogonal projector on $\mathrm{im}(\mathbf{\Sigma}_{XZ})^\perp$. Denoting $V^\parallel = \mathrm{im}(\mathbf{\Sigma}_{XZ})$, which has dimension $k-1$ (as Z is one-hot), and $V^\perp$ its orthogonal complement in $\mathbb{R}^d$, we decompose the representation space $\mathbb{R}^d$ as $V^\perp \oplus V^\parallel$.

We typically have $k \ll d$ which means that $\mathbf{P}$ erases only a tiny fraction of the information in X, namely that information which could be used to predict Z using a linear predictor. Figure 1 illustrates the geometric situation when Z can take $k = 2$ values. In particular the component $x^\parallel$ contains information allowing to predict Z (linearly) whereas the component $x^\perp$ does not.

To define our CFR $x_{Z\leftarrow z}$ for an initial representation x we first set $x^\perp_{Z\leftarrow z} := x^\perp$, thus keeping the component without linear information on Z unchanged. Next, we define $x^\parallel_{Z\leftarrow z}$ by linearly regressing $x^\parallel$ on $x^\perp$, on the subset of texts for which $Z(s) = z$. The CFR $x_{Z\leftarrow z}$ will thus be close (in quadratic mean) to representations of real sentences s having $Z(s) = z$ as illustrated in Figure 1. More precisely, we define $x^\parallel_{Z\leftarrow z}$ using one multivalued least square regression for each possible target value z. Summarizing, our CFR's are thus defined by

$$x_{Z\leftarrow z}(x) := \begin{bmatrix} x^\parallel_{Z\leftarrow z}(x) \\ x^\perp_{Z\leftarrow z}(x) \end{bmatrix} := \begin{bmatrix} \mathbf{W}(z)\, x^\perp(x) + \mathbf{b}(z) \\ x^\perp(x) \end{bmatrix}. \tag{4}$$

3.3 Relation with Pearl's Framework

In this section we will argue that the CFR $x_{Z\leftarrow z}$ defined above fits naturally into Pearl's causal inference framework. For this we will exhibit an appropriate structural causal model (SCM) the definition of which we now briefly recall to fix the notations, referring to [19, 20] for more details.

[3] Non-linear concept erasure is still largely an open problem. Three lines of work have tackled it recently. Adversarial approaches [11], kernelized versions of linear erasure methods [22] and more recently approaches leveraging rate distortion theory [3].

[4] In Belrose et al. [4] this freedom is used to make $X^\perp = \mathbf{P}X$ as close as possible to the original X in quadratic mean.

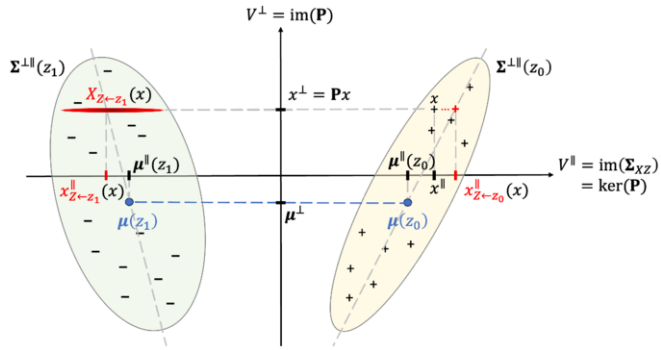

Figure 1: The representation space when Z takes $k = 2$ values. Representations of texts for which $Z(s) = z_0$ are shown as $+$ and those for which $Z(s) = z_1$ as $-$, they form two clusters. The representation x is associated with a text for which $Z = z_0$. Once projected by $\mathbf{P}$ on $V^\perp$ we obtain a representation $x^\perp$ from which it is impossible to recover the value z of the protected attribute Z using a linear predictor. This information is contained in $x^\parallel$. Our CFRs $x_{Z\leftarrow z_0}$ and $x_{Z\leftarrow z_1}$ for x corresponding to setting $Z = z_0$ or z_1 are obtained by regressing $x^\parallel$ on $x^\perp$ on observations for which $Z = z_0$ and z_1 respectively (oblique dashed lines). The random variable $X_{Z\leftarrow z_1}(x)$ is the $Z = z_1$ non deterministic Pearl counterfactual for x. Its expectation value corresponds to our CFR $x_{Z\leftarrow z_1}(x)$. The remaining notations are defined in equations (5), (6) and (7).

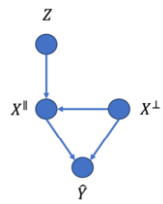

Figure 2: The DAG G which corresponds to the SCM generating model of text documents.

A SCM specifies the causal mechanism that generates data. It is defined by a DAG G where each node $i \in G$ is associated to an observed variable O_i and a noise variable U_i. A set $\{f_1, \ldots, f_{|G|}\}$ of functions specifies how each variable O_i depends on its parent variables PA_i in G and on the noise variable U_i, namely $O_i = f_i(\mathrm{PA}_i\,; U_i)$. The set of observed and noise variables are denoted by O and U respectively. At last, let $P(U)$ denote the joint probability distribution of the noise variables U, which are assumed independent and which are the only source of randomness for the observed variables O. Let $P(O)$ denote the induced probability distribution. An intervention on a variable $Z = O_i \in O$ which sets its value to z is defined by replacing $f_i(\mathrm{PA}_i\,; U_i)$ by the constant function $f_i \equiv z$. If $X \in O$ is another variable of interest, such an intervention induces a modified distribution on X which we denote by $P(X_{Z\leftarrow z})$. If we observe that some variables $E \subset O$ have values e this induces a conditional distribution $P(U|E = e)$ on the noise variables (which need not be independent anymore). What would have been the value of the variable X if the value of Z had been equal to z? The answer to this counterfactual question is given by first conditioning the noise variables U on the evidence $E = e$ (abduction) and then by an intervention which sets $Z = z$ on this modified SCM. More precisely, the counterfactual distribution for X is defined by $P(X_{Z\leftarrow z}|E = e)$.

In our case $O = \{Z, X^\perp, X^\parallel, \widehat{Y}\}$. The DAG G of the relevant SCM is shown in Figure 2. It expresses the fact that the prediction $\widehat{Y}$ depends on the variable Z only through $X^\parallel$ and that $X^\parallel$ and $X^\perp$ are correlated as revealed by experiment and displayed in Figure 1.

The protected variable we act upon is Z. Let us assume it is a balanced categorical variable with k values $Z \sim \mathrm{Cat}(\frac{1}{k}, \cdots, \frac{1}{k})$. Let us moreover assume that $X = (X^\perp, X^\parallel)$ is distributed as a multivariate Gaussian whose mean $\boldsymbol{\mu}(z)$ and covariance $\boldsymbol{\Sigma}(z)$ depend on z. The conditional distribution $P(X|Z=z)$ is thus given by

$$P(X|z) = \mathcal{N}(\boldsymbol{\mu}(z), \boldsymbol{\Sigma}(z)),$$
$$\boldsymbol{\mu}(z) := \begin{bmatrix} \boldsymbol{\mu}^\perp \\ \boldsymbol{\mu}^\parallel(z) \end{bmatrix}, \quad \boldsymbol{\Sigma}(z) := \begin{bmatrix} \boldsymbol{\Sigma}^{\perp\perp} & \boldsymbol{\Sigma}^{\perp\parallel}(z) \\ \boldsymbol{\Sigma}^{\parallel\perp}(z) & \boldsymbol{\Sigma}^{\parallel\parallel}(z) \end{bmatrix}, \tag{5}$$

where both $\boldsymbol{\mu}^\perp$ and $\boldsymbol{\Sigma}^{\perp\perp}$ are independent of z because $X^\perp$ is not impacted by z. Using standard properties of multivariate Gaussian we infer from (5) that the conditional $P(X^\parallel|X^\perp = x^\perp, Z=z)$ is linear-Gaussian for each z

$$P(X^\parallel|x^\perp, z) = \mathcal{N}\left(\boldsymbol{\mu}^\parallel(x^\perp, z), \boldsymbol{\Sigma}^\parallel(z)\right),$$
$$\boldsymbol{\mu}^\parallel(x^\perp, z) := \mathbf{W}(z)\, x^\perp + \mathbf{b}(z), \tag{6}$$

where $\boldsymbol{\Sigma}^\parallel(z), \mathbf{W}(z), \mathbf{b}(z)$ can be expressed as closed-form expressions involving the components of $\boldsymbol{\mu}(z)$ and $\boldsymbol{\Sigma}(z)$. An SCM which is compatible with the above can now be proposed by introducing appropriate noise variables $U_Z, U^\perp, U^\parallel$ and linear functions $f_Z, f_{X^\perp}, f_{X^\parallel}$ associated with the $Z, X^\perp, X^\parallel$ nodes in G. The predictor $\widehat{Y}$ is identified with $f_{\widehat{Y}}$ and has no associated noise variable. As we shall argue below, the noise variable $U^\parallel := (U_1^\parallel, \ldots, U_k^\parallel)$ for $X^\parallel$ should have as many components as the number of values Z can take which is k. We shall write $U^\parallel(Z)$ to mean $U_z^\parallel$ when $Z=z$. Using the definitions for $\boldsymbol{\mu}^\perp, \boldsymbol{\mu}^\parallel(x^\perp, z)$ and $\boldsymbol{\Sigma}^{\perp\perp}$ introduced in (5) and (6) we then define the SCM which implements the distribution $P(O)$ and the causality relations defined by G as

$$\begin{aligned}
Z &= f_Z(U_Z) := U_Z, && U_Z \sim \mathrm{Cat}(\tfrac{1}{k}, \ldots, \tfrac{1}{k}), \\
X^\perp &= f_{X^\perp}(U^\perp) := \boldsymbol{\mu}^\perp + \boldsymbol{\Sigma}^{\perp\perp} U^\perp, && U^\perp \sim \mathcal{N}(\mathbf{0}, \mathbf{1}), \\
X^\parallel &= f_{X^\parallel}(X^\perp, Z; U^\parallel) && U_z^\parallel \sim \mathcal{N}(\mathbf{0}, \mathbf{1}) \\
&:= \boldsymbol{\mu}^\parallel(X^\perp, Z) + \boldsymbol{\Sigma}^\parallel(Z) U^\parallel(Z) && \text{for } z = 1, \ldots, k, \\
\widehat{Y} &= f_{\widehat{Y}}(X^\perp, X^\parallel).
\end{aligned} \tag{7}$$

Suppose now that we observe the evidence $e := (Z(s), X^\perp(s), X^\parallel(s)) := (z, x^\perp, x^\parallel)$ for some text s. Abduction amounts to reading off the values $u^\perp$ and $(u_1^\parallel, \ldots, u_k^\parallel)$ of the noise variables from (7). Once conditioned on the evidence, $U_z^\parallel$ and $U^\perp$ are obviously not random anymore, so neither are Z and $X^\perp$. The $X^\parallel$ variable on the other hand remains stochastic because knowing that $Z=z$ only freezes $U_z^\parallel$ but not $U_{z_1}^\parallel$ for $z_1 \neq z$. Next, the counterfactual distribution $P(X_{Z \leftarrow z_1}|E=e)$ is defined by acting on Z to set its value to $z_1 \neq z$ in the SCM (7) in which the only remaining noise is $U_{z_1}^\parallel$. The distribution $P(X_{Z\leftarrow z_1}^\parallel | E=e)$ is thus given by (6) by replacing z by z_1. Its expectation $\boldsymbol{\mu}^\parallel(x^\perp, z)$ is nothing but the $\parallel$ component of our CFR $x_{Z\leftarrow z}(x)$ as defined by a linear regression in (4). The $\perp$ component is deterministic $X_{Z\leftarrow z}^\perp = x^\perp$ in agreement with the second component in (4). In other words, our CFR is nothing but the expectation of a counterfactual as defined in Pearl's causal inference framework for the SCM (7), thus justifying our claim.

If we had refrained from defining as many noise variables $U_z^\parallel$ as there are different values of z, then the evidence e would have fully determined the value of $U^\parallel$, making the CFR fully deterministic. However, this would induce a geometric relationship between the locations of a representation x and that of its counterfactual $x_{Z\leftarrow z}$ for which there is no justification whatsoever in any text generation mechanism.

4 Datasets and training details

To assess our CFR generation model, we conducted a series of experiments on both synthetic and real world datasets. We first introduce a synthetic dataset, named EEEC+, to provide a ground truth in the form of CFs defined at the text level to which we later compare our CFRs. Next, we test CFRs for assessing causal effects on the realistic benchmark dataset CEBaB [1], which poses a greater challenge than EEEC+ because it is much smaller, sparse in concept labels and in which concept values are not determined by a local signal in the samples. At last, we leverage the real world dataset BiasInBios [7] to challenge our CFRs to provide useful substitutes for CFs in cases these are not available. This will motivate the practical usefulness of our CFRs as a tool for classifier explainability.

4.1 Datasets

EEEC+ We introduce a new synthetic dataset, EEEC+, as an extension of the existing EEEC dataset [11]. Both are well suited for evaluating the impact of protected attributes (the gender or perceived race of the individual referred to in a text) on downstream mood state classification. Compared with EEEC, besides increasing the diversity of templates, we also turned the binary race concept into a ternary one to extend the scope of evaluation of our CFR model. Information on the creation and structure of EEEC+ can be found in Supplementary Material A [16].

Each observation in EEEC+ is labelled with a binary gender (male of female), a ternary race (white American, Afro-American or Asian-American, which incidentally allows to go beyond just flipping a pair of races) and a mood state (joy, fear, sadness, anger or neutral).

We built both a balanced and an aggressive version of EEEC+. In the balanced version, mood state is uncorrelated with gender or race. In the aggressive version, a correlation has been induced by assigning 80% of 'joy' states and 20% of other mood states one specific value of the protected attribute (female for gender or Afro-American for race). Each observation in the balanced version of EEEC+ has been assigned one genuine CF generated by randomly selecting a counterfactual value for the protected attribute and automatically editing the text accordingly. Every EEEC+ version comprises 40,000 observations distributed across three stratified-by-mood-states splits, with 26,000 training (65%), 6,000 validation (15%), and 8,000 test samples (20%).

CEBaB This realistic dataset is well suited for evaluating the causal effect of a concept on a sentiment classification task [1]. It includes both 2,299 original restaurant reviews from OpenTable and human-edited counterfactual reviews in which an aspect of the dining experience (food, service, ambiance or noise) was modified. The analysis of causal effects in CEBaB is facilitated through the creation of edit pairs. These are pairs of observations from the same edit set that differ in the value of one aspect. An edit set comprises an original observation and all observations edited from that original observation. Observations have been annotated with multiply-validated sentiment ratings at the aspect level (mostly Positive, Negative, or Unknown labels) and at the review level (1 star (very negative) to 5 stars (very positive)). In this article, we use CEBaB's exclusive train set described in Abraham et al. [1] as training data. So we conducted the study on 5,117 observations distributed across three splits, with

1,755 training (34%), 1,673 validation (32%), and 1,689 test samples (33%).

BiasInBios This real world dataset is suited for studying gender biases in biography classification tasks [7]. It consists of short biographies collected through web scraping and labeled with binary gender and occupation (28 occupations in total). This dataset is notoriously gender-biased. We have used the dataset version introduced in Ravfogel et al. [21] which contains over 98% of the original dataset, as the full version is no longer available on the web. It comprises 399,423 biographies distributed across three stratified-by-occupation splits, with 255,710 training (65%), 39,369 validation (10%), and 98,344 test samples (25%).

4.2 Training details

In the subsequent analysis, each genuine observation is represented by the last hidden state of a frozen BERT (bert-base-uncased) [8] over the [CLS] token. The component $X^\perp$ results from an orthogonal projection onto $V^\perp$. The computation of $\boldsymbol{\mu}^\parallel(x^\perp, z)$ results from a linear regression via stochastic gradient descent with mean squared error objective and L^2-regularization for each value z of Z. Unless specified, we use the deterministic version of $x^\parallel_{Z\leftarrow z}(x^\perp)$ defined in (4).

In EEEC+, Z corresponds either to the gender (k=2) or to the race (k=3) while Y is a mood state with 5 discrete values. In CEBaB, Z corresponds to an aspect of the dining experience while Y is a sentiment rating with 5 discrete values. Aspects in CEBaB can be treated as ternary attributes ($k = 3$, that we refer to as a ternary setting) or binary ($k = 2$, a binary setting), depending on whether the 'Unknown' label is considered as a proper concept value. In BiasInBios, Z corresponds to the gender (k=2) and Y to an occupation with 28 discrete values.

Classifiers $\widehat{Y}$ and $\widehat{Z}$ are trained as one-vs-all logistic regressions with L^2-regularization. Validation data was used for shallow optimization. For EEEC+, an aggressive and a balanced training scenario were defined by training the classifier respectively on the aggressive and balanced version of EEEC+. Evaluations were conducted on test data (on balanced test data for EEEC+, irrespective of the training data distribution). When available, genuine counterfactuals of test data were used solely for evaluation purposes. Further training details are provided in Supplementary Material F [16].

5 Evaluation and results

5.1 Direct evaluation of CFRs on synthetic data

On our EEEC+ synthetic dataset, for which genuine CFs are available, we first evaluate the ability of our CFRs to mimic real observations by comparing the predictions of the $\widehat{Y}$ classifier when representations $X(s_{Z\leftarrow z})$ of reference CFs are replaced with their fictional counterpart $X(s)_{Z\leftarrow z}$. One possible metric for this evaluation is the proportion of observations for which predictions coincide. For a finer analysis, we can also evaluate the average distance in total variation between the probability distributions predicted by the classifiers $\widehat{Y}$ and $\widehat{Z}$.

Let $\mathcal{S} := \{(s_i, z_i)\}$ be a set of couples of text documents s_i and of CF values $z_i \neq Z(s_i)$. Define the proportion of identical predictions (PIP) by

$$\mathrm{PIP}_{\widehat{Y}}[\mathcal{S}] := \frac{1}{|\mathcal{S}|} \sum_{(s,z)\in\mathcal{S}} \mathbf{1}\left[\widehat{Y}(X(s_{Z\leftarrow z})) = \widehat{Y}(X(s)_{Z\leftarrow z})\right]. \quad (8)$$

Table 1: PIP and ATV for EEEC+ for each training scenario.

Training scenario		$\mathrm{PIP}_{\widehat{Y}}$	$\mathrm{ATV}_{\widehat{Y}}$	$\mathrm{PIP}_{\widehat{Z}}$	$\mathrm{ATV}_{\widehat{Z}}$
gender	balanced	82.66%	0.158	93.89%	0.067
gender	aggressive	71.83%	0.237	95.01%	0.057
race	balanced	82.86%	0.161	92.30%	0.105
race	aggressive	73.88%	0.228	91.15%	0.134

Table 2: $\mathrm{ATE}_{\widehat{Y}}$ and $\widehat{\mathrm{ATE}}_{\widehat{Y}}$ for EEEC+ for each training scenario.

Training scenario		$\mathrm{ATE}_{\widehat{Y}}$	$\widehat{\mathrm{ATE}}_{\widehat{Y}}$
gender	balanced	0.159	0.013
gender	aggressive	0.225	0.280
race	balanced	0.161	0.020
race	aggressive	0.192	0.211

The range of the PIP metric is $[0, 1]$, closer to 1 being better. Similarly, we define $\mathrm{PIP}_{\widehat{Z}}[\mathcal{S}]$.

Let $p_{\widehat{Y}}(x)$ be the probability distribution over Y-values used by the classifier $\widehat{Y}$. Define the average total variation (ATV) distance by

$$\mathrm{ATV}_{\widehat{Y}}[\mathcal{S}] := \frac{1}{|\mathcal{S}|} \sum_{(s,z)\in\mathcal{S}} \underbrace{\frac{1}{2}\left|p_{\widehat{Y}}(X(s_{Z\leftarrow z})) - p_{\widehat{Y}}(X(s)_{Z\leftarrow z})\right|}_{=:\,\mathrm{TV}_{\widehat{Y}}(s,z)}. \quad (9)$$

The range of the ATV metric is $[0, 1]$, closer to 0 being better. Similarly, we define $\mathrm{ATV}_{\widehat{Z}}[\mathcal{S}]$.

Results The results are shown in Table 1. For the $\widehat{Z}$ classifier, in all cases $\mathrm{PIP}_{\widehat{Z}} > 0.9$ and $\mathrm{ATV}_{\widehat{Z}}$ is close to 0, indicating that CFs and CFRs are largely processed in a similar way.

For $\widehat{Y}$'s, the results are more nuanced. In the balanced scenarios, CFs and CFRs lead to very similar predictions. However, the ability of CFRs to mimic CFs seems to deteriorate with the strength of the correlation between the predicted variable Y and the attribute Z being manipulated.

Results do not significantly improve if we use the stochastic version of the CFRs which takes into account the variance of the $X^\parallel$ component of the CFR in (7).

5.2 Treatment effect on synthetic data

In this subsection we use our synthetic dataset EEEC+ to argue that, in the relevant biased cases, our CFRs can be used to define good estimates for both the average treatment effect (ATE) at the population level $\mathcal{S}$ and, more significantly, for the treatment effect (TE) on each individual observation s.

Let's thus define the estimator $\widehat{\mathrm{ATE}}_{\widehat{Y}}$ and the corresponding estimator $\widehat{\mathrm{TE}}_{\widehat{Y}}(s, z)$ for individual effects by

$$\widehat{\mathrm{ATE}}_{\widehat{Y}}[\mathcal{S}] := \frac{1}{|\mathcal{S}|} \sum_{(s,z)\in\mathcal{S}} \underbrace{\frac{1}{2}\left|p_{\widehat{Y}}(X(s)_{Z\leftarrow z} - p_{\widehat{Y}}(X(s))\right|}_{=:\,\widehat{\mathrm{TE}}_{\widehat{Y}}(s,z)}. \quad (10)$$

Both $\widehat{\mathrm{ATE}}_{\widehat{Y}}$ and $\widehat{\mathrm{TE}}_{\widehat{Y}}$ should be thought of as estimators for corresponding quantities $\mathrm{ATE}_{\widehat{Y}}$ and $\mathrm{TE}_{\widehat{Y}}$ defined just as in (10) except that the representations $X(s_{Z\leftarrow z})$ of the true CFs are substituted for the CFRs $X(s)_{Z\leftarrow z}$.

Results Let us first notice that results in Table 2 show, as expected, that aggressive training scenarios yield higher $\mathrm{ATE}_{\widehat{Y}}$ and $\widehat{\mathrm{ATE}}_{\widehat{Y}}$ than balanced ones. Moreover, when $\widehat{Y}$ is not Z-biased the $\widehat{\mathrm{ATE}}_{\widehat{Y}}$

is close to 0 while the $\mathrm{ATE}_{\widehat{Y}}$ is close to the $\mathrm{ATV}_{\widehat{Y}}$ in Table 1 as expected.

In aggressive training scenarios, $\widehat{\mathrm{ATE}}_{\widehat{Y}}[\mathcal{S}]$ overestimates $\mathrm{ATE}_{\widehat{Y}}[\mathcal{S}]$ and we suspect that a small fraction of the observations for which CFRs are poor substitutes for CFs degrade the estimator. Our aim is thus to show that there is actually a large fraction of observations in $\mathcal{S}$ for which $\widehat{\mathrm{ATE}}_{\widehat{Y}}$ is a good estimate of the true $\mathrm{ATE}_{\widehat{Y}}$. More precisely, we show that the observations for which the estimate is bad coincide with a tiny fraction for which the $\mathrm{TV}_{\widehat{Y}}$ defined in (9) are the largest. We do this by constructing a sequence of $|\mathcal{S}|$ nested subsets $\mathcal{S}_1 \subset \ldots \mathcal{S}_n \subset \mathcal{S}_{n+1} \subset \ldots \mathcal{S}_{|\mathcal{S}|} = \mathcal{S}$ along which $\mathrm{ATV}[\mathcal{S}_n] < \mathrm{ATV}[\mathcal{S}_{n+1}]$.

A correlation analysis confirm that $\widehat{\mathrm{ATE}}_{\widehat{Y}}$ is a very good estimator for $\mathrm{ATE}_{\widehat{Y}}$ over a large fraction of the observations in $\mathcal{S}$ in aggressive scenarios. This is not by chance but it is a consequence of a strong linear correlation between individual effects estimations $\widehat{\mathrm{TE}}_{\widehat{Y}}$ and their actual values $\mathrm{TE}_{\widehat{Y}}$ within most subsets $\mathcal{S}_n$ of $\mathcal{S}$. For gender, 66% of the observations have a very strong correlation in the sense that their correlation coefficient $\rho > 0.75$, while 91% have $\rho > 0.5$ (see also Figure 3 in Supplementary Material B.1 [16]). Moreover the regression coefficient never deviates much from 1 along the nested $\mathcal{S}_n$. Similar result hold for the race. These facts help build confidence in the potential to use our CFRs as reliable substitutes for CFs in practice.

5.3 Treatment effect on realistic data

In this section, we use the CEBaB dataset [1] to demonstrate that our CFRs can be used as reliable substitutes for real counterfactuals in realistic settings. We will also make the case that our CFRs are good candidates for a strong and easy-to-implement baseline for future work on explainability.

Recall that Y corresponds to a sentiment rating in CEBaB. Let's thus define an ATE for this rating along the same lines as in (10). Our definition is also meant to facilitate comparison with [1].[5] Let $\mathcal{S}^{(z_1,z_2)}$ be the set of texts s for which $Z(s) = z_1$ and for which a CF exists such that $Z(s_{Z\leftarrow z_2}) = z_2$.

$$\widehat{\mathrm{ATE}}^{\mathrm{score}}_{\widehat{Y}}[\mathcal{S}^{(z_1,z_2)}] := \frac{1}{|\mathcal{S}^{(z_1,z_2)}|} \sum_{s\in\mathcal{S}^{(z_1,z_2)}} \Big(\widehat{Y}(X(s)_{Z\leftarrow z}) - \widehat{Y}(X(s))\Big). \tag{11}$$

The quantity $\widehat{\mathrm{ATE}}^{\mathrm{score}}_{\widehat{Y}}$ can be thought of as an estimator for an $\mathrm{ATE}^{\mathrm{score}}_{\widehat{Y}}$ defined as in (11) except that the representations $X(s_{Z\leftarrow z_2})$ of true CFs are used instead of CFRs $X(s)_{Z\leftarrow z_2}$.

To assess how well CFRs account for individual causal effects and also having various approaches to explainability in mind, we adapt the error measure introduced in [1] (definition 3) to our CFRs:

$$\mathrm{Error}_{\widehat{Y}}[\mathcal{S}] = \frac{1}{|\mathcal{S}|} \sum_{(s,z)\in\mathcal{S}} \mathrm{Dist}\Big(p_{\widehat{Y}}\big(X(s_{Z\leftarrow z})\big) - p_{\widehat{Y}}\big(X(s)\big)\,,\; p_{\widehat{Y}}\big(X(s)_{Z\leftarrow z}\big) - p_{\widehat{Y}}\big(X(s)\big)\Big). \tag{12}$$

where Dist is a distance between the observed individual effects and the individual effects estimated using CFRs. Following [1], we consider three distance measures: the *cosine* distance which is influenced only by the directions of the effects, the *normdiff* which is the absolute difference between the Euclidian norms of each effect and is influenced only by the magnitude of the effects and at last the L^2 distance which is the norm of the difference of effects and is influenced by both the magnitude and direction of the effects.[6]

[5] $\widehat{\mathrm{ATE}}^{\mathrm{score}}_{\widehat{Y}}$ corresponds to the evaluation of the scalar version of $\widehat{\mathrm{CaCE}}_{\widehat{Y}}$ in Abraham et al. [1] (definition 4)

Table 3: Average treatment effects (and standard deviations) averaged over 10 different seeds. Rows are concepts, columns are concept interventions, and each entry indicates how the average rating increases or decreases when the concept is intervened on with the given direction. Aspect labels are Positive, Negative or Unknown. Our CFRs were trained in a ternary setting.

(a) $\mathrm{ATE}^{\mathrm{score}}_{\widehat{Y}}$ (reference)

	Neg. to Pos.	Neg. to Unk.	Pos. to Unk.
food	1.83 (±0.02)	0.93 (±0.02)	−0.81 (±0.02)
service	1.36 (±0.03)	0.84 (±0.02)	−0.42 (±0.02)
ambiance	1.24 (±0.03)	0.76 (±0.02)	−0.45 (±0.01)
noise	0.73 (±0.02)	0.46 (±0.02)	−0.19 (±0.02)

(b) $\widehat{\mathrm{ATE}}^{\mathrm{score}}_{\widehat{Y}}$ (using CFRs)

	Neg. to Pos.	Neg. to Unk.	Pos. to Unk.
food	2.15 (±0.12)	0.86 (±0.11)	−0.57 (±0.20)
service	2.02 (±0.13)	0.85 (±0.10)	−0.37 (±0.15)
ambiance	1.73 (±0.21)	1.15 (±0.05)	−0.33 (±0.06)
noise	0.53 (±0.12)	0.20 (±0.07)	−0.24 (±0.04)

Table 4: $\mathrm{Error}_{\widehat{Y}}$ (and standard deviations) for a 5-way sentiment linear classifier on top of a frozen bert-base-uncased previously fine-tuned for this task. Rows are distances. Columns are explanatory methods. **Lower is better**. Best results per metric are highlighted in bold. Results are averaged over 10 random initializations. The *random* explainer takes the difference between two random probability vectors as the predicted effect.

	random	approximate counterfactuals	CFR (binary setting)	CFR (ternary setting)
cosine	1.00 (±0.01)	**0.83 (± 0.03)**	0.86 (±0.05)	0.87 (±0.03)
normdiff	0.67 (±0.08)	0.49 (±0.06)	0.49 (±0.05)	**0.41 (± 0.04)**
L^2	0.93 (±0.11)	0.81 (±0.14)	0.81 (±0.14)	**0.71 (± 0.10)**

The above metrics can be easily adapted to another counterfactual generation method by replacing appropriately $X(s)_{Z\leftarrow z}$ in (11) or (12) thus allowing comparison. For this last purpose, we adapt the so-called *approximate counterfactuals* method introduced in [1], which is a baseline for explanatory methods and surprisingly proves to be the best-performing one. Starting with an edit pair comprising an original observation and a genuine CF, this method consists in selecting as approximate CF another original observation that has the same labels for concepts as the genuine CF. More details on this method are given in Supplementary Material E [16].

Results First, we note that the linear classifier captures the real-world effects well, as confirmed by the results in Table 3a which are well-aligned with the empirical estimates of the causal effect (see Table 3d in [1]).

Next, the evaluations of $\widehat{\mathrm{ATE}}^{\mathrm{score}}_{\widehat{Y}}$ in Table 3b using CFRs as substitutes for genuine counterfactuals are well-aligned with the reference results in Table 3a. Results achieved using CFRs trained in a binary setting are also well-aligned (see complementary results in Supplementary Material B.2 [16]). In a realistic context, this confirms the explanatory power of using CFRs as substitutes for genuine counterfactuals to estimate real-world causal effects.

The results in Table 4 show that CFRs provide a better overall estimate of individual causal effects in terms of the different metrics considered than approximate CFs. Moreover, the values for *normd-*

[6] Comparison of individual effects in section 5.2 is based on the distance in total variation rather than the Euclidian norm of the effects, which in both cases amounts to considering magnitude only.

Table 5: Pairs of occupations with the largest values of $\widehat{\Pi}_{\text{male},(y_f,y_t)}$ (top) and $\widehat{\Pi}_{\text{female},(y_f,y_t)}$ (bottom), i.e., the percentage of men's (resp. women's) biographies that are only correctly predicted by a linear classifier as y_t when their gender attribute is swapped for which the predicted label changes from the wrong prediction y_f. In bold, the pairs already identified in [7]

y_f (false prediction)	y_t (true occupation)	$\widehat{\Pi}_{\text{male},(y_f,y_t)}$
architect	**interior designer**	38.46%
attorney	**paralegal**	33.33%
professor	**dietitian**	13.95%
professor	psychologist	9.54%
teacher	**yoga teacher**	7.50%
professor	teacher	6.03%
surgeon	chiropractor	5.88%
photographer	**interior designer**	5.13%
professor	yoga teacher	5.00%
surgeon	dietitian	4.65%

y_f (false prediction)	y_t (true occupation)	$\widehat{\Pi}_{\text{female},(y_f,y_t)}$
physician	**surgeon**	10.77%
physician	chiropractor	10.53%
teacher	**pastor**	9.47%
professor	**surgeon**	9.43%
nurse	dietitian	8.29%
journalist	comedian	7.20%
dietitian	personal trainer	6.54%
model	comedian	6.25%
model	dj	5.88%
nurse	surgeon	5.72%

iff and L^2 are quantitatively close to the best values reported in [1] on a closely-related task. Thus CFRs, because they are computationally inexpensive and easy-to-implement, seem to us to be an ideal candidate for a baseline in future works on explainability.

5.4 Explaining predictions on real-world data

In this subsection we investigate the usefulness of our CFRs as a practical tool for providing a detailed bias analysis in a real world example. We select the BiasInBios dataset [7] because it is notoriously biased and approximate genuine counterfactuals can be generated from the observations. De-Arteaga et al. [7] create these CFs by simply swapping the gender z for its opposite $\bar{z}$ in each biography[7]. These CFs thus provides us with a ground truth against which we can compare our CFRs.

The bias analysis in De-Arteaga et al. [7] revolves around a somewhat involved miss-classification rate that we now embark to adapt for our purpose. Let's fix a gender z and two occupations $y_f \neq y_t$. First consider the subset of sentences s with a gender z which are misclassified as y_f when the classifier $\widehat{Y}$ uses the original representation $X(s)$ while it makes a correct prediction y_t when using the CFR $X(s)_{Z\leftarrow\bar{z}}$ for the swapped gender $\bar{z}$. We next consider the larger subset where we relax the misclassification constraint. We then define a misclassification rate $\widehat{\Pi}_{z,(y_f,y_t)}$ as a ratio between the cardinalities of the former to the latter

$$\widehat{\Pi}_{z,(y_f,y_t)} := \frac{\left|\left\{s | \widehat{Y}(X(s)) = y_f, \widehat{Y}(X(s)_{Z\leftarrow\bar{z}}) = Y(s) = y_t, Z(s) = z\right\}\right|}{\left|\left\{s | \widehat{Y}(X(s)_{Z\leftarrow\bar{z}}) = Y(s) = y_t, Z(s) = z\right\}\right|}. \tag{13}$$

This misclassification rate $\widehat{\Pi}_{z,(y_f,y_t)}$ can be thought of as an estimator for a quantity $\Pi_{z,(y_f,y_t)}$ defined as in (13) except that the representations $X(s_{Z\leftarrow\bar{z}})$ of the genuine CFs are used instead of the CFRs. At last we define $\widehat{\Pi}_z^{\max}$ as the maximum of $\widehat{\Pi}_{z,(y_f,y_t)}$ over all possible pairs (y_f, y_t).

Results Results in Table 5 align with those in [7]. We recover 8 of the 10 pairs of occupations (y_f, y_t) that were identified in this study when we use our CFRs as substitutes of the genuine CFs. These results qualitatively reflect a tropism that favors the prediction of occupations such as 'nurse' for women working in the medical field, or 'model' for those in the arts. Similarly, the results clearly reflect a tendency of the classifier to associate a man in the medical field with the occupation of 'surgeon', or a man in the education field with the occupation of 'professor'.

5.5 CFRs beyond explainability

Part of the usefulness of our CFRs stems from the possibility to compute them even in circumstances where no explicit text CF would make sense. However, as a simple consistency check, it is tempting to ask how CFR work on single word representations such as GloVe embeddings, which are notoriously gender-biased [5, 22]. More precisely, for a word s with a given $Z(s) = z$ we can ask which word s' has the closest embedding $X(s')$ to the CFR $X(s)_{Z\leftarrow z'}$, thus providing an explicit approximate textual counterfactual. We performed many such checks in Supplementary Material C [16] when Z corresponds to a gender bias. For example the word s = "bridesmaids" becomes s' = "groomsmen" through such an indirect gender switch. Most examples are indeed convincing explicit gender counterfactuals.

Another classic use of counterfactuals is to improve the fairness of classifiers. On the BiasInBios dataset introduced above, we show in Supplementary Material D [16] that CFRs can be leveraged to mitigate the bias in a downstream classification task via data augmentation, i.e. by integrating CFRs into an unbalanced training set to make it more balanced.

6 Conclusion

In this paper, we propose a straightforward approach, based on linear regressions in the representation space, to generate minimally disruptive counterfactual representations (CFRs) for text documents. These CFRs offer an effective way of altering the value of a protected text attribute, even in scenarios where constructing a corresponding meaningful sentence explicitly proves impossible. The theoretical soundness of these CFRs is demonstrated by their alignment with the definition within Pearl's causal inference framework for a natural SCM.

These CFRs can be harnessed to provide fine-grained explanations for the decisions made by a text classifier. In various synthetic and realistic contexts, they also prove very useful for quantitatively assessing causal effects linked to changes in concept values in textual data. They could in particular come in handy as a strong baseline for such tasks. Furthermore, in contexts where the fairness of a text classifier is crucial, CFRs offer a method to augment a training set with additional observations, thereby making it more balanced. This confirms both the practical usefulness and the quality of our CFRs.

An interesting avenue for future research involves enhancing our CFRs by using non-linear regressions. This development is likely to require a parallel exploration of non-linear erasure methods, which is an open problem by itself.

[7] These CFs are flawed because other factors correlated with gender are not modified by swapping gender indicators.

References

[1] E. D. Abraham, K. D'Oosterlinck, A. Feder, Y. Gat, A. Geiger, C. Potts, R. Reichart, and Z. Wu. CEBaB: Estimating the causal effects of real-world concepts on nlp model behavior. *Advances in Neural Information Processing Systems*, 35:17582–17596, 2022.

[2] B. Barr, M. R. Harrington, S. Sharpe, and C. B. Bruss. Counterfactual explanations via latent space projection and interpolation. *arXiv preprint arXiv:2112.00890*, 2021.

[3] S. Basu Roy Chowdhury, N. Monath, K. A. Dubey, A. Ahmed, and S. Chaturvedi. Robust concept erasure via kernelized rate-distortion maximization. *Advances in Neural Information Processing Systems*, 36, 2023.

[4] N. Belrose, D. Schneider-Joseph, S. Ravfogel, R. Cotterell, E. Raff, and S. Biderman. LEACE: Perfect linear concept erasure in closed form. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[5] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[6] N. Calderon, E. Ben-David, A. Feder, and R. Reichart. DoCoGen: Domain counterfactual generation for low resource domain adaptation. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7727–7746, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[7] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai. 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, 2019.

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, volume 1, pages 4171—-4186, Minneapolis, MN, USA, 2019. Association for Computational Linguistics.

[9] Y. Elazar and Y. Goldberg. Adversarial removal of demographic attributes from text data. In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21, Brussels, Belgium, Oct.-Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1002. URL https://aclanthology.org/D18-1002.

[10] Y. Elazar, S. Ravfogel, A. Jacovi, and Y. Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

[11] A. Feder, N. Oved, U. Shalit, and R. Reichart. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386, 2021.

[12] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226, 2019.

[13] D. Kaushik, E. Hovy, and Z. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.

[14] D. Kaushik, A. Setlur, E. H. Hovy, and Z. C. Lipton. Explaining the efficacy of counterfactually augmented data. In *International Conference on Learning Representations*, 2021.

[15] M. J. Kusner, J. Loftus, C. Russell, and R. Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[16] P. Lemberger and A. Saillenfest. Explaining text classifiers with counterfactual representations. *arXiv preprint arXiv:2402.00711, Full version of this paper*, 2024. URL https://arxiv.org/abs/2402.00711.

[17] N. Madaan, I. Padhi, N. Panwar, and D. Saha. Generate your counterfactuals: Towards controlled counterfactual generation for text. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35 (15), pages 13516–13524, 2021.

[18] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617, 2020.

[19] J. Pearl. *Causality*. Cambridge university press, 2009.

[20] J. Peters, D. Janzing, and B. Schölkopf. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

[21] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.647. URL https://aclanthology.org/2020.acl-main.647.

[22] S. Ravfogel, F. Vargas, Y. Goldberg, and R. Cotterell. Adversarial concept erasure in kernel space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6034–6055, 2022.

[23] S. Ravfogel, Y. Goldberg, and R. Cotterell. Log-linear guardedness and its implications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9413–9431, 2023.

[24] S. Shao, Y. Ziser, and S. B. Cohen. Gold doesn't always glitter: Spectral removal of linear and nonlinear guarded attribute information. In A. Vlachos and I. Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1611–1622, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-main.118.

[25] V. Veitch, A. D'Amour, S. Yadlowsky, and J. Eisenstein. Counterfactual invariance to spurious correlations in text classification. *Advances in Neural Information Processing Systems*, 34:16196–16208, 2021.

[26] Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig. Controllable invariance through adversarial feature learning. *Advances in neural information processing systems*, 30, 2017.

[27] X. Zeng, Y. Li, Y. Zhai, and Y. Zhang. Counterfactual generator: A weakly-supervised method for named entity recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7270–7280, 2020.