CausaLM: Causal Model Explanation Through Counterfactual Language Models

Amir Feder
feder@campus.technion.ac.il
Nadav Oved
nadavo@campus.technion.ac.il
Uri Shalit
urishalit@technion.ac.il
Roi Reichart
roiri@technion.ac.il

Understanding predictions made by deep neural networks is notoriously difficult, but also crucial to their dissemination. As all ML-based methods, they are as good as their training data, and can also capture unwanted biases. While there are tools that can help understand whether such biases exist, they do not distinguish between correlation and causation, and might be ill-suited for text-based models and for reasoning about high level language concepts. A key problem of estimating the causal effect of a concept of interest on a given model is that this estimation requires the generation of counterfactual examples, which is challenging with existing generation technology. To bridge that gap, we propose CausaLM, a framework for producing causal model explanations using counterfactual language representation models. Our approach is based on fine-tuning of deep contextualized embedding models with auxiliary adversarial tasks derived from the causal graph of the problem. Concretely, we show that by carefully choosing auxiliary adversarial pre-training tasks, language representation models such as BERT can effectively learn a counterfactual representation for a given concept of interest, and be used to estimate its true causal effect on model performance. A byproduct of our method is a language representation model that is unaffected by the tested concept, which can be useful in mitigating unwanted bias ingrained in the data. [1]

1. Introduction

The rise of deep learning models (DNNs) has produced better prediction models for a plethora of fields, particularly for those that rely on unstructured data, such as computer vision and natural language processing (NLP) (Peters et al. 2018; Devlin et al. 2018). In recent years, variants of these models have disseminated into many industrial applications, varying from image recognition to machine translation (Szegedy et al. 2016; Wu et al. 2016; Aharoni, Johnson, and Firat 2019). In NLP, they were also shown to produce better language models, and are being widely used both for language representation and for classification in nearly every sub-field (Tshitoyan et al. 2019; Gao et al. 2019; Lee et al. 2020; Feder et al. 2020).

---

[1] Our code and data are available at: https://amirfeder.github.io/CausaLM/

While DNNs are very successful, this success has come at the expense of model explainability and interpretability. Understanding predictions made by these models is difficult, as their layered structure coupled with non-linear activations do not allow to reason about the effect of each input feature on the model's output. In the case of text-based models this problem is amplified. Basic textual features are usually comprised of n-grams of adjacent words, but these features alone are limited in their ability to encode meaningful information conveyed in the text. While abstract linguistic concepts, such as topic or sentiment, do express meaningful information, they are usually not explicitly encoded in the model's input. Such concepts might push the model towards making specific predictions, without being directly modeled and therefore interpreted. Such interpretability problems affect the dissemination of DNNs in a variety of fields, particularly in scientific applications to fields such as healthcare and the social sciences that rely on model interpretability for deployment.

Recently, there have been many attempts to build tools that allow for DNN explanations and interpretations (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017), which have developed into a sub-field often referred to as Blackbox-NLP (Linzen et al. 2019). These tools can be roughly divided into *local explanations*, where the effect of a feature on the classifier's prediction for a specific example is tested, and *global explanations*, which measure the general effect of a given feature on a classifier. A prominent research direction in DNN explainability involves utilizing network artifacts such as attention mechanisms, which are argued to provide a powerful representation tool (Vaswani et al. 2017) to explain how certain decisions are made (Jain and Wallace 2019; Wiegreffe and Pinter 2019). Alternatively, there have been attempts to estimate simpler, more easily-interpretable models, around test examples or their hidden representations (Ribeiro, Singh, and Guestrin 2016; Kim et al. 2018).

Unfortunately, existing model explanation tools often rely on local perturbations of the input and compute shallow correlations, which can result in misleading, and sometimes wrong, interpretations. This problem arises, for example, in cases where two concepts that can potentially explain the predictions of the model are strongly correlated with each other. As an example, in Section 2 we discuss the problem of measuring the impact of two concepts: *Adjectives* and *Political Figure*, on a sentiment classifier. An explanation model that only considers correlations might show that the mention of a political figure is strongly correlated with the prediction, leading to worries about the classifier having political bias. However, such a model cannot indicate whether the political figure is in fact the cause of the prediction, or whether it is actually the type of adjectives used that is the true cause of the classifier output, suggesting that the classifier is not politically biased.

A natural solution would be to generate counterfactual examples and compare the model prediction for each example with its prediction for the counterfactual. That is, one needs a controlled setting where it is possible to compute the difference between an actual observed text, and what the text would have been had a specific concept (e.g. a political figure) not existed in it. However, in natural language this is often too hard to do automatically with generative models, and too costly to do manually, particularly for abstract concepts such as topics or sentiment.

There have been some attempts to construct counterfactuals for generating local explanations. Specifically, Goyal et al. (2019b) proposed changing the pixels of an image to those of another image classified differently by the classifier, in order to compute the effect of those pixels. However, as this method takes advantage of the spatial structure of images, it is hard to replicate their process with texts. Vig et al. (2020) offered to use mediation analysis to study which parts of the DNN are pushing towards specific decisions by querying the language model. While their work further highlights the usefulness of counterfactual examples as a basis for answering causal questions in model interpretation, they create counterfactual examples manually, by changing specific tokens in the original example. Unfortunately, such strategies do not support automatic estimation of the causal effect that high-level concepts have on model performance.

To alleviate these difficulties, in this paper we propose, we propose to manipulate the representation of the text and not the text itself. By creating a text encoder that is not affected by a specific concept of interest, we can compute the *counterfactual representation*. Our explanation method, which we name *Causal Model Explanation through Counterfactual Language Models* (*CausaLM*), receives the classifier's training data and a concept of interest as input, and outputs the causal effect of the concept on the classifier in the test set. It does that by pre-training an additional instance of the language representation model employed by the classifier, with an adversarial component designed to "forget" the concept of choice, while keeping the other "important" concepts represented. Following the additional training step, the representation produced by this counterfactual model can be used to measure the concept's effect on the classifier's prediction for each test example, by comparing the classifier performance with the two representations.

We start by providing motivation for why causal inference is crucial for model explanations, using two real-world examples (Section 2). Then, we dive into the link between causality and interpretability (Section 3), and discuss how to estimate causal effects using language representations (Section 4). After defining the causal estimator and discussing the challenges of producing counterfactual examples, we discuss methods for generating such examples, pointing on the challenges of this approach (Section 5.1). With those options laid out, we move to describing how we can approximate counterfactual examples through manipulation of the language representation (Section 5.2).

To test our method, we introduce in Section 6 four novel datasets, three of which include counterfactual examples for a given concept. Building on those datasets, we present in Section 7 four cases where a BERT-based representation model can be modified to ignore concepts such as *Adjectives*, *Topics*, *Gender* and *Race*, in various settings involving sentiment and mood state classification (Section 7). To prevent a loss of information on correlated concepts, we further modify the representation to remember such concepts while forgetting the concept whose causal effect is estimated. While in most of our experiments we test our methods in controlled settings, where the true causal concept effect can be measured, our approach can be used in the real-world, where such ground truth does not exist. Indeed, in our analysis we provide researchers with tools to estimate the quality of the causal estimator without access to gold standard causal information.

Using our newly created datasets, we estimate the causal effect of concepts on a BERT-based classifier utilizing our intervention method and compare to the ground truth causal effect, computed with manually created counterfactual examples (Section 8). To equip researchers with tools for using our framework in the real-world, we provide an analysis of what happens to the language representation following the intervention, and discuss how to choose adversarial training tasks effectively (Section 8.2). As our approach relies only on interventions done prior to the supervised task training stage, it is not dependent on BERT's specific implementation and can be applied whenever a pre-trained language representation model is used. We also show that our counterfactual models can be used to mitigate unwanted bias in cases where its effect on the classifier can negatively affect outcomes. Finally, we discuss the strengths and limitations of our approach, and propose how to use causal inference to further improve model interpretations in NLP (Section 9).

We hope that this research will spur more interest in the usefulness of causal inference for better-understanding DNNs and for creating more robust models, within the NLP community and beyond.

2. Motivation

Causal and concept-based explanations are crucial for scientific applications, and hinder further use of useful prediction models in many domains. Failing to account for the causal effect of

concepts on text classifiers can potentially lead to biased, unfair, misinterpreted and incorrect predictions. As models are dependent on the data they are trained on, a bias existing in the data could potentially result in a model that under-performs when this bias no longer holds in the test set. In clinical settings this risk is amplified, as using models that rely on unwanted concepts such as a doctor's writing style can put unnecessary risk on patients.

To illustrate these problems, consider the example presented in Figure 1, which will be our running example throughout the paper. Suppose we have a binary classifier, trained to predict the sentiment conveyed in news articles. Say we hypothesize that the choice of adjectives is driving the classification decision, something that has been discussed previously in computational linguistics (Pang, Lee, and Vaithyanathan 2002). However, if the text is written about a controversial figure, it could be that the presence of its name, or the topics that it induces are what is driving the classification decision, and not the use of adjectives. The text in the figure is an example of such a case, where both adjectives and the mentioning of politicians seem correlated, and could be driving the classifier's prediction. Estimating the effect of Donald Trump's presence in the text on the predictions of the model is also hard, as this presence clearly affects the choice of adjectives, the other political figures mentioned in the text and probably many additional textual choices.

President **Trump** did his best imitation of **Ronald Reagan** at the State of the Union address, falling just short of declaring it Morning in America, the **iconic** imagery and message of a campaign ad that **Reagan** rode to re-election in 1984. **Trump** talked of Americans as pioneers and explorers; he lavished praise on members of the military, several of whom he recognized from the podium; he **optimistically** declared that the best is yet to come. It was a **masterful** performance – but behind the **sunny** smile was the same old **Trump**: **petty**, **angry**, **vindictive** and **deceptive**. He refused to shake the hand of House Speaker **Nancy Pelosi**, a snub she returned in kind by ostentatiously ripping up her copy of the President's speech at the conclusion of the address, in full view of the cameras.

Figure 1: An example of a political commentary piece published at `https://edition.cnn.com`. Highlighted in **blue** and **red** are names of political figures from the US Democratic and Republican parties, respectively. Adjectives are highlighted in **green**.

Training a generative model to condition on a concept, such as the choice of adjectives, and produce counterfactual examples that only differ by this concept is still intractable in most cases involving natural language (see Section 5.1 for a more detailed discussion). While there are instances where this seems to be improving (Semeniuta, Severyn, and Barth 2017; Fedus, Goodfellow, and Dai 2018), generating a version of the example presented in Figure 1 where a different political figure is being discussed while keeping other concepts unaffected is very hard (Radford et al. 2018, 2019). Alternatively, our key technical observation is that instead of generating a counterfactual text we can more easily generate a counterfactual textual representation, based on adversarial training.

It is important to note that it is not even necessarily clear what are the concepts that should be considered as the "generating concepts" of the text. In the example above we only consider adjectives and the political figure, but there are other concepts that generate the text, such as the topics being discussed, the sentiment being conveyed and others. The number of concepts that would be needed and their coverage of the generated text are also issues that we touch on below. The choice of such *control concepts* depends on our model of the world, as in the *causal graph* example presented in Figure 3. In our experiments we control for such concepts, as our model of the world dictates both *treated concepts* and *control concepts*.

While failing to estimate the causal effect of a concept on a sentiment classifier is harmful, it pales in comparison to the potential harm of wrongfully interpreting clinical prediction models. If a model is trained on clinical notes to predict clinically important factors, the need for understanding the model is amplified. If it were the case that the model is relying on textual features that are doctor or hospital specific, it could lead to devastating implications.

For example, we can look at the (fake) clinical note presented in Figure 2. In this note, the patient's mental health is discussed extensively, with a verbose description and much detail. As the description is lengthy and sometimes repetitive, it could be summarized without losing too much clinically relevant information. In this case, a classifier that heavily relies on the doctor's verbose style could fail when given a short and concise note which still contains the same clinical information.

In this note, we highlight words by their length, a feature described previously as a proxy for writing style (Sari, Stevenson, and Vlachos 2018). Looking at the words highlighted in red and in orange, it is clear that changing the writing style of the note would require a significant intervention. At the same time, deleting long words or replacing them would significantly affect the structure and content of the note. Moreover, long words are correlated with the note's section (such as the *Vital Signs* section, which contains very short words), which could be a potential confounder. If we intervene and replace longer words with shorter synonyms, we might also change some concepts alongside the ones we mean to change, and there is no test that will tell us that ex-ante.

Concepts that influence both the label and other concepts, also known as *confounders*, could be extremely risky. Imagine a case where a doctor receives on average more patients that are of certain type, such as individuals with severe depression. In that case, a DNN could learn to associate this writing style as a signal for a depressed patient. Measuring the model's performance on notes written by that doctor would show promising results, but deploying such a model would risk patients health when used on patients of a different doctor. Without measuring the causal effect of the doctor's writing style on the classifier, we would not be able to tell to what extent the model is relying on it. [2]

Other, more complex relationships, might exist between concepts. For example, if a clinical note is describing the doctor's clinical treatment suggestion based on the patient's condition (i.e. depression, anxiety etc.), it would be hard to disentangle the clinical treatment suggestion from a specific condition (the causal graph for this example is presented in the bottom graph in Figure 5). Alternatively, it could also be that only the patient's depression or lack of it is causing the doctor's treatment suggestion, and the text is generated based on that suggestion alone (see the bottom graph in Figure 5). This would make it impossible to imagine a counterfactual text, where the upstream concept (depression) is changed but the one generated by it (treatment suggestion) remains fixed. In Section 5.3 we discuss alternative causal graphs that can be modeled and highlight the power and limitations of using a world model such as those discussed here to interpret DNNs.

3. Previous Work

Previous work on the intersection of DNN interpretations and causal inference, specifically in relation to NLP is rare. While there is a vast and rich literature on each of those topics alone, the gap between interpretability, causality and NLP is only now starting to close (Vig et al. 2020). To ground our work in those pillars, we survey here previous work in each. Specifically, we discuss

---

2 Note that while clinical notes are an important application domain, we do not consider them in our experiments as they were not publicly available to us. We plan to create such synthetic data in future work.

Status of patient:

Julie **is** worse today.

Target Symptoms:

Julie reports **that depressive** symptoms **continue. Her symptoms, she** reports, **are more** frequent **or more** intense. Anergia **is** present. **Increased** symptoms **of anhedonia are** present. Julie's **difficulty with concentrating has not** changed. Julie reports **that she continues to feel sad.** Guilty feelings **are described by** Julie. **'I** should **have been with my** sister, **I had no idea she was suicidal.'** Sleep **has** improved **with the use of PRN** Ambien **CR at HS.** Julie **convincingly** denies suicidal ideas **or intentions.**

Basic **Behaviors:**

Medication has been taken **regularly. She** needs **help with** ADLs. **When she** attends **activities participation is** minimal. Prn's **are used occasionally and are described as effective for her headaches. Impulsive behaviors are occurring, but less frequently.** Julie **has diminished food and** fluid intake. Julie **has not been confused. A good** night's sleep **is described.**

Additional Signs **or** Possible **Side** Effects:

Sedative effects **of the medication are described.** Patient reports **a dry** mouth. **No** other **side** effects **are** reported **or in evidence.**

MENTAL STATUS:

Julie presents **as** glum, **downcast, inattentive, minimally communicative, and** looks unhappy. **She** appears listless **and** anergic. **She** appears **downcast.** Thought content **is depressed.** Slowness **of** physical movement helps reveal **depressed** mood. Facial **expression and** general demeanor reveal **depressed** mood. **She** denies having suicidal ideas. There **are no** apparent signs **of hallucinations, delusions,** bizarre **behaviors, or** other **indicators of psychotic** process. **Associations are** intact, thinking **is** logical, **and** thought content appears **appropriate.** There **are** signs **of** anxiety. Patient **is** fidgety **in a way that is suggestive of anxiety.**

Special **Circumstances:**

Julie **continues to have an** unsteady gait, **especially** after **midnight. Call** light **is** within **her** reach. **She has been instructed to ring for the** nurse **to** assist **her when ambulating to bathroom.**

Vital Signs:

Sitting blood pressure **is 150 / 85.** Sitting pulse **rate is 80. Respiratory rate is 18 per** minute. Temp. **is 98+ F.** Weight **is 155 lbs.** (70.3 **Kg).**

Figure 2: A fake example of a *Nursing Progress Note* taken from `https://www.examples.com/business/progress-note.html`. Highlighted in **red** and **orange** are words with length in the 90-100 and 75-90th quantiles, respectively. **Green** words are of length that is below the 25th quantile. Qunatiles are measured based on the frequency of all words in the clinical note.

how to use causal inference in NLP (Keith and O'Connor 2020), and describe the current state of research on model interpretations and debiasing in NLP. Finally, we discuss our contribution in light of the relevant work.

3.1 Causal Inference and NLP

There is a rich body of work on causality and on causal inference, as it has been at the core of scientific reasoning since the writings of Plato and Aristotle (Woodward 2005). The questions that drive most researchers interested in understanding human behavior are causal in nature, not associational (Pearl et al. 2009). They require some knowledge or explicit assumptions regarding the data-generating process, such as the world model we describe in the causal graph presented in Figure 3. Generally speaking, causal questions cannot be answered using the data alone, or through the distributions that generate it (Pearl et al. 2009).

Even though causal inference is widely used in the life and social sciences, it has not had the same impact on machine learning and NLP in particular (Angrist and Pischke 2008; Dorie et al. 2019; Gentzel, Garant, and Jensen 2019). This can mostly be attributed to the fact that using existing frameworks from causal inference in NLP is challenging (Keith and O'Connor 2020). The high-dimensional nature of language does not easily fit into the current methods, specifically as the treatment whose effect is being tested is often binary (D'Amour et al. 2017; Athey et al. 2017). Recently, this seems to be changing, with substantial work being done on the intersection of causal inference and NLP (Tan, Lee, and Pang 2014; Fong and Grimmer 2016; Egami et al. 2018; Wood-Doughty, Shpitser, and Dredze 2018; Veitch, Sridhar, and Blei 2019).

Specifically, researchers have been looking into methods of measuring other confounders via text (Pennebaker, Francis, and Booth 2001; Saha et al. 2019), or using text as confounders (Johansson, Shalit, and Sontag 2016; De Choudhury et al. 2016; Roberts, Stewart, and Nielsen 2018). In this strand of work, a confounder is being retrieved from the text and used to answer a causal question, or the text itself is used as a potential confounder, with its dimensionality reduced. Another promising direction is causally-driven representation learning, where the representation of the text is designed specifically for the purposes of causal inference. This is usually done when the treatment affects the text, and the model architecture is manipulated to incorporate the treatment assignment (Roberts et al. 2014; Johansson, Shalit, and Sontag 2016; Roberts, Stewart, and Nielsen 2018). Recently, Veitch, Sridhar, and Blei (2019) added to the BERT's fine-tuning stage an objective that estimates propensity scores and conditional outcomes for the treatment and control variables, and used a model to estimate the treatment effect. As opposed to our work, they are interested in creating low-dimensional text embeddings that can be used as variables for answering causal questions, not in interpreting what affects an existing model.

While previous work from the causal inference literature used text to answer causal questions, to the best of our knowledge we are the first (except for (Vig et al. 2020)) that are using this framework for causal model explanation. Specifically, we build in this research on a specific subset of causal inference literature – counterfactual analysis (Pearl 2009). That is, we ask causal questions aimed at inferring what would have been the predictions of a given neural model had conditions been different. We present this kind of counterfactual analysis as a method for interpreting DNNs to understand what affects the decisions of the model. By intervening on the textual representation, we provide a framework for answering causal questions regarding the effect of low and high level concepts on text classifiers without having to generate counterfactual examples.

Vig et al. (2020) also suggest using ideas from causality for DNN explanations, but focus on understanding how information flows through different model components, while we are interested in understanding the effect of textual concepts on classification decisions. They are dependant on manually constructed queries, such as comparing the language model's probability for a male pronoun to that of a female, for a given masked word. As their method can only be performed by manually creating counterfactual examples such as this query, it is exposed to all the problems involving counterfactual text generation (see Section 5.1). Also, they do not compare model predictions on examples and their counterfactuals, and only measure the

difference between the two queries, neither of which are the original text. In contrast, we propose a generalized method for providing a causal explanation for any textual concept, and present datasets where any causal estimator can be tested and compared to a ground truth. We also generate a language representation which approximates counterfactuals for a given concept of interest on each example, thus allowing for a causal model explanation without having to manually create examples.

3.2 Model Interpretations and Debiasing in NLP

Model interpretability is the degree to which a human can consistently predict the model's outcome (Kim, Khanna, and Koyejo 2016; Doshi-Velez and Kim 2017). The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made. An explanation usually relates the feature values of an instance to its model prediction in a humanly understandable way, usually referred to as a *local explanation*. Alternatively, it can be comprised of an estimation of the global effect of a certain feature on the model's predictions.

There is an abundance of recent work on model explanations and interpretations, especially following the rise of DNNs in the past few years (Lundberg and Lee 2017; Ribeiro, Singh, and Guestrin 2016). Vig et al. (2020) divide interpretations in NLP into structural and behavioral methods. Structural methods try to identify the information encoded in the model's internal structure by using its representations to classify textual properties (Adi et al. 2017; Hupkes, Veldhoen, and Zuidema 2018; Conneau et al. 2018). For example, Adi et al. (2017) find that representations based on averaged word vectors encode information regarding sentence length. Behavioral methods evaluate models on specific examples that reflect an hypothesis regarding linguistic phenomena they capture (Sennrich 2017; Isabelle, Cherry, and Foster 2017; Naik et al. 2019). Sennrich (2017), for example, discover that neural machine translation systems perform transliteration better than models with byte-pair encoding (BPE) segmentation, but are worse in terms of capturing morphosyntactic agreement.

Both structural and behavioral methods generally do not offer ways to directly measure the effect of the structure of the text or the linguistic concepts it manifests on model outcomes. They often rely on token level analysis, and do not account for counterfactuals. Still, there has been very little research in NLP on incorporating tools from causal analysis into model explanations (Vig et al. 2020) (see above), something which lies at the heart of our work. Moreover, there's been, to the best of our knowledge, no work on measuring the effect of concepts on models' predictions in NLP (see Kim et al. (2018) and Goyal et al. (2019a) for a discussion in the context of computer vision).

Closely related to model interpretability, debiasing is a rising sub-field that deals with creating models and language representations that are unaffected by unwanted biases that might exist in the data (Kiritchenko and Mohammad 2018; Elazar and Goldberg 2018; Gonen and Goldberg 2019; Ravfogel et al. 2020). DNNs are as good as the training data they are fed, and can often learn associations that are in direct proportion to the distribution observed during training (Caliskan, Bryson, and Narayanan 2017). While debiasing is still an ongoing effort, there are methods for removing some of the bias encoded in models and language representations (Gonen and Goldberg 2019). Model debiasing is done through manipulation of the training data (Kaushik, Hovy, and Lipton 2019), by altering the training process (Huang et al. 2019) or by changing the model (Gehrmann et al. 2019).

Recently, Ravfogel et al. (2020) offered a method for removing bias from neural representations, by iteratively training linear classifiers and projecting the representations on their nullspaces. Their method does not provide causal model explanation, but instead reveals correlations

between certain textual features and the predictions of the model. Yet, their method could be used as an alternative to adversarial training in our framework for causal model explanation.

In our work, we present datasets where bias can be computed directly by comparing predictions on examples and their counterfactuals. Comparatively, existing work measures model bias using observational, rather than interventional measures (Rudinger, May, and Van Durme 2017; De-Arteaga et al. 2019; Ravfogel et al. 2020). To compare methods for causal model explanations, the research community would require datasets where we can intervene on specific textual features and test whether candidate methods can estimate their effect. Our work is the first to provide datasets where such comparisons are possible. Yet, in future work we plan to develop richer, more complex datasets that would allow for even more realistic counterfactual comparisons.

4. Causal Model Explanation

While usually in scientific endeavors causal inference is the main focus, we rely here on a different aspect of causality - causal model explanation. That is, we attempt to estimate the causal effect of a given variable (also known as the *treatment*) on the model's predictions, and present such effects to explain the observed behavior of the model. Here we formalize model explanation as a causal inference problem, and propose a method to do that through language representations.

We start by providing a short introduction to causal inference and its basic terminology, focusing on its application to NLP. To ground our discussion within NLP, we follow the *Adjectives* example from Section 1 and present in Figure 3 a *casual diagram*, a graph that could describe the data-generating process of that example. Building on this graph, we discuss its connection to Pearl's *structural causal model* and the *do*-operator (Pearl et al. 2009), the operator we are approximating in this paper through adversarial pre-training. Typically, causal models are built for understanding real-world outcomes, while model interpretability efforts deal with the case where the classification decision is the outcome, and the intervention is on a feature present in the model's input. As we are the first, to the best of our knowledge, to propose a comprehensive causal framework for model interpretations in NLP, we link the existing literature on both.

4.1 Causal Inference and Language Representations

Confounding Factors and the do-operator. Continuing with the first example from Section 1 (presented in Figure 1), imagine we observe a text X and have trained a model to classify each example as either positive or negative, corresponding to the conveyed sentiment. We also have information regarding the *Political Figure* discussed in the text, and tags for the parts of speech in it. Given a set of concepts, which we hypothesize might affect the classification decision, we denote the set of binary variables $C = \{C_j \in \{0,1\} | j \in \{0,1,\ldots,k\}\}$, where each variable corresponds to the existence of a predefined concept in the text, i.e. if $C_j = 1$ then the j-th concept appears in the text. We further assume a pre-trained language representation model ϕ (such as BERT), and wish to assert how our trained classifier f is affected by the concepts in C, where f is a classifier that takes $\phi(X)$ as input and outputs a class $l \in L$. As we are interested in the effect on the probability assigned to each class by the classifier f, we measure the class probability of our output for an example X, and denote it for a class $l \in L$ as z_l. When computing differences on all L classes, we use $\vec{z}(f(\phi(X)))$, the vector of all z_l probabilities.

Computing the effect of a concept C_j on $\vec{z}(f(\phi(X)))$ seems like an easy problem. We can simply feed to our model examples with and without the chosen concepts, and compute the difference between the average $\vec{z}(\cdot)$ in both cases. For example, if our concept of interest is positive *Adjectives*, we can feed the model with examples that include positive *Adjectives* and
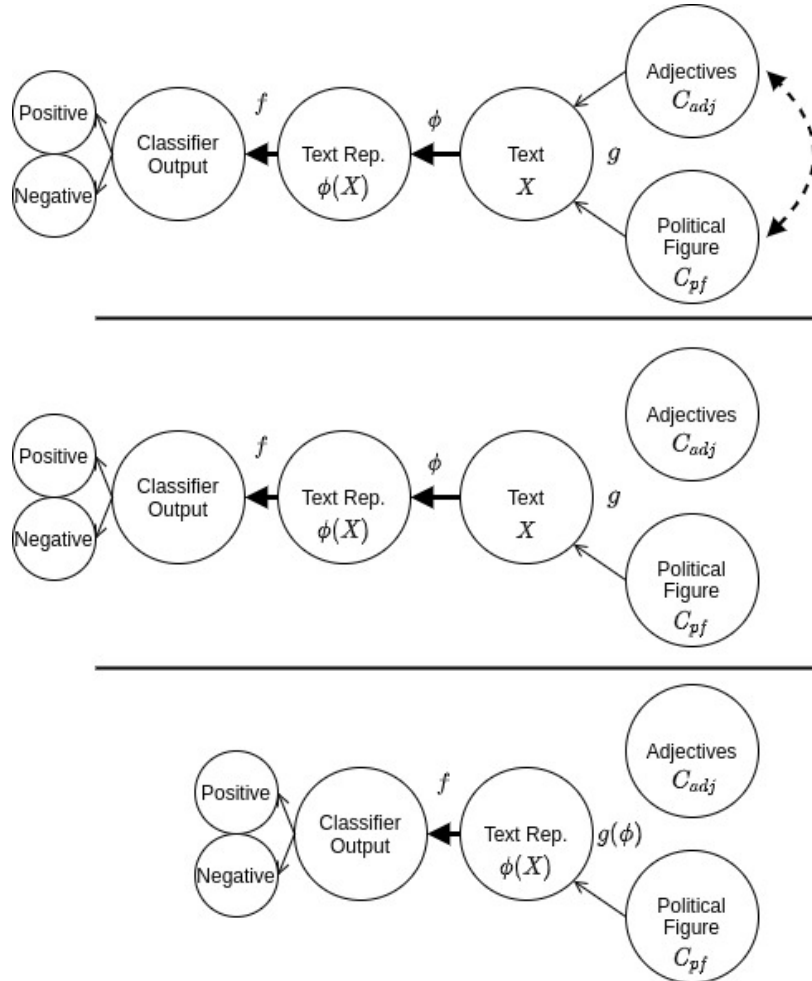
Figure 3: Three causal graphs relating the concepts of *Adjectives* and *Political Figure*, texts, their representations and classifier output. The top graph describes the original data generating process. The middle graph describes the case of directly manipulating the text. In this case, after performing the *do*-operator the *Adjectives* do not affect the textual representation. The bottom graph describes our approach, where we manipulate the representation mechanism and not the actual text. The dashed edge indicates a possible hidden confounder of the two concepts.

examples that do not. Then, we can compare the difference between the averaged $\vec{z}(\cdot)$ in both sets and conclude that this difference is the effect of positive *Adjectives*.

Now, imagine the case where the use of positive and negative *Adjectives* is associated with the *Political Figure* that is being discussed in the texts given to the model. An obvious example is a case where a political commentator with liberal-leaning opinions is writing about a conservative politician, or vice-versa. In that case, it would be reasonable to assume that the *Political Figure* being discussed would affect the text through other concepts besides its identity. The author can then choose to express her opinion through *Adjectives* or in other ways, and these might be correlated. In such cases, comparing examples with and without positive *Adjectives* would result in an inaccurate measurement of their effect on the classification decisions of the model.

The problem with our correlated concepts is that of *confounding*. It is illustrated in the top graph of Figure 3 using the example of *Political Figure* and *Adjectives*. In causal inference, a *confounder* is a variable that affects other variables and the predicted label. In our case, the *Political Figure* (C_{pf}) being discussed in the texts is a confounder of the *Adjectives* concept, as it directly affects both C_{adj} and $f(\phi(X))$. As can be seen in this figure, we can think of texts as originating from a list of concepts. While we plot only two, *Adjectives* and *Political Figure*, there could be many concepts generating a text. We denote the potential confoundedness of the concepts by dashed arrows, to represent that one could affect the other.

Alternatively, if it was the case that a change of the *Political Figure* would not affect the usage of *Adjectives* in the text, we could have said that C_{adj} and C_{pf} are not confounded. In the middle plot of Figure 3, we imagine an alternative data-generating process, where we can intervene and change the *Adjectives* in a way that does not affect the *Political Figure* and the textual representation. This is the case where intervening on C_{adj}, such as by having the author write a text without using positive *Adjectives*, would not induce a text that contains a different *Political Figure*. In causal terms, this is the case where:

$$\vec{z}(f(\phi(X)|do(C_{adj}))) = \vec{z}(f(\phi(X)|C_{adj})) \tag{1}$$

Where $do(C_{adj})$ stands for an external intervention that compels the change of C_{adj}. In contrast, the class probability distribution $\vec{z}(f(\phi(X)|C_{adj}))$ represents the distribution resulting from a passive observation of C_{adj}, and rarely coincides with $\vec{z}(f(\phi(X)|do(C_{adj})))$. Indeed, not using positive *Adjectives* decreases the probability assigned by the classifier to a positive sentiment, but it does not necessarily "cause" it.

Counterfactual Text Representations. The act of manipulating the text to change the *Political Figure* in focus or the *Adjectives* used in the text is derived from the notion of *counterfactuals*. In the *Adjectives* example (presented in Figure 1), a counterfactual text is such an instance where we intervene on one concept only, holding everything else equal. It is the equivalent of imagining what could have been the text, had it been written about a different *Political Figure*, or about the same *Political Figure* but with different *Adjectives*.

In the case of *Adjectives*, we can simply detect all of them in the text and change them to a random alternative, or delete them altogether. [3] For the concept highlighting the *Political Figure* being discussed this is much harder to do manually, as the chosen figure induces the topics being described in the text and is hence likely to affect other important concepts that generate the text. Implementing the *do*-operator on *Adjectives* as presented in the middle graph of Figure 3 relies on our ability to create a conditional generative model, one that makes sure a certain concept is or is not represented in the text. Since this is often hard to do (see Section 5.1), we propose a solution that is based on the language representation $\phi(X)$. As shown in the bottom causal graph of Figure 3, we assume that the concepts generate the representation $\phi(X)$ directly. This approximation shares some similarities with the idea of *Process Control* described in Pearl et al. (2009). While Pearl presents *Process Control* as the case of intervening on the process affected by the treatment, it is not discussed in relation to language representations or model interpretations. Interventions on the process that is generating the outcomes are also discussed in Chapter 4 of Bottou et al. (2013), in the context of multi-armed bandits and reinforcement learning.

By intervening on the language representation, we attempt to bypass the process of generating a text given that a certain concept should or should not be represented in that text. We

---

3 This would still require the modeler to control some confounding concepts, as *Adjectives* could be correlated with other variables (such as some *Adjectives* used to describe a specific politician).

take advantage of the fact that modern NLP systems use pre-training to produce a language representation, and generate a counterfactual language representation $\phi_{cf}(X)$ that is unaffected by the existence of a chosen concept. That is, we try to change the language representation such that we get for a binary C_k:

$$\vec{z}(f(\phi_{cf_k}(X))) = \vec{z}(f(\phi_{cf_k}(X'))) \qquad (2)$$

Where X and X' are identical for every generating concept, except for the concept C_k, on which they might or might not differ. In Section 5, we discuss how we intervene in the fine-tuning stage of the language representation model (BERT in our case) to produce the counterfactual representation using an adversarial component.

We now formally define the *causal concept effect* (CaCE), first introduced in Goyal et al. (2019a) in the context of computer vision. The process required to calculate CaCE is presented in the middle graph of Figure 3, and requires a conditional generative model. As presented in the bottom graph of Figure 3, we use an adversarial method, inspired by the idea of *Process Control*, first introduced by Pearl (2009), to approximate the *do*-operator through the text representation. We then define the *Textual Representation-based Average Treatment Effect* (TReATE), which is estimated using our method, and compare it to the standard *Average Treatment Effect* (ATE) estimator from the causal literature.

4.2 The Textual Representation-based Average Treatment Effect (TReATE)

When estimating causal effects, researchers commonly measure the *average treatment effect*, which is the difference in mean outcomes between the treatment and control groups. Using *do*-calculus (Pearl 1995), we can define it in the following way:

Definition 1 (Average Treatment Effect (ATE))
The average treatment effect of a binary treatment T on the outcome Y is:

$$\text{ATE}_T = \mathbb{E}\left[Y|do(T=1)\right] - \mathbb{E}\left[Y|do(T=0)\right] \qquad (3)$$

Following the notations presented in the beginning of Section 4.1, we define the following Structural Causal Model (SCM, Pearl (2009)) for a document X:

$$(C_0, C_1, \ldots, C_k) = h(\epsilon_C)$$
$$\phi(X) = g(C_0, C_1, \ldots, C_k, \epsilon_X)$$
$$C_j \in \{0, 1\}, \forall j \in K$$

$$\qquad (4)$$

Where, as is standard in SCMs, ϵ_C and ϵ_X are independent variables. The function h is the generating process of the concept variables from the random variable ϵ_C and is not the focus here. The SCM in Equation (4) makes an important assumption, namely that it is possible to intervene atomically on C_j, the *treated concept* (TC), while leaving all other concepts untouched. We denote expectations under the interventional distribution by the standard *do*-operator notation $\mathbb{E}_g\left[\cdot|do(C_j = a)\right]$, where the subscript g indicates that this expectation also depends on the choice of the generative process g. We can now use these expectations to define *CaCE*:

Definition 2 (Causal Concept Effect (CaCE) (Goyal et al. 2019a))
The causal effect of a concept C_j on the class probability distribution $\vec{z}$ of the classifier f under the generative process $g(X)$ is:

$$\text{CaCE}_{C_j} = \langle \mathbb{E}_g \left[\vec{z}\big(f(\phi(X))\big)|do(C_j = 1)\right] - \mathbb{E}_g \left[\vec{z}\big(f(\phi(X))\big)|do(C_j = 0)\right]\rangle \qquad (5)$$

Where the operator $\langle \rangle$ denotes a summation over the absolute values of vector coordinates.[4]

Implementing *CaCE* requires counterfactual example generation, as shown in the middle graph of Figure 3. As this is often intractable in NLP (see Section 5.1), we do not compute *CaCE* here. We instead generate a counterfactual language representation, a process which is inspired by the idea of *Process Control* introduced by Pearl (2009) for dynamic planning. This is the case where we can only control the process generating $\phi(X)$ and not X itself. Here, we follow the bottom graph from Figure 3, and intervene only on the language representation $\phi(X)$. Following this intervention, we compute the *Textual Representation-based Average Treatment Effect* (*TReATE*).

Definition 3 (Textual Representation-based Average Treatment Effect (TReATE))
The causal effect of a concept C_j on the class probability distribution $\vec{z}$ of the classifier f under the generative process $g(\phi(X))$ is:

$$\text{TReATE}_{C_j} = \langle \mathbb{E}_g \left[\vec{z}\big(f(\phi(X))|do(\phi(X) = \phi(X|C_j = 0))\big)\right] \qquad (6)$$
$$- \mathbb{E}_g \left[\vec{z}\big(f(\phi(X))\big)\right]\rangle$$

In the case where we would also like to control that the concept C_k is represented in the text, we compute the following *TReATE*:

$$\text{TReATE}_{C_j, C_k} = \langle \mathbb{E}_g \left[\vec{z}\big(f(\phi(X))|do(\phi(X) = \phi(X|C_j = 0, C_k = 1))\big)\right] \qquad (7)$$
$$- \mathbb{E}_g \left[\vec{z}\big(f(\phi(X))\big)\right]\rangle$$

In our framework, we would like to use the tools defined here to measure the casual effect of one or more concepts $\{C_0, C_1, \cdots, C_k\}$ on the predictions of the classifier f. We will do that by measuring *TReATE*, which is a special case of the *average treatment effect* (ATE) defined in Equation 3, where the intervention is performed via the textual representation. While *ATE* is usually used to compute the effect of interventions in randomized experiments, here we use *TReATE* to explain the predictions of a text classification model in terms of concepts.

5. Representation-Based Counterfactual Generation

In this section we discuss the reason we choose to approximate the *do*-operator through the language representation mechanism, as an alternative to directly implementing this operator through synthetic example generation. We present two existing approaches for generating such synthetic examples and explain why they are often implausible in NLP. We then introduce our approach, an intervention on the language representation, designed to ignore a particular set of concepts while preserving the information from another set of concepts. Finally, we describe how to approximate the *do*-operator using this counterfactual language representation

---

4 For example, for a three class prediction problem, where the model's probability class distribution for the original example is $(0.7, 0.2, 0.1)$, while for the counterfactual example it is $(0.5, 0.1, 0.4)$, $CaCE_{C_j}$ is equal to:
$|0.7 - 0.5| + |0.2 - 0.1| + |0.1 - 0.4| = 0.2 + 0.1 + 0.3 = 0.6$.

5.1 Generating Synthetic Examples

Comparing model predictions on examples to the predictions on their counterfactuals is what allows the estimation of causal explanations. Without producing a version of the example that does not contain the treatment (i.e concept or feature of interest), it would be hard to ascertain whether the classifier is using the treatment or other correlated information (Kaushik, Hovy, and Lipton 2019). To the best of our knowledge, there are two existing methods for generating counterfactual examples: manual augmentation and automatic generation using generative models.

Manual augmentation can be straight-forward, as one needs to manually change every example of interest to reflect the absence or presence of a concept of choice. For example, when measuring the effect of *Adjectives* on a sentiment classifier, a manual augmentation could include changing all positive *Adjectives* into negative ones, or simply deleting all *Adjectives*. While such manipulations can sometime be easily done with human annotators, they are costly and time consuming and therefore implausible for large datasets. Also, in cases such as the clinical note example presented in Figure 2, it would be hard to manipulate the text such that it uses a different writing style, making it even harder to manually create the counterfactual text.

Using generative models has been recently discussed in the case of images (Goyal et al. 2019a). In this paper, Goyal et al. propose using a conditional generative model, such as a conditional VAE (Lorberbom et al. 2019), to create counterfactual examples. While in some cases, such as those presented in their paper, it might be plausible to generate counterfactual examples, in most cases in NLP it is still too hard to generate realistic texts with conditional generative models (Lin et al. 2017; Che et al. 2017; Rajeswar et al. 2017; Guo et al. 2018). Also, for generating local explanations it is required to produce a counterfactual for each example such that all the information besides the concept of choice is preserved, something that is even harder than producing two synthetic examples, one from each concept class, and comparing them.

As an alternative to manipulating the actual text, we propose to intervene on the language representation. This does not require generating more examples, and therefore does not depend on the quality of the generation process. The fundamental premise of our method is that comparing the original representation of an example to this counterfactual representation is a good approximation of comparing an example to that of a synthetic counterfactual example that was properly manipulated to ignore the concept of interest.

5.2 The *Do*-Operator in Language Representation Models

Since the introduction of pre-trained word-embeddings, there have been an explosion of research on choosing pre-training tasks and understanding their effect (Jernite, Bowman, and Sontag 2017; Logeswaran and Lee 2018; Ziser and Reichart 2018; Dong et al. 2019; Chang et al. 2019; Sun et al. 2019; Rotman and Reichart 2019). The goal of this process is to generate a representation that captures valuable information for solving downstream tasks, such as sentiment classification, entity recognition and parsing. Recently, there has also been a shift in focus towards pre-training contextual language representations (Liu et al. 2019; Yang et al. 2019).

Contextual embedding models typically follow three stages: **(1)** Pre-training: Where a DNN (encoder) is trained on a massive unlabeled dataset to solve self-supervised tasks; **(2)** Fine-tuning: An optional step, where the encoder is further trained on different tasks or data; and **(3)** Supervised task training: Where task specific layers are trained on labeled data for a downstream task of interest.

Our intervention is focused on Stage 2. In this stage, we continue training the encoder of the model on the tasks it was pre-trained on, but add auxiliary tasks, designed to forget some concepts and remember others. In Figure 4 we present an example of our proposed Stage 2, where we train our model to solve the original BERT's *Masked Language Model* (MLM) and

Next Sentence Prediction (NSP) tasks, along with a *Treated Concept* objective, denoted in the figure as *TC*. In order to preserve the information regarding a potentially confounding concept, we use an additional task denoted in the figure as *CC*, for *Controlled Concept*.
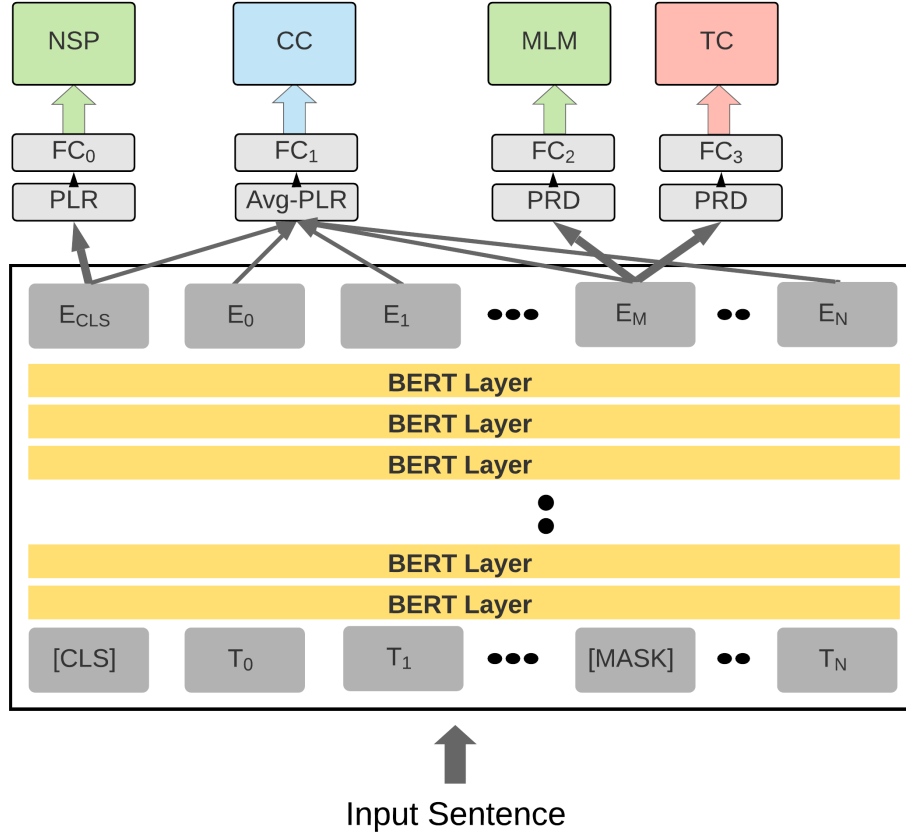

Figure 4: An illustration of our Stage 2 fine-tuning procedure for our counterfactual representation model (*BERT-CF*). In this representative case, we add a task, named *Treated Concept* (TC), which is trained adversarially. This task is designed to "forget" the effect of the treated concept, as in the IMA adversarial task discussed in Section 7. To control for a potential confounding concept (i.e. to "remember" it), we add the *Control Concept* (CC) task, which predicts the presence of this concept in the text, as in the PF task discussed below. PRD and PLR stand for the BERT prediction head and the pooler head respectively, $AVG - PLR$ for an average pooler head, FC is a fully connected layer, and $[MASK]$ stands for masked tokens embeddings. NSP and MLM are the BERT's next prediction and masked language model objectives. The results of this training stage is our counterfactual *BERT-CF* model.

To illustrate our intervention, we can revisit the *Adjectives* example, introduced in Figure 1, and consider a case where we want to test whether their existence in the text affects the classification decision. To be able to estimate this effect, we traditionally would have to produce for each example in the test-set an equivalent example that does not contain *Adjectives*. In terms of our intervention on the language representation, we should be able to produce a representation that is unaffected by the existence of *Adjectives*, meaning that the representation of a sentence

that contains *Adjectives* would be identical to that of the same sentence where *Adjectives* are excluded. Taking that to the fine-tuning stage, we could use adversarial training to "forget" *Adjectives*.

Concretely, we add to BERT's loss function a negative term for the target concept and a positive term for each control concept we consider. As shown in Equation 8, in the case of the example from Figure 1, this would entail augmenting the loss function with two terms: adding the loss for the *Political Figure* classification PF (the *CC* head), and subtracting that of the *Is Masked Adjective* (IMA) task (the *TC* head). As we are using the IMA objective term in our *Adjectives* experiments (Section 7), and not only in the running example, we describe the task below. For the *Political Figure* (PF) concept, we could simply use a classification task where for each example we predict the political orientation of the politician being discussed. With those tasks added to the loss function, we have that:

$$\mathcal{L}(\theta_{bert}, \theta_{mlm}, \theta_{nsp}, \theta_{tc}, \theta_{cc}) = \tag{8}$$

$$\frac{1}{n} \Big(\sum_{i=1}^{n} \mathcal{L}_{mlm}^{i}(\theta_{bert}, \theta_{mlm})$$

$$+ \sum_{i=1}^{n} \mathcal{L}_{nsp}^{i}(\theta_{bert}, \theta_{nsp})$$

$$+ \sum_{i=1}^{n} \mathcal{L}_{cc}^{i}(\theta_{bert}, \theta_{cc})$$

$$- \lambda \sum_{i=1}^{n} \mathcal{L}_{tc}^{i}(\theta_{bert}, \theta_{tc}) \Big)$$

Where θ_{bert} denotes all of BERT's parameters, except those devoted to $\theta_{mlm}, \theta_{nsp}, \theta_{tc}$ and θ_{cc}. λ is a hyper-parameter which controls the relative weight of the adversarial task, as discussed in Ganin et al. (2016).

One way of implementing the IMA *TC* head is inspired by the BERT's MLM head. That is, masking *Adjectives* and *Non-adjectives*, then predicting whether the masked token is an adjective. Following the *gradient reversal* method presented in Ganin et al. (2016), [5] we add this task with a layer which leaves the input unchanged during forward propagation, yet reverses its corresponding gradient by multiplying it with a negative scalar during back propagation. *Gradient reversal* ensures that the features over the two text types (with/without *Adjectives*) are made similar (as indistinguishable as possible for the IMA classifier), thus resulting in an adjective-invariant representation. By optimizing this objective, the parameters of the underlying language representation are simultaneously optimized in order to minimize the MLM loss and maximize the IMA loss, encouraging adjective-invariant features to emerge. For the *CC* objective, we can add any of the classification tasks suggested above for PF (*CC*), following the definition of the world model (i.e. the causal graph) the researcher is assuming.

Having optimized the loss functions presented in Equation 8, we can now use the resulting counterfactual representation model and compute the *individual treatment effect* (ITE) on an example as follows. We compute the predictions of two different models: One that employs the original BERT, that has not gone through our counterfactual fine-tuning, and one that employs

---

5 See equation $9 - 10$ and $13 - 15$ in Ganin et al. (2016).

the counterfactual BERT model (BERT-CF). *The Textual Representation-based ITE* (TRITE) is then the average of the absolute differences between the probabilities assigned to the possible classes by these models:[6]

$$TR\hat{I}TE_{TC}^{i} = \langle \vec{z}\big(f(\phi_{TC=0}(X = x_i))\big) - \vec{z}\big(f(\phi(X = x_i))\big) \rangle \tag{9}$$

Where x_i is the specific example, ϕ is the original language representation model and $\phi_{C_{TC}=0}$ is the counterfactual *BERT-CF* representation model, where the intervention is such that *TC* has no effect. $\vec{z}\big(f(\phi(X))\big)$ is the class probability distribution of the classifier f when using ϕ as the representation model for example X. As *TReATE* is presented in Equation 6 in expectation form, we compute our estimated $TRe\hat{A}TE$ by summing over $TR\hat{I}TE$ for the set of all test-set examples, I:

$$TRe\hat{A}TE_{TC} = \frac{1}{|I|} \sum_{i \in I} \langle \vec{z}\big(f(\phi_{TC=0}(X = x_i))\big) - \vec{z}\big(f(\phi(X = x_i))\big) \rangle \tag{10}$$

5.3 Alternative Causal Graphs and Limitations

Our ability to intervene on the *treated concept* and estimate $TRe\hat{A}TE_{TC}$ is dependent on the world model we assume, as presented in the causal graph. For the examples presented in Figures 1 and 2 we have suggested causal graphs (Figure 3) where the relationship between the concepts generating the text is rather simple, as all concepts generate the text without any inheritance relations (i.e where one concept causes the other). In many interesting cases, the relationship between concepts is not as straight-forward, and might affect our ability to intervene on some concepts. In Figure 5, we consider two such cases, where one or more concepts $A_j | j \in \{0, 1, \ldots, k\}$ cause a concept B, which in turn generates a text X. In the clinical note example, it could be the case that a patient's condition is causing a doctor to recommend a specific treatment, and this decision induces the doctor to write the note. In Figure 5 we propose two causal graphs that model this data-generating process, for the case where many conditions cause the doctor's decision (top graph) and for the case where only the patient's depression affects the decision (bottom graph).

Intervening on specific concepts and computing the causal effect of those concepts on the decisions made by the classifier is not as straight-forward. In the first case, where *Anxiety*, *Depression* and *Suicidal* are causing the doctor's *Clinical Recommendation*, intervening on *Depression* would also affect the *Clinical Recommendation*, as it is one of its causes. While our method can accommodate for this case, our estimator will measure both the direct effect of *Depression* and its indirect effect, through the *Clinical Recommendation*. In such a case it would be impossible to preserve all the information resulting from the *Clinical Recommendation* concept while omitting the information from the *Depression* concept, due to the inheritance relation between the two concepts. Intervening on the *Clinical Recommendation* will be even more problematic, as a text without a recommendation will have to be blank according to this graph, regardless of the patient's underlying condition.

The causal graph presented in the bottom of Figure 5 is also possible, meaning that it could be a reasonable world model in some cases. However, if we were to intervene on *Depression*, we will not be able to know what the doctor would have recommended. In this case we could not estimate the causal effect of *Depression* or that of the *Clinical Recommendation* on the classifier's

---

6 In order to avoid multiple similar equations, we do not explicitly write the equation for the case where we also control for a CC concept.
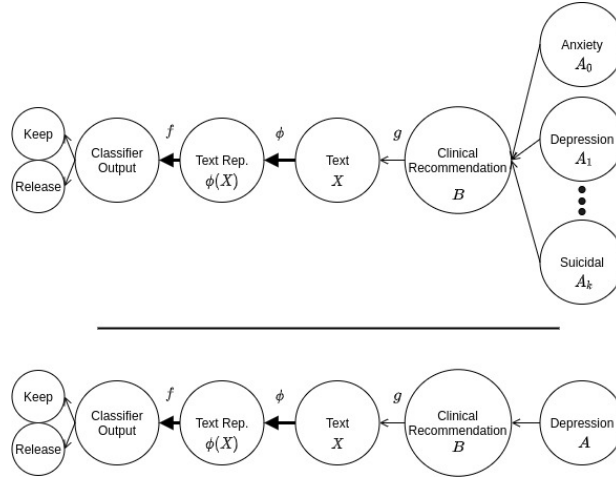
Figure 5: Two plausible causal graphs for a case where a patient's condition (*Anxiety*, *Depression* and *Suicidal*) is causing the doctor's *Clinical Recommendation*, which is then generating a text. The classifier then uses this text to decide if the patient should be kept in the hospital (*Keep*) or released (*Release*). The top graph represents a data-generating process where many conditions cause the doctor's *Clinical Recommendation*, such as *Anxiety*, *Depression* and *Suicidal*. The bottom graph represents the scenario where only *Depression* causes the *Clinical Recommendation*.

decision. In Section 8.2 we suggest several sanity checks to help modelers understand if the world model they are using has successfully learned to forget the *treated concept* while remembering the *control concepts*, but there are certainly cases where the ability to perform such interventions is limited.

6. Data

When evaluating a trained classification model, we usually have access to a test-set, consisting of manually labeled examples that the model was not trained on, and can hence be used for evaluation. Estimating causal effects is often harder in comparison, as we do not have access to the ground truth. In the case of causal inference, we can generally only identify effects if our assumptions on the data-generating process, such as those presented in Figure 3, hold. This means that at the core of our causal model explanation paradigm is the availability of a causal graph that encodes our assumptions about the world. Notice, however, that non-causal explanation methods that do not make assumptions about the world are prone to finding arbitrary correlations, a problem that we are aiming to avoid with our method.

To allow for ground-truth comparisons and to spur further research on causal inference in NLP, we propose here four cases where causal effects can be estimated. In three out of those cases, we have constructed datasets with counterfactual examples so that the causal estimators can be compared to the ground truth. We start here by introducing the datasets we created and discuss the choices made in order to allow for proper evaluation. In Section 7 we present the tasks for which we estimate the causal effect on and the experiments we conduct, all using these datasets. [7]

---

[7] Our datasets are available at: `https://www.kaggle.com/amirfeder/causalm`.

6.1 Product and Movie Reviews

Following the running example of Section 2, we start by looking for prominent sentiment classification datasets. Specifically, we look for datasets where the domain entails a rich description where *Adjectives* could play a vital role. With enough variation in the structure and length of examples, we hope that *Adjectives* would have a significant effect. Another key aspect is the number of training examples. To be able to amplify the correlation between the treated concept (*Adjectives*) and the label, we need to be able to omit some training examples. For instance, if we omit most of the positive texts describing a *Political Figure*, we can create a correlation between the negative label and that politician. We need a dataset that will allow us to do that and still have enough training data to properly train modern DNN classification models.

Another concept we wish to estimate its causal effect on sentiment classification is *Topics* (see Section 7 for an explanation on how we compute the topic distribution). To be able to observe the causal effect of *Topics*, some variation is required in the *Topics* discussed in the texts. For that, we use data originating from several different domains, where different, unrelated products or movies are being discussed. In this section we focus on the description of the dataset we have generated, and explain how we manipulate the data in order to generate various degrees of concept-label correlations.

Considering these requirements and the concepts for which we wish to estimate the causal effect on model performance, we choose to combine two datasets, spanning five domains. The product dataset we choose is widely used in the NLP domain adaptation literature, and is taken from Blitzer, Dredze, and Pereira (2007). It contains four different domains: *Books*, *DVD*, *Electronics* and *Kitchen Appliances*. The movie dataset is the IMDB movie review dataset, taken from Maas et al. (2011). In both datasets, each example consists of a review and a rating (0-5 stars). Reviews with $rating > 3$ were labeled positive, those with $rating < 3$ were labeled negative, and the rest were discarded because their polarity was ambiguous. The product dataset is comprised of $1,000$ positive and $1,000$ negative examples for each of the four domains, for a total of $4,000$ positive and $4,000$ negative reviews. The *Movies* dataset is comprised of $25,000$ negative and $25,000$ positive reviews. To construct our combined dataset, we randomly sample $1,000$ positive and $1,000$ negative reviews from the *Movies* dataset and add these alongside the product dataset reviews. Our final combined dataset amounts to a total of $10,000$ reviews, balanced across all five domains and both labels.

We tag all examples in both datasets for the Part-of-Speech (*PoS*) of each word with the automatic tagger available through *spaCy*,[8] and use the predicted labels as ground truth. For each example in the combined dataset, we generate a counterfactual example for *Adjectives*. That is, for each example we create another instance where we delete all words that are tagged as *Adjectives*, such that for the example: "*It's a lovely table*", the counterfactual example will be: "*It's a table*". Finally, we count the number of *Adjectives* and other *PoS* tags, and create a variable indicating the ratio of *Adjectives* to *Non-adjectives* in each example, which we use in Section 7 to bias the data.

For the *Topic* concepts, we train an LDA topic model (Blei, Ng, and Jordan 2003)[9] on all the data in our combined dataset and optimize the number of topics for maximal coherence (Lau, Newman, and Baldwin 2014), resulting in a set of $T = 50$ topics. We then search for the topic that is relatively most associated with each of the five domains, i.e. the topic with the largest difference between the probability assigned to examples from that domain and the probability

---

[8] https://spacy.io/
[9] Using the *gensim* library (Řehůřek and Sojka 2010).

assigned to examples outside of that domain, using the following equation:

$$t_{TC}(d) = \arg\max_{t \in T} \Big(\frac{1}{|I_{d_+}|} \sum_{i \in I_{d_+}} \theta_t^i - \frac{1}{|I_{d_-}|} \sum_{i \in I_{d_-}} \theta_t^i \Big) \qquad (11)$$

Where d is the domain of choice, t is a topic from the list of topics T, θ_t is the probability of topic t and I_{domain} is the set of examples for a given domain. I_{d+} is the set of examples in domain d, and I_{d-} the set of examples outside of domain d. After choosing t_{TC}, we remove the chosen topic from T and use the same calculation to the choose t_{CC}, our *control concept*.

For each *Topic*, we also compute the median probability on all examples, and define a binary variable indicating for each example whether the *Topic* probability is above or below its median. This binary variable can then be used for the *TC* and *CC* tasks described in Section 7.

In Table 1 we present some descriptive statistics for all five domains, including the *Adjectives* to *Non-adjectives* ratio and the median probability (θ_{domain}) of the $t_{TC}(d)$ topic for each domain. As can be seen in this table, there is a significant number of *Adjectives* in each domain, but the variance in their number is substantial. Also, *Topics* are domain specific, with the most correlated topic $t_{TC}(d)$ for each domain being substantially more visible in its domain compared with the others. In Table 2 we provide the top words for all *Topics*, to show how they capture domain specific information.

Domain	Min. $r(adj)$	Med. # $r(adj)$	Max. # $r(adj)$	σ of # $r(adj)$	θ_b	θ_d	θ_e	θ_k	θ_m
Books	0.0	0.135	0.444	0.042	0.311	0.014	0.052	0.052	0.014
DVD	0.0	0.138	0.425	0.042	0.014	0.225	0.045	0.045	0.225
Electronics	0.0	0.136	0.461	0.049	0.01	0.003	0.08	0.08	0.003
Kitchen	0.0	0.142	0.5	0.052	0.007	0.002	0.075	0.075	0.002
Movies	0.0	0.138	0.666	0.0333	0.01	0.281	0.045	0.045	0.281

Table 1: Descriptive statistics for the Sentiment Classification datasets. $r(adj)$ denotes the ratio of *Adjectives* to *Non-adjectives* in an example. θ_{domain} is the median probability of the topic that is most observed in that domain which will also serve as our *treated topic*. b, d, e, k, m are abbreviations for Books, DVD, Electronics, Kitchen and Movies.

Our sentiment classification data allows for a natural setting for testing our methods and hypotheses, but it has some limitations. Specifically, in the case of *Topics*, we cannot generate realistic counterfactual examples and therefore compute ATE_{gt}, the ground-truth estimator of the causal effect. This is because creating counterfactual examples would require deleting the topic from the text without affecting the grammaticality of the text, something which cannot be done automatically. In the case of *Adjectives*, we are hoping that removing *Adjectives* will not affect the grammaticality of the original text, but are aware that this sometimes might not be the case. While this data provides a real-world example of natural language, it is hard to automatically generate counterfactuals for it. To allow for a more accurate estimation of the ground truth effect, we would need a dataset where we can control the data-generating process.

6.2 The Enriched Equity Evaluation Corpus (EEEC)

Understanding and reducing gender and racial bias encapsulated in classifiers is a core task in the growing literature of interpretability and debiasing in NLP (see Section 3). There is an ongoing effort to both detect such bias and to mitigate its effect, which we see from a causal perspective as

#	Top 10 Words								
1	set	box	wait	20	making	flat	worth	longer	disappoint	spend
2	pan	phone	computer	work	does	use	still	non	battery	problem
3	great	use	just	months	problem	bought	time	good	years	ago
4	classic	stories	know	great	really	book	definitely	reading	writing	long
5	item	dull	returned	expect	given	fit	did	ridiculous	run	matter
6	kids	crap	turned	fun	children	making	point	needs	understand	truly
7	dvd	version	video	player	original	screen	release	quality	features	cover
8	book	real	second	school	author	going	page	shows	past	light
9	machine	software	uses	issues	using	help	problems	makes	device	bought
10	mind	fine	despite	pages	author	lost	books	book	read	especially
11	book	reading	information	read	quot	books	better	author	know	does
12	just	did	know	ll	does	ve	think	got	times	work
13	product	buy	amazon	bought	plastic	did	reviews	cheap	work	ve
14	does	man	just	woman	story	women	way	stop	time	like
15	expected	star	series	rest	terrible	simply	pretty	watching	paid	wait
16	away	water	stay	model	dog	good	difficult	like	right	just
17	broke	replacement	warranty	month	send	weeks	days	called	week	product
18	people	god	says	mr	life	like	world	person	american	way
19	return	garbage	single	different	unless	given	oh	hot	plastic	thought
20	play	does	power	light	white	little	used	make	drive	large
21	bad	good	pretty	really	ve	just	worst	seen	10	best
22	movie	film	like	movies	acting	bad	watch	just	plot	scenes
23	fan	wrote	fans	years	special	true	humor	day	disappoint	novel
24	order	received	monster	performance	ordered	sent	said	better	later	returned
25	book	long	ll	just	tell	totally	later	reader	given	great
26	book	job	person	poor	read	kept	thought	trying	boring	good
27	new	piece	tried	stopped	junk	worked	working	work	brand	maybe
28	line	john	coming	certainly	early	true	films	enjoy	like	write
29	book	read	books	author	pages	novel	writing	reader	history	interesting
30	killer	card	camera	car	shows	stupid	series	tv	picture	better
31	coffee	mouse	stand	products	use	like	make	decided	finally	tried
32	john	writing	movie	book	waste	time	plot	make	did	line
33	quot	written	self	does	things	view	needs	like	new	hope
34	book	let	good	make	did	interesting	does	say	self	great
35	unit	device	purchased	features	works	house	returned	running	warranty	hear
36	does	hand	nice	need	small	clean	time	sex	look	things
37	quality	poor	daughter	cable	low	design	control	sound	bad	good
38	boring	long	time	end	story	rest	stop	slow	minutes	good
39	old	year	horrible	great	got	food	beautiful	boy	said	instead
40	hard	happy	sure	disappoint	writing	music	bad	reviews	days	uses
41	known	christian	truth	like	feel	store	novel	remember	stay	able
42	mouse	design	15	agree	purchased	given	job	happened	order	making
43	world	war	words	self	old	word	attempt	needed	title	life
44	lost	christian	guys	despite	turn	getting	mind	decent	war	fine
45	music	ipod	weak	car	30	battery	playing	takes	able	major
46	like	just	really	did	characters	story	character	love	little	make
47	money	waste	time	save	thought	worth	spend	better	good	just
48	disappointed	feel	fast	little	bit	good	job	parts	matter	complete
49	day	black	sound	hours	like	just	minutes	bread	went	getting
50	service	support	customer	told	product	check	company	called	terrible	hold

Table 2: Top 10 words in each of the 50 topics. A topic model was trained on all texts in all domains combined. Topic #22, our θ_m, is highlighted in red, topic #38, θ_b, is highlighted in blue, topic #8, θ_d, is highlighted in green, topic #2, θ_e, is highlighted in orange and topic #13, θ_k, is highlighted in purple.

a call for action. By offering a way to estimate the causal effect of *Gender* and *Race* on classifiers, we enable researchers a way to avoid using classifiers with unwanted bias.

In order to evaluate the quality of our causal effect estimation method, we need a dataset where we can control test examples such that for each text we have a counterfactual text that differs only by the *Gender* or *Race* of the person it discusses. We also need to be able to control the data-generating process in the training set, so that we can create such a bias for the model to pick up. A dataset that offers such control exists, and is called the Equity Evaluation Corpus (EEC) (Kiritchenko and Mohammad 2018).

It is a benchmark dataset, designed for examining inappropriate biases in system predictions, and it consists of 8,640 English sentences chosen to tease out *Racial* and *Gender* related bias. Each sentence is labeled for the mood state it conveys, a task also known as *Profile of Mood States* (POMS). Each of the sentences in the dataset is comprised using one of eleven templates, with placeholders for a person's name and the emotion it conveys. For example, one of the original

templates is: *"<Person> feels <emotional state word>."*. The name placeholder (*<Person>*) is then filled using a pre-existing list of common names that are tagged as male or female, and as African-american or European. The emotion placeholder (*<emotional state word>*) is filled using lists of words, each list corresponding to one of four possible mood states: *Anger*, *Sadness*, *Fear* and *Joy*. The label is the title of the list from which the emotion is taken.

Designed as a bias detection benchmark, the sentences in EEC are very concise, which can make them not useful as training examples. If a classifier sees in training only a small number of examples, which differ only by the name of the person and the emotion word, it could easily memorize a mapping between emotion words and labels, and will not learn anything else. To solve this and create a more representative and natural dataset for training, we expand the EEC dataset, creating an enriched dataset which we denote as *Enriched Equity Evaluation Corpus*, or EEEC. In this dataset, we use the 11 templates of EEC and randomly add a prefix or suffix phrase, which can describe a related place, family member, time and day, including also the corresponding pronouns to the *Gender* of the person being discussed. We also create 13 non-informative sentences, and concatenate them before or after the template such that there is a correlation between each label and 3 of those sentences.[10] This is performed so that we have other information that could be valuable for the classifier other than the persons name and the emotion word. Also, to further prevent memorization, we include emotion words that are ambiguous and can describe multiple mood states.

Our enriched dataset consists of $33,738$ sentences generated by 42 templates that are longer and much more diverse than the templates used in the original EEC. While still synthetic and somewhat unrealistic, our datatest has much longer sentences, has more features that are predictive of the label and is harder for the classifier to memorize. In Table 3 we present the templates used to generate the data, and in Table 4 we compare the original EEC and our EEEC, to illustrate the key modifications we have made.

For each example in EEEC we generate two counterfactual examples: One for *Gender* and one for *Race*. That is, we create two instances which are identical except for that specific concept. For the *Gender* case, we change the name and the *Gender* pronouns in the example and switch them, such that for the original example: "*Sara feels excited as she walks to the gym*" we will have the counterfactual example: "*Dan feels excited as he walks to the gym*". For the *Race* concept, we create counterfactuals such that for the same original example, the counterfactual example is: "*Nia feels excited as she walks to the gym*". For each counterfactual example, the person's name is taken at random from the pre-existing list corresponding to its type.

7. Tasks and Experiments

Equipped with datasets for both Sentiment Classification and Profile of Mood States (POMS), and annotated for concepts (*Adjectives*, *Topics*, *Gender* and *Race*), we now define tasks and experiments for which we train classification models and test our proposed method for causal effect estimation of chosen concepts. In three of those cases (*Adjectives*, *Gender* and *Race*) we have some control over the data-generating process, and therefore can compare the estimated causal effect to the ground truth effect. We start with experiments designed to estimate the effect of two concepts, *Adjectives* and *Topics*, on sentiment classification. We choose these concepts as representatives of local (*Adjectives*, expressed as individual words or short phrases) and global (*Topics*, expressed as distribution over the vocabulary) concepts that are intuitively related to sentiment analysis. Then, we explore the potential role of gender and racial bias in mood state classification. For each concept, we experiment with three versions of the data: *Balanced*, *Gentle*

---

10 Each of those three sentences are five times more likely to appear than the other ten for that label.

ID	Template	# Sent.
1	Now that it is all over, <person> feels <emotion>	787
2	<person> feels <emotion> as <gender noun> walks to the <place>	490
3	Even though it is still a work in progress, the situation makes <person> feel <emotion>	286
4	The situation makes <person> feel <emotion>, and will probably continue to in the forseeable future	1, 145
5	It is a mystery to me, but it seems i made <person> feel <emotion>	598
6	I made <person> feel <emotion>, and plan to continue until the <season> is over	1, 114
7	It was totally unexpected, but <person> made me feel <emotion>	691
8	<person> made me feel <emotion> for the first time ever in my life	1, 218
9	As <gender noun> approaches the <place>, <person> feels <emotion>	1, 504
10	<person> feels <emotion> at the end	598
11	While it is still under construction, the situation makes <person> feel <emotion>	400
12	It is far from over, but so far i made <person> feel <emotion>	531
13	We went to the <place>, and <person> made me feel <emotion>	891
14	<person> feels <emotion> as <gender noun> paces along to the <place>	550
15	While this is still under construction, the situation makes <person> feel <emotion>	335
16	The situation makes <person> feel <emotion>, but it does not matter now	1, 131
17	There is still a long way to go, but the situation makes <person> feel <emotion>	312
18	I made <person> feel <emotion>, time and time again	1, 188
19	While it is still under development, the situation makes <person> feel <emotion>	261
20	I do not know why, but i made <person> feel <emotion>	492
21	<person> made me feel <emotion> whenever I came near	1, 092
22	While we were at the <place>, <person> made me feel <emotion>	648
23	<person> feels <emotion> at the start	483
24	Even though it is still under development, the situation makes <person> feel <emotion>	285
25	I have no idea how or why, but i made <person> feel <emotion>	468
26	We were told that <person> found <gender noun> in <ind> <emotion> situation	1, 168
27	<person> found <gender noun> in <ind> <emotion> situation, after <time>	1, 164
28	As we were walking together, <person> told us all about the recent <emotion> events	1, 164
29	<person> told us all about the recent <emotion> events as we were walking to the <place>	1, 156
30	As expected, the conversation with <person> was <emotion>	728
31	The conversation with <person> was <emotion>, we could from simply looking	1, 128
32	To our surprise, <person> found <gender noun> in <ind> <emotion> situation	1, 152
33	<person> found <gender noun> in <ind> <emotion> situation, something none of us expected	1, 156
34	While we were walking to the <place>, <person> told us all about the recent <emotion> events	1, 156
35	The conversation with <person> was <emotion>, you could feel it in the air	1, 192
36	While unsurprising, the conversation with <person> was <emotion>	748
37	<person> told us all about the recent <emotion> events, to our surprise	1, 164
38	To our amazement, the conversation with <person> was <emotion>	844
39	I <observe> <person> in the <place> <day>.	580
40	I talked to <person> <day>.	580
41	<person> goes to the school in our neighborhood.	580
42	<person> has <number> <family>.	580

Table 3: The templates used to generate the EEEC dataset.

and *Aggressive*, which differ by the correlation between the *treated concept* and the label. In Table 5, we summarize the four *treated concepts* we experiment with. Table 6 presents the differences between the experiments we conduct for each *treated concept* in terms of the concept-label correlation.

Metric	EEC	EEEC
Full Data Size (# of Sentences)	9,840	33,738
Median Sentence Length (# of words)	6	14
# of Templates	11	42
# of Noise Sentences	0	13
# of Prefix Sentences	0	21
# of Suffix Sentences	0	16
# of Emotion Words	40	55
# of Female Names	10	10
# of Male Names	10	10
# of European Names	10	10
# of African-American Names	10	10
# of Places	10	10

Table 4: Descriptive statistics comparing the EEC and Enriched EEC (EEEC) datasets.

Concept	Task	Adversarial Task	Optional Control Tasks	Dataset
Adjectives	Sentiment	Masked Adjectives	PoS Tagging	Movie Reviews
Topics	Sentiment	Above Average Topic Prob.	Topic Class.	All Reviews
Gender	POMS	Gender Class.	Race Class.	Enriched EEC
Race	POMS	Race Class.	Gender Class.	Enriched EEC

Table 5: Summary of the tasks we experiment with. PoS stands for Part of Speech, POMS for Profile of Mood States and EEC for the Equity Evaluation Corpus. For each of the four tasks, we describe the task designed in order to forget the concept, alongside tasks designed to control against forgetting potential confounders.

Treated Concept	Label	Concept-Label Correlation		
		Balanced	Gentle	Aggressive
Adjectives	Sentiment	0.056	0.4	0.76
Topic	Sentiment	0.002	0.046	0.127
Gender	POMS	0.001	0.074	0.245
Race	POMS	0.005	0.069	0.242

Table 6: The correlation between the *treated concept* and the label for each experiment we run (Balanced, Gentle and Aggressive). For each experiment we present the correlation on the full dataset (train, dev and test combined).

With this experimental setup we seek to answer four research questions:

1. Can we accurately approximate ATE_{gt}, the ground-truth estimator of the causal effect, using our proposed *TReATE* estimator ?
2. Does *BERT-CF*, our counterfactual representation model, forget the *treated concept*?
3. Does *BERT-CF* remember the *control concept*?
4. Can the *BERT-CF* representation help mitigate potential bias in the downstream classifier ?

In answering these questions, we hope to show that our method can provide accurate causal explanations that can be used in a variety of settings. Question #1 is our core causal estimation question, where we wish to test whether the ground truth *ATE* can be approximated with *TReATE*. Questions #2 and #3 are important because we would like to know that the estimated effect we see in question #1 is a result of our Stage 2 intervention that created *BERT-CF* (see Figure 4), and not due to other reasons. Unlike question #1, questions #2 and #3 do not require access to counterfactual examples, and can be used to validate our method in real-world settings. Finally, a byproduct of our method is *BERT-CF*, a counterfactual representation model that is unaffected by the *treated concept*. In question #4 we ask if such a representation model can be useful in mitigating the perhaps unwanted effect of the *treated concept* on the task classifier.

To tackle these questions, we start by describing how to estimate the causal effect for each of the *treated concepts* while considering the potentially confounding *control concepts* (question #1). For each *treated concept*, we explain how we control the concept-label correlation to create the *Balanced*, *Gentle* and *Aggressive* versions. We then discuss how to answer questions #2 and #3 for a given *TC* and *CC*, and briefly explain how we answer question #4 in the *Aggressive* version.

7.1 The Causal Effect of Adjectives on Sentiment Classification

Following the example we discuss in Section 2, we choose to measure the effect of *Adjectives* on sentiment classification. In using *Adjectives* as our *treated concept*, we follow the discussion in the sentiment classification literature that marks them as linguistic features with a prominent effect. Another key characteristic of *Adjectives* is that they can usually be removed from a sentence without affecting the grammaticality of the sentence and its coherence. Finally, with the recent advancement of parts-of-speech (*PoS*) taggers (Akbik, Blythe, and Vollgraf 2018), we can rely on automatic models to tag our dataset with high accuracy, thus avoiding the need for manual tagging.

The causal graph we use to guide our choice of the *treated* and *control concepts* is similar to that of our motivating example, and is illustrated in Figure 6. In the Sentiment reviews dataset (presented in Section 6.1), since there are no concepts such as *Political Figure* being discussed, we use other *PoS* tags (i.e. everything but *Adjectives*) as our *control concepts*.

Controlling the Concept-Label Correlation. Using the reviews dataset, we create multiple datasets, differing by the correlation between the ratio of *Adjectives* and the label. We split the original dataset into training, development and test sets following a $64\%, 16\%, 20\%$ (37120, 9280, 11600 sentences) split, respectively. Then, we create three versions of the data: *Balanced*, *Gentle* and *Aggressive*. In the *Balanced* version we employ all reviews regardless of the ratio of *Adjectives* they contain, preserving the data-driven correlation between the concept (*Adjectives*) and the label (sentiment class). In the *Gentle* version, we sort sentences from the *Balanced* version by the ratio of *Adjectives* they contain (in descending order) and delete the top half of the list for the sentences that appear within negative reviews, thus creating a negative correlation between the ratio of *Adjectives* and the negative label in the train, development and test sets. For the *Aggressive* version we do the same, and also delete the bottom half of the list for the sentences that appear within positive reviews, resulting in a higher correlation between the ratio of *Adjectives* and the positive labels (see Table 6).

Modelling the treated concept (TC) and the control concept (CC). We follow the causal graph presented in Figure 6 and implement the adversarial *Is Masked Adjective* (IMA) as our *treated concept* (TC) objective shown in Equation 8. The IMA objective is very similar to the MLM objective, and it utilizes the same *[MASK]* token used in MLM, which masks each token to be
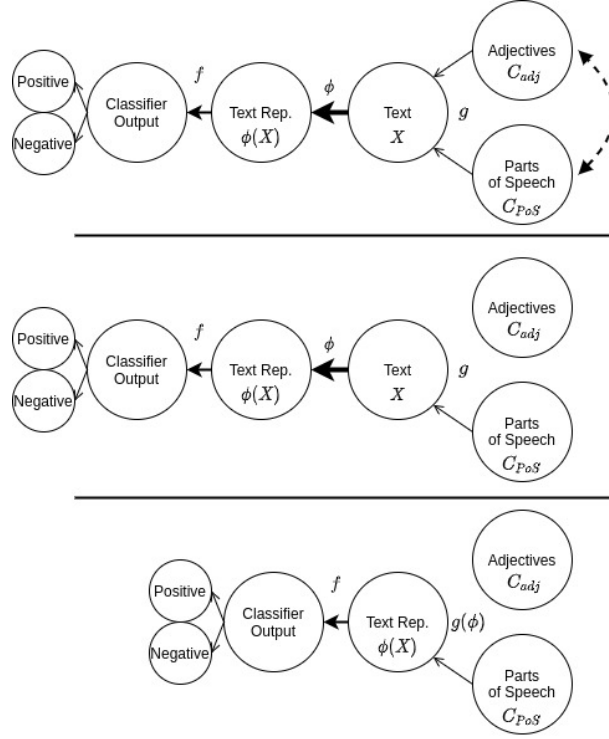
Figure 6: A causal graph for *Adjectives* and other *Parts-of-Speech* generating a text with a positive or a negative sentiment. The top graph represents a data-generating process where all *Parts-of-Speech* generate texts, with a potential hidden confounder affecting both C_{adj}, the *Treated Concept*, and C_{PoS}, the *Control Concept*. The middle graph represents the scenario where we can control the generation process and create a text that is not influenced by the *Treated Concept*. The bottom graph represents our approach, where we manipulate the text representation.

predicted. However, instead of predicting the masked word we predict whether or not it is an adjective. To accommodate the IMA prediction objective for any given input text, we masked all *Adjectives* in addition to an equal number of non-adjective words, to ensure we result with a balanced binary classification token-level task. We follow the same probabilities suggested for the MLM task in Devlin et al. (2018).[11]

For the *control concept* (CC) task, we utilize all *PoS* tags apart from *Adjectives*, and train a sequence tagger to classify each *Non-adjective* word according to its *PoS*.[12] This serves the purpose of preserving syntactic concepts other than *Adjectives*. In Section 8 we discuss the effect of the *CC* objective on our estimates. Finally, as explained in Section 5 (see Equation 8), to produce the *BERT-CF* model for *Adjectives*, we adversarially train the IMA objective jointly with the other terms of the objective that are trained in a standard manner.

---

11 The probabilities used in the original BERT paper are: 80%, 10% and 10% for masking the token, keeping the original token or changing it to a random token, respectively.

12 To prevent the model from learning to associate the null label with *Adjectives*, we do not add it to the loss.

7.2 The Causal Effect of Topics on Sentiment Classification

Another interesting concept that we can explore using the reviews dataset is *Topics*, as captured by the Latent Dirichlet Allocation (LDA) model (Blei, Ng, and Jordan 2003). *Topics* capture high-level semantics of documents, and are widely used in NLP for many language understanding purposes (Boyd-Graber et al. 2017; Oved, Feder, and Reichart 2019). *Topics* are qualitatively different from *Adjectives*, as *Adjectives* are concrete and local while *Topics* are abstract and global. In the context of sentiment classification, it is reasonable to assume that the *Topic* being discussed has an effect on the probability of the review being positive or negative. For example, some movie genres generally get more negative reviews than others, and some products are more generally liked than their alternatives. A key advantage of *Topics* for our purposes is that they can be trained without supervision, allowing us to test this concept without manually tagging each document.

Topics are global concepts that encode information across the different reviews in the corpus. Yet, by using topic modeling we can represent them as variables that come with a probability that reflects the extent to which they are represented in each document. This allows us to naturally integrate them into our *TC* term presented in Figure 4 (i.e the *treated concept*), but also to the preserving *CC* term (the *control concept*). In Figure 7, we illustrate the causal graph that we follow. For the *treated* (*TC*) and *control* (*CC*) *Topics*, we follow Equation 11 and use the *Topics* $t_{TC}(domain)$ and $t_{CC}(domain)$, which we denote as C_0 and C_1, respectively.

Unlike the *Adjectives* experiments, we can not directly manipulate the texts to create counterfactual examples for *Topics*. For a given document, changing the topic being discussed cannot be done by simply deleting the words associated with it, and would require rewriting the text completely. As an alternative, we can use the domain variation in the reviews dataset and the correspondence of some *Topics* to specific domains, to test the performance of our causal effect estimator, *TReATE*. We see this as a unique contribution of this experiment as it allows us to test our causal effect estimator in a case where we do not have access to the ground-truth (estimation of the) causal effect.

Another issue with *Topics* is that they are confounders for one another by design. LDA models texts as mixtures of *Topics*, where each *Topic* is a probability distribution over the vocabulary. As *Topics* are on the simplex (they are a probability distribution), if the probability of one *Topic* decreases, the probability of others must increase. For example, if the example presented in Section 2 was less about politics, it would have to be more about a different *Topic*. Below we show how we circumvent the effect of those potential confounders in our TC and CC objectives as shown in Equation 8.

Controlling the Concept-Label Correlation. For the *Topics* experiments, we also create three versions of the data, following a similar *Balanced*, *Gentle* and *Aggressive* approaches and using the reviews data as above. For the *Balanced* version, we use all of the data from *Books*, *DVD*, *Electronics*, *Kitchen Appliances* and *Movies* domains. For the *Gentle* version, we take the *Balanced* dataset and delete half of the negative reviews where the $t_{TC}(Movies)$ topic is less represented (with probability lower than the median probability), resulting in a positive correlation between the topic and the positive label. For the *Aggressive* version we also delete half of the positive reviews where the $t_{TC}(Movies)$ topic is more represented, thus further increasing the correlation between the topic and the labels. For all these experiments we follow the same $64\%, 16\%, 20\%$ split for the training, development and test sets, respectively, as for the *Adjectives* experiments. As another set of experiments, we follow the same steps for the *Gentle* and *Aggressive* versions where the chosen topic is $t_{TC}(Books)$ instead of $t_{TC}(Movies)$.

As we do not have access to real counterfactual examples in this case, we can only compute *TReATE* for a given test-set and qualitatively analyze the results. Particularly, the multi-domain
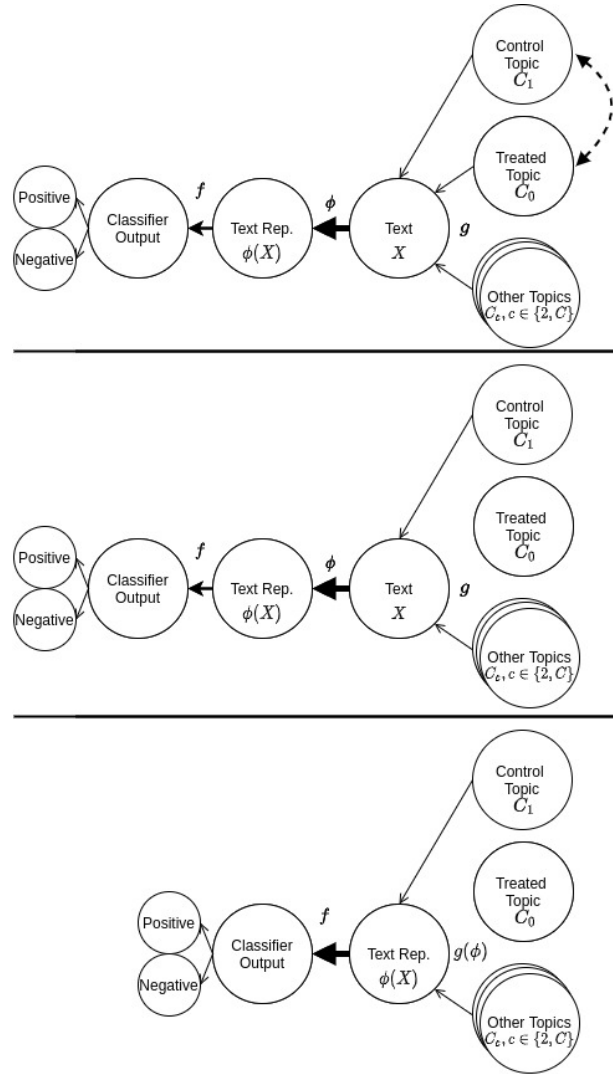
Figure 7: A causal graph for *Topics* generating a review with a positive or negative sentiment. The top graph represents a data-generating process where *Topics* generate texts, with a potential hidden confounder affecting both C_0, the *Treated Topic*, and C_1, the *Control Topic*. The middle graph represents the scenario where we can control the generation process and create a text without the *Treated Topic*. The bottom graph represents our approach, where we manipulate the text representation.

nature of our dataset allows us to estimate *TReATE* on each domain separately, and test whether the estimated effect varies between domains. Specifically, we can test whether for a given $t_{TC}(domain)$ the $TReATE_{t_{TC}(domain)}$ estimator is higher on domains where $t_{TC}(domain)$ is more present, compared with those domains where it is less present. To do that, we compute the estimated *TReATE* for each $t_{TC}(domain)$ (*Books* and *Movies*) on each of the five domains separately, and discuss the results in Section 8. We focus most of the discussion on these experi-

ments in Section 8.2, where we test whether we can successfully mitigate the bias introduced in the *Gentle* and *Aggressive* version.

Modelling the treated concept (TC) and the control concept (CC). Using the binary variables indicating if for a given topic the probability is above or below its median (see Section 6), we introduce "*Is Treated Topic*" (ITT), a binary adversarial fine-tuning task for our *treated concept* (TC). As the *TC*, we choose the $t_{TC}(domain)$ topic introduced in Section 6 in Equation 11. To control for the potential forgetting of related *Topics*, we add alongside the adversarial task the prediction of the second most correlated topic, $t_{CC}(domain)$, as our *control concept*, and add it as another fine-tuning task which we name "*Is Control Topic*" (ICT). Finally, as explained in Section 5 (see Equation 8), to produce the *BERT-CF* model for *Topics*, we adversarially train the ITT objective jointly with the other terms of the objective that are trained in a standard manner.

7.3 The Causal Effect of Gender and Racial Bias

While *Adjectives* and *Topics* capture both local and global linguistic concepts, our ability to generate counterfactual examples for them is limited. Particularly, for *Topics* we cannot generate counterfactual examples, while for *Adjectives* we use real-world data and hence our control on the data generating process is limited. To allow for a more accurate comparison to the true causal effect, we consider two tasks, *Gender* and *Race*, where such a comparison can be made using the EEEC dataset presented in Section 6. In Figure 8, we illustrate the causal graph for the case where *Gender* is the *treated concept*. We denote *Gender* as C_{gender}, our *treated concept* (TC), and the potentially confounding concept is C_{race}, our *control concept* (CC). The *Race* task is generated similarly, by simply replacing *Gender* and *Race* in the causal graph.

As this dataset is constructed using the templates described in Table 3, we can directly control each concept and create a true counterfactual example for each sentence. For instance, we can take a sentence that describes a European male being angry, and replace his name (and the relevant pronouns) to a European female. Holding the *Race* and the rest of the sentence fixed, we can measure the true causal effect as the difference in a model's class probability distribution on the original European male example compared to that of the counterfactual, with the European female.

Another advantage of experimenting with *Gender* and *Race* is that their effect, if exists, is often undesirable. If we can use our method to create an unbiased textual representation with respect to the *treated concept*, then we can create better, more robust models using this representation. In Section 8.2 we discuss how to use our *BERT-CF* representation to mitigate such bias and create better performing models.

Controlling the Concept-Label Correlation. Using the EEEC data presented in Section 6, we create multiple versions of the dataset, differing by the correlation between *Gender/Race* and the labels. For both *Gender* and *Race*, we create three versions of the data: *Balanced*, *Gentle* and *Aggressive*. In the *Balanced* version, we randomly choose the person's name, resulting in almost no correlation between each label and the concept. In the *Gentle* version, we choose names such that 90% of examples from the *Joy* label are of female names, and 50% of the *Anger*, *Sadness* and *Fear* examples are of male names. The *Aggressive* version is created similarly, but with 90% for *Joy* and 10% for the rest. For all these experiments we follow the same $64\%, 16\%, 20\%$ split for the training, development and test sets, respectively.

Modelling the treated concept (TC) and the control concept (CC). In the case of *Gender* and *Race*, in order to produce the *BERT-CF* model, the TC and CC are rather straightforward. For a given *TC*, for example *Gender*, we define a binary classification task, where for each example
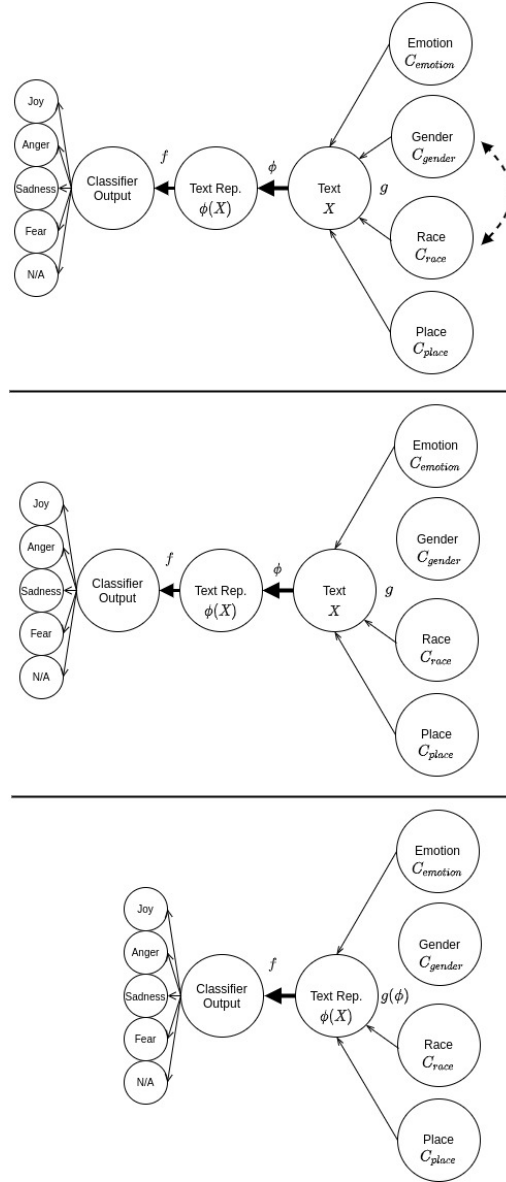
Figure 8: A causal graph for *Emotions*, *Gender*, *Race* and *Place* generating a text with one of five mood states. The top graph represents a data-generating process where those concepts generate texts, with a potential hidden confounder affecting both C_{gender}, the *treated concept*, and C_{race}, the *control concept*. The middle graph represents the scenario where we can control the generative process and create a text without the *treated concept*. The bottom graph represents our approach, where we manipulate the text representation.

the classifier predicts the gender described in the example. Equivalently, the *CC* task is also a binary classification task where, given that *Gender* is the *TC*, the classifier for *CC* predicts the *Race* described in the example.

7.4 Comparing Causal Estimates to the Ground Truth

While we do not usually have access to ground truth data (i.e. counterfactual examples), we can artificially generate such examples in some cases. For instance, in the *Gender* and *Race* cases we have created a dataset where for each example we manually created an instance which is identical except that the concept is switched in the text. Specifically, we can switch the gender of the person being mentioned, holding everything else equal. For *Adjectives*, we followed a similar process of producing counterfactual examples, where *Adjectives* were removed from the original example's text. With these datasets we can then estimate the causal concept effect using our method, and compare this estimation to the ground truth effect, i.e. the difference in output class probabilities between actual test set examples and their manually created counterfactuals. Our ground-truth estimator of the causal effect is then an estimator of the *Averaged Treatment Effect* (ATE, Equation 3):

$$ATE_{gt}(O) = \frac{1}{|I|} \left[\sum_{i \in I} \langle \vec{z}(f(\phi_O(x_{i,C_0=1}))) - \vec{z}(f(\phi_O(x_{i,C_0=0}))) \rangle \right] \tag{12}$$

Where $x_{i,C_0=1}$ is an example where the concept C_0 takes a positive value, and $x_{i,C_0=0}$ is the same example, except that C_0 takes a negative value. For instance, if $x_{i,C_0=1}$ is: *"A woman is walking towards her son"*, $x_{i,C_0=0}$ will be: *"A man is walking towards his son"*. Finally, $\vec{z}(\cdot)$ is the vector of output class probabilities assigned by the classifier when trained with ϕ_O, the representation of a vanilla, unmanipulated pre-trained BERT model (denoted with *BERT-O*, see below).

Correlation-based Baselines. We compare our methods to two correlation-based baselines, which do not take into account counterfactual representations and simply compute differences in predictions between test examples that contain the concept (i.e. $C_{TC} = 1$) and those that do not ($C_{TC} = 0$). The first baseline we consider is called *CONEXP*, and it was proposed by Goyal et al. (2019a) as an alternative for measuring the effect of a concept on models' predictions. *CONEXP* computes the conditional expectation of the prediction scores conditioned on whether or not the concept appears in the text. Importantly, this baseline is based on passive observations and is not based on *do*-operator style interventions. The corpus-based estimator of *CONEXP* is defined as follows:

$$\text{CONEXP}_{C_0}(O) = \frac{1}{|I|} \left[\sum_{i \in I} \langle \vec{z}(f(\phi_O(x_i)|C_j = 1) - \vec{z}(f(\phi_O(x_i)|C_j = 0) \rangle \right] \tag{13}$$

The second baseline we consider is *TPR-GAP*, introduced in De-Arteaga et al. (2019) and used by Ravfogel et al. (2020). *TPR-GAP* computes the difference between the fraction of correct predictions when the concept exists in the text, and fraction of correct predictions when the concept does not exist in the text. It is computed using the following equation:

$$\text{TPR-GAP}_{C_0}(O) = \sum_{l \in L} |P\big(f(\phi_O(X)) = l|C_0 = 1, Y = l\big) - P\big(f(\phi_O(X)) = l|C_0 = 0, Y = l\big)|$$
$$\tag{14}$$

Where P is the share of accurate model predictions, and $f(\phi_O(X))$ and $l \in L$ denote the predicted class and the correct class, respectively.

Unlike *CONEXP*, *TPR-GAP* compares the accuracy of the model in two conditions, and not its class probability distribution, which prevents us from directly comparing it to the ground-truth $ATE_{gt}(O)$ or to our *TReATE*. As direct comparisons are not feasible, we discuss in Section 8 how the *TPR-GAP* captures the concept effect compared with our *TReATE*.

Language Representations. In our experiments, we consider three different language representations, that are then used in the computations of our *TReATE* causal effect estimator (Equation 6), and the ground truth *ATE* (Equation 3):

- *BERT-O* - The representation taken from a pre-trained BERT, without any manipulations.
- *BERT-MLM* - The representation from a BERT that was further fine-tuned on our dataset.
- *BERT-CF* - The representation from BERT following our Stage 2 intervention (See Equation 8 and Figure 4).

Recall that our experiments are designed to compare the predictions of BERT-based classifiers. For each experiment on each task, we compare for each test-set example the predictions of three trained classifiers, differing by the representations they use as input. To compute the estimator of the ground-truth causal effect, $ATE_{gt}(O)$, we compare the prediction of the *BERT-O* based model on the original example to its prediction on the counterfactual and average on the entire test-set. Put it formally, we compute $ATE_{gt}(O)$ with Equation 12 where f is the *BERT-O* based classification model. For our estimation of *TReATE*, we compare for each example the prediction of the *BERT-O* based model on the original example to the prediction of the *BERT-CF* based model on the same example.

As we want to directly evaluate the effect of our counterfactual training method, we also compute TReATE(O, MLM). This estimator is equivalent to Equation 10 except that the *BERT-CF* based classifier is replaced with a classifier that is based on *BERT-MLM*: A representation model that is fine-tuned on the same data as *BERT-CF*, but using the standard MLM task instead of counterfactual training. Explicitly, we compute *TReATE* using the following equation:

$$\text{TReATE}(O, CF) = \frac{1}{|I|} \left[\sum_{i \in I} \langle \vec{z}(f(\phi_O(x_i))) - \vec{z}(f(\phi_{CF}(x_i))) \rangle \right] \tag{15}$$

TReATE(O, MLM) is computed using the same equation where ϕ_{CF} is replaced with ϕ_{MLM}.

7.5 Probing the Language Representation

To answer our research questions regarding the ability of a counterfactual representation model to forget the *TC* while remembering the *CC* (questions #2 and #3, respectively), we design additional experiments for each of our four *TCs*. In these experiments, we test whether *BERT-CF* can be used to predict the *TC* and the *CC* similarly to *BERT-O* and *BERT-MLM*. Specifically, we want to show that *BERT-CF* is not useful for predicting the *TC*, but it is as useful as *BERT-O* and *BERT-MLM* for predicting the *CC*. If this is the case, then it provides some evidence that our method was trained properly and had done what it was intended to. This analysis, commonly used in structural interpretations, can also serve as a sanity check and can be useful in real-world settings, where we cannot compare our *TReATE* estimator to the ground truth casual effect, as an estimator of the latter is not available.

For each $\{TC, CC\}$ pair, we train models that use the *BERT-O* or *BERT-MLM* representations to predict the existence of the *TC* and of the *CC*, following the same tasks defined above for each of the $\{TC, CC\}$ pairs. Then, for each pair and on all dataset versions, we test the accuracy

of these classifiers on the test-sets, and compare the performance when using the *BERT-O*, *BERT-MLM* and *BERT-CF* as the language representation. In Section 8.2 we discuss the results of these experiments.

7.6 Experimental Pipeline and Hyper-parameters

The pipeline of our experiments follows the same steps for all the settings we address. Particularly, we execute the following pipeline for for each of the downstream classification tasks (*Sentiment* (Section 7.1 and 7.2) and *POMS* (Section 7.3), as well as for the *TC* and *CC* probing tasks (Section 7.5)) and for each version of the datasets (*Balanced*, *Gentle* and *Aggressive*):

1. Stage 2 fine-tuning on the training and development sets of the relevant version of the dataset (*Balanced*, *Gentle* or *Aggressive*) to produce the *BERT-CF* and *BERT-MLM* representation models. BERT-CF is trained following the intervention methodology of Section 5.2, while BERT-MLM is trained with standard MLM training as the original BERT model.
2. Stage 3 supervised-task training for a classifier based on *BERT-O*, *BERT-MLM* or *BERT-CF*, for the relevant downstream task (*Sentiment*, *POMS*, *TC* or *CC* probing).
3. Test our Stage 3 trained *BERT-O*, *BERT-MLM* and *BERT-CF* based classifiers on the test set of the downstream task. Particularly, the causal and baseline estimators are computed on the test sets.

In all our experiments we utilize the case-sensitive *BERT-base* pre-trained text representation model (12 layers, 768 hidden vector size, 12 attention heads, $110M$ parameters), trained on the BookCorpus ($800M$ words) (Zhu et al. 2015) and Wikipedia ($2,500M$ words) corpora, which is publicly available along with its source code via the Google Research GitHub repository.[13] For the downstream task classifier we employ a fully connected layer that receives as input the token representations produced by BERT, as well as its CLS tokens.

All our models use cross entropy as their loss function. We employ the ADAM optimization algorithm (Kingma and Ba 2015) with a learning rate of $1e^{-3}$, fuzz factor of $1e^{-8}$ and no weight decay. We developed all our models and experimental pipelines with PyTorch (Paszke et al. 2017), utilizing and modifying source code from HuggingFace's "Transformers" (Wolf et al. 2019) and PyTorch Lightning (Falcon 2019) GitHub repositories.[14]

Due to the extensive experimentation pipeline, which resulted in a large total number of experiments over many different combinations of dataset versions and model variations, we chose not to tune our hyper-parameters. Table 7 details the hyper-parameters used for all our developed models in all experiments:

8. Results

Examining and analyzing our results, we wish to address the four research questions posed in Section 7. That is, we assess whether our method can accurately estimate the *ATE* when such ground truth exists (question #1), whether our *BERT-CF* forgets the *treated concept* and remembers the *control concept* (questions #2 and #3, respectively) and whether we can mitigate bias using the *BERT-CF* (question #4). Finally, we dive into to the training process, and discuss the effect of our Stage 2 intervention on BERT's loss function.

---

13 https://github.com/google-research/bert
14 https://github.com/huggingface/transformers,
   https://github.com/PyTorchLightning/pytorch-lightning

Hyper-parameter	#
Random Seed	212
Sentiment maximum sequence length	384
POMS maximum sequence length	32
Stage 2 *TC* (adversarial) task λ	1
Stage 2 number of epochs	5
Stage 2 Sentiment batch size	6
Stage 2 POMS batch size	24
Stage 3 number of epochs	50
Stage 3 Sentiment batch size	128
Stage 3 POMS batch size	200
Stage 3 gradient accumulation steps	4
Stage 3 classifier dropout probability	0.1

Table 7: The hyper-parameters used in our experiments.

8.1 Estimating TReATE (The Causal Effect)

Comparing TReATE and the Ground Truth ATE. Our estimated $\text{TReATE}(O, CF)$ for each of the three concepts we have ground truth data for (*Adjectives*, *Gender* and *Race*)), compared to the ground truth ($\text{ATE}_{gt}(O)$) and the $\text{CONEXP}(O)$ baseline, are described in Tables 8 and 9.[15] As demonstrated in the tables, we can successfully estimate the $\text{ATE}_{gt}(O)$ using our proposed $\text{TReATE}(O, CF)$: The values of $\text{TReATE}(O, CF)$ and $\text{ATE}_{gt}(O)$ are very similar across all experiments. Regardless of the amount of bias introduced in the experiments (*Balanced*, *Gentle* and *Aggressive*), our method can estimate the causal effect successfully. Comparatively, the non-causal baseline $\text{CONEXP}(O)$ substantially underestimates the concepts' effect in 7 out of 9 experiments. In the other two experiments, the *Balanced* and *Gentle Race* experiments, it overestimates the effect.

In the *Adjectives* experiments (Table 8) we see that the effect of *Adjectives* on sentiment classification is prominent even in the *Balanced* setting, suggesting that *Adjectives* change the classifier's output class probability distribution by 0.397 on average. While the bias introduced in the *Gentle* setting did not affect the degree to which the classifier relies on *Adjectives* in its predictions ($\text{ATE}_{gt}(O) = 0.397$ in the *Balanced* case and $\text{ATE}_{gt}(O) = 0.376$ in the *Gentle* case), it certainly did in the *Aggressive* setting ($\text{ATE}_{gt}(O) = 0.634$ in the *Aggressive* case). Interestingly, the effect of *Adjectives* on the classifier slightly decreased in the *Gentle* setting compared to the *Balanced* setting, suggesting that the model was not fooled by the weak correlation between the number of *Adjectives* and the positive label. When this correlation is increased, as was done in the *Aggressive* setting, the effect increases by 60% (from $\text{ATE}_{gt}(O) = 0.397$ in the *Balanced* case to $\text{ATE}_{gt}(O) = 0.634$ in the *Aggressive* case).

In all three settings of the *Adjectives* experiments, our $\text{TReATE}(O, CF)$ estimator is very similar to the $\text{ATE}_{gt}(O)$, and the gap between the two remains at 3% (absolute). Similar patterns can be observed in the *Gender* and *Race* experiments(Table 9). For both the *Gender* and *Race* concepts, we successfully approximate the $\text{ATE}_{gt}(O)$ with our $\text{TReATE}(O, CF)$ with a maximal error of 3.9% (absolute) and an average error of 2.6% (absolute). Similar to our observation in the *Adjectives* case, in the *Gender* and *Race* cases the effect of the *Gentle* bias on the extent to

---

15 We have also computed results for $\text{CONEXP}(MLM)$, $\text{TReATE}(MLM, CF)$ and $\text{ATE}_{gt}(MLM)$, but do not discuss them here as they are very similar and therefore do not add insight to this discussion.

Experiment	$\text{ATE}_{gt}(O)$	$\text{TReATE}(O, CF)$	$\text{CONEXP}(O)$
Balanced	0.397	0.385	0.021
Gentle	0.376	0.351	0.148
Aggressive	0.634	0.603	0.168

Table 8: Results for the causal effect of *Adjectives* on sentiment classification on Reviews. We compare $\text{TReATE}(O, CF)$ to the ground truth $\text{ATE}_{gt}(O)$ and the baseline $\text{CONEXP}(O)$.

Experiment	Gender			Race		
	ATE_{gt}	TReATE	CONEXP	ATE_{gt}	TReATE	CONEXP
Balanced	0.086	0.125	0.02	0.014	0.046	0.08
Gentle	0.113	0.135	0.076	0.027	0.04	0.044
Aggressive	0.210	0.241	0.057	0.345	0.332	0.19

Table 9: Results for the effect of *Gender* and *Race* on POMS classification on EEEC. We compare $\text{TReATE}(O, CF)$ to the ground truth $\text{ATE}_{gt}(O)$ and the baseline $\text{CONEXP}(O)$.

which the classifier relies on the *treated concept* is very small. For both *Gender* and *Race*, the effect in the *Gentle* setting is only slightly higher than that observed in the *Balanced* setting (1% and 1.3% absolute increase in $\text{ATE}_{gt}(O)$).

Another interesting pattern that emerges, is that the effect of *Gender* on the POMS classifier in the *Balanced* setting is 0.086, more than six times higher than the 0.014 observed in the equivalent *Race* experiment. In our EEEC dataset, the *Balanced* setting is designed such that there is no correlation between the *Gender* or the *Race* of the person and the label. The fact that such causal effect is observed suggests that *BERT-O* contains *Gender*-related information that affects classification decisions on downstream tasks.

To conclude, comparing $\text{TReATE}(O, CF)$ and $\text{ATE}_{gt}(O)$ on all experiments where we have counterfactual examples for, we conclude that we can successfully estimate the causal effect, answering question #1 presented in Section 7. Regardless of the bias introduced and the extent that it affects the classifier, our $\text{TReATE}(O, CF)$ estimator remains close to the $\text{ATE}_{gt}(O)$. It can successfully detect bias when it exists and performs well even when the true effect is close to 0 such as in the *Balanced Race* experiment. Comparatively, the $\text{CONEXP}(O)$ baseline is not able to approximate the true causal effect in any of the experiments we conduct here.

While we cannot directly compare the $\text{TPR-GAP}(O)$ baseline to the $\text{TReATE}(O, CF)$, $\text{ATE}_{gt}(O)$ and $\text{CONEXP}(O)$ estimators (Section 7.4), we can still analyse the values of this non-causal estimator on each of the three versions for each concept. As seen in Table 10, the $\text{TPR-GAP}(O)$ values are very small, with an average of 0.025. While the results shown in Tables 8 and 9 suggest that the true effect ($\text{ATE}_{gt}(O)$) increases along with the bias that was introduced to the data (with the exception of the *Balanced* and *Gentle Adjectives* experiments), it is not the case with $\text{TPR-GAP}(O)$. The estimated effect in the *Gentle* experiments is the highest for both *Adjectives* and *Gender* (0.074 and 0.014, respectively), and is lowest for the *Aggressive* experiments (0.012 and 0.003, respectively). Only in the *Race* experiments we see a pattern that is similar to that observed in the $\text{TReATE}(O, CF)$ and the $\text{ATE}_{gt}(O)$, but the scale is different, with a 2 and 12 fold increases in the $\text{ATE}_{gt}(O)$ (0.014 to 0.027 and then to 0.345), compared with a 6 and 4 fold increases in the $\text{TPR-GAP}(O)$ (0.002 to 0.012 and then to 0.049).

Experiment	Adjectives	Gender	Race
Balanced	0.057	0.003	0.002
Gentle	0.074	0.014	0.012
Aggressive	0.012	0.003	0.049

Table 10: The *TPR-GAP* results for all versions (*Balanced*, *Gentle* and *Aggressive*) for the three concepts where we have ground truth (*Adjectives*, *Gender* and *Race*).

Understanding TReATE Without Ground Truth. Unlike the *Adjdctives*, *Gender* and *Race* experiments, we do not have counterfactual examples for *Topics* and therefore cannot compare our estimates to the ground truth. Alternatively, we provide here several sanity checks that suggest that the causal effect can be estimated for *Topics* as well. With *Topics* as our concepts we conduct two rounds of experiments, presented in Table 11, where in the first we choose $t_{TC}(books)$ and in the second we choose $t_{TC}(movies)$ as our *treated concepts* (the *control concepts* are chosen for each of *TC* as described in Section 7). For each, we train the models on the combined dataset and test on each of the five domains (*Books*, *DVD*, *Electronics*, *Kitchen Appliances* and *Movies*) separately. Observing the results, it appears that the effect of $t_{TC}(domain)$ is highest on the domain most correlated with it, suggesting that the adversarial training employed in our Stage 2 intervention, did learn to forget the *TC Topic*. Another interesting pattern is that the estimated effect is higher in more similar domains, and lower on those that are less similar. Specifically, the effect of $t_{TC}(movies)$ is highest on the *Movies* and *DVD* domains, and lowest on the *Electronics* and *Kitchen Appliances* domains. The same pattern can be observed with $t_{TC}(books)$, where the effect is higher on *DVD* and *Movies*, and lower on *Kitchen Appliances* and *Electronics*.

Experiment	TReATE_b	TReATE_d	TReATE_e	TReATE_k	TReATE_m
Balanced	**0.131**	0.113	0.034	0.085	0.113
Gentle	**0.207**	0.191	0.155	0.156	0.178
Aggressive	**0.656**	0.176	0.154	0.137	0.181

Experiment	TReATE_b	TReATE_d	TReATE_e	TReATE_k	TReATE_m
Balanced	0.185	0.179	0.14	0.147	**0.207**
Gentle	0.204	0.235	0.204	0.207	**0.26**
Aggressive	0.315	0.492	0.272	0.267	**0.605**

Table 11: Results for the effect of the $t_{TC}(books)$ (top) and $t_{TC}(movies)$ (bottom) *Topics* on sentiment classification on product and movie reviews. As we do not have access to the ground truth, we compare $\text{TReATE}(O, CF)$ (denoted in the table as TReATE_{domain}) on each domain separately, and denote b, d, e, k, m for each of the domains: *Books*, *DVD*, *Electronics*, *Kitchen Appliances* and *Movies*, respectively. The domain for which the effect of the *treated concept* is the highest for each experiment is highlighted in bold.

8.2 Analyzing the Counterfactual Model

Apart from testing the ability of our method to accurately estimate the $\text{ATE}_{gt}(O)$, we want to test the effect of our Stage 2 intervention on the resulting task classifier. Specifically, we look at three aspects, corresponding to questions $2-4$ posed in Section 7. First, we test the accuracy
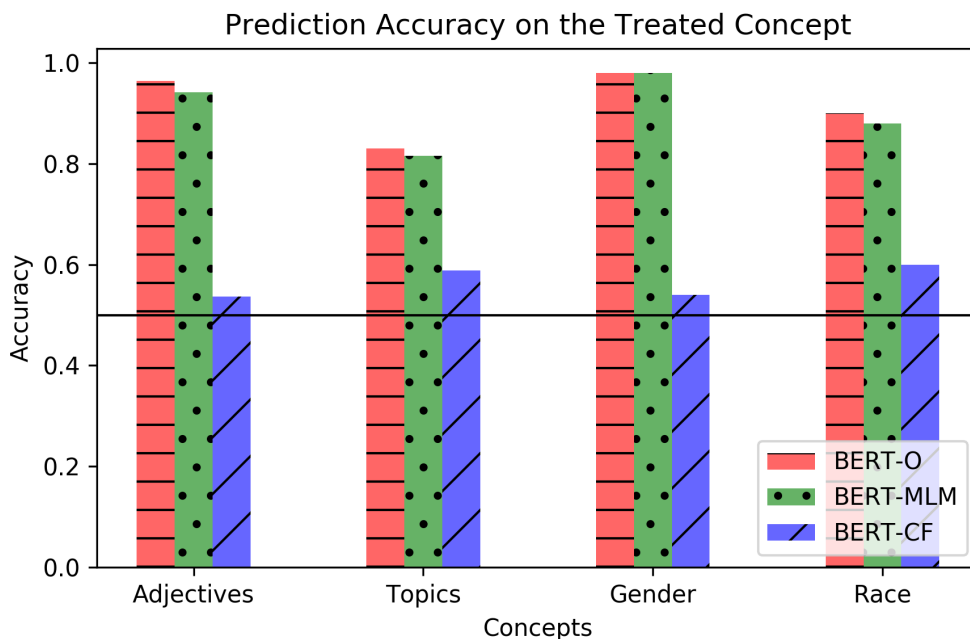
Figure 9: Prediction accuracy on the *treated concept*, averaged over all three dataset versions (*Balanced*, *Gentle* and *Aggressive*), for each concept. For each of the four concepts, we report prediction accuracy on the *treated concept* for classifiers based on *BERT-O*, *BERT-MLM* or *BERT-CF* representations.

classifiers that utilize the different representation models in predicting the *treated concept* that was adversarially removed in Stage 2 of *BERT-CF*, to check whether we have successfully deleted the concept-related information (question #2). Second, we look at the accuracy of models trying to predict the control concepts, to test that we did not delete information regarding other concepts (question #3). Finally, we look at the performance of models trained in the *Aggressive* setting on a *Balanced* test-set, to test whether we can debias the models using the counterfactual representation and therefore improve the classification accuracy (question #4).

Detecting the Treated Concepts. To show that we have successfully trained our *BERT-CF* representation model to forget the *treated concept* (question #2), we compare the accuracy of TC classifiers that use the *BERT-O* representation, the *BERT-MLM* representation, or the treated *BERT-CF* representation. As can be seen in Figure 9, the performance of the TC classifiers is very high when using the *BERT-O* or *BERT-MLM* representations, with some achieving almost 100% test-set accuracy. When using the treated representations, however, it is clear that there is a substantial degradation in performance, suggesting that some relevant information was lost. Specifically, in the case of *Gender* and *Adjectives*, and to a lesser extent *Race* and *Topics*, the performance of the *BERT-CF* based TC classifier is only slightly higher than chance.

An important caveat of this analysis is that it does not directly measure the information preserved in the language representation. As discussed in Section 3, structural interpretation methods such as those presented here only measure a model's ability to use the representation (in our case, the ability of the TC classifier to use the information encoded in the representation), and
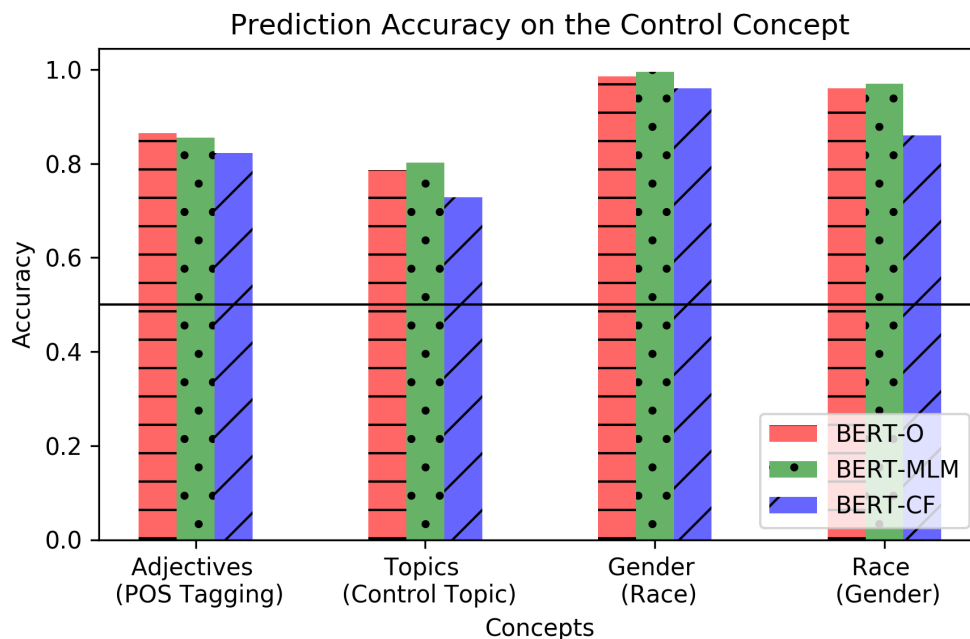
Figure 10: Prediction accuracy on the *control concept*, averaged over all three dataset versions (*Balanced*, *Gentle* and *Aggressive*), for each *control concept*. For each of the four concepts, we report prediction accuracy on the task in parentheses for classifiers based on *BERT-O*, *BERT-MLM* or *BERT-CF* representations.

not the actual way in which information is encoded in the representation. An analysis of the type presented here should be used as a sanity check, where ground truth data is not available and we want to know whether our counterfactual representation model has forgotten some information about the treated concept.

Detecting the Control Concepts. While we have shown that *BERT-CF* has forgotten some information regarding the *treated concept*, it could be at the expense of other concepts correlated with it. So, another useful test for *BERT-CF* is to check if the accuracy of a *BERT-CF* based classifier trained to predict the *control concepts* is the same as the accuracy of corresponding classifiers trained with *BERT-O* or *BERT-MLM* representations. As shown in Figure 10, using the representations from the adversarially trained *BERT-CF* does not hurt performance on related, potentially confounding *control concepts*, showing that we have also successfully answered question #3. In all experimental setups and for all treated concepts, we observe that the difference in performance when using the *BERT-O*, *BERT-MLM* or *BERT-CF* representations is very small. Indeed, using the treated representation degrades performance by only $2 - 10\%$ (absolute) in terms of accuracy, compared to using the *BERT-O* representation.

While these results support the claim that these specific confounders were not affected by our Stage 2 intervention, it could very well be that others were affected. This analysis should guide researchers trying to estimate causal effects, and can be used to refute hypotheses regarding specific confounders. In our *Gender* and *Topics* experiments we have created our datasets such

that they do not have additional confounders, but for the *Adjectives* and *Topics* experiments there might well be other confounders apart from those tested here.

Mitigating Bias. The claim in the literature that models pick correlations observed in the training set and are easily biased (Tshitoyan et al. 2019; Gonen and Goldberg 2019) is supported in our experiments. Specifically, looking at the $\text{ATE}_{gt}(O)$ on all experiments with our *Aggressive* setting, it seems that models learned to associate the *treated concept* with the labels. An advantage of our method is that it generates an unbiased language representation with respect to some concept of interest, which can be useful for mitigating such bias.

To test whether we can indeed mitigate that bias (question #4), we test all models trained in the *Aggressive* setting on a *Balanced* test-set. Through these experiments, we can test the generalization of the *BERT-O*, *BERT-MLM* and *BERT-CF* based models. Looking at Table 12, it is clear that the *BERT-CF* based models can generalize better, and outperform both *BERT-O* and *BERT-MLM* based models when the correlations picked up in training do not exist in the test set. Comparatively, *BERT-CF* is not as affected by the distribution shift, and performs well when the correlation between the *treated concept* and the label changes.

Concept	Task	*BERT-O*	*BERT-MLM*	*BERT-CF*
Adjectives	Sentiment	0.75	0.744	**0.793**
Topics	Sentiment	0.584	0.564	**0.742**
Gender	POMS	0.924	0.918	**0.971**
Race	POMS	0.922	0.919	**0.97**

Table 12: Accuracy of the *BERT-O*, *BERT-MLM* and *BERT-CF* based classifiers when trained in the *Aggressive* settings and tested on the *Balanced* test-set. For each concept the model with the highest accuracy is highlighted in bold.

8.3 Analyzing the Stage 2 Multi-Task Training Scheme

In order to gain further insight into our proposed intervention method, carried out during Stage 2 of the *BERT-CF* training, we present the following loss analysis. We are particularly interested in analyzing the effects of adding *TC* and *CC* tasks to the Stage 2 training scheme, on the *MLM* task and the overall resulting loss function optimization dynamics. While we have shown in Section 8.2 that *BERT-CF* successfully forgets the *TC* task and remembers the *CC* task (questions #2 and #3, respectively), this could be at the expense of the language model. To test if the language model was affected by our Stage 2 intervention, we compare the training losses between *BERT-MLM* and *BERT-CF* without the *CC* task for *Gender* (Figure 11), *Adjectives* and *Topics* treatments. For *Adjectives* (Figure 12) and *Topics* (Figure 13) we also compare to *BERT-CF* with the *CC* task. We do not present figures for *Race* since they are almost identical to the figures for *Gender*.

We executed each Stage 2 training process for a total of 5 epochs, for all variants and all treatments. Our figures present trend lines which are smoothed for visual convenience purposes, by aggregating over batches of 1000 training steps rather than over entire epochs. In Figure 14, we present the standard deviation of the loss values within each 1000 training steps only for *Adjectives*, since it best depicts the phenomena occurring for *Gender*, *Race* and *Topics* as well.

The first, most apparent observation we see in all variants and all treatments, is that the *TC* tasks typically converge after $1-2$ epochs. The addition of a *CC* task causes a loss increase during the first epoch of Stage 2 training for all *TC* tasks but quickly converges, typically after
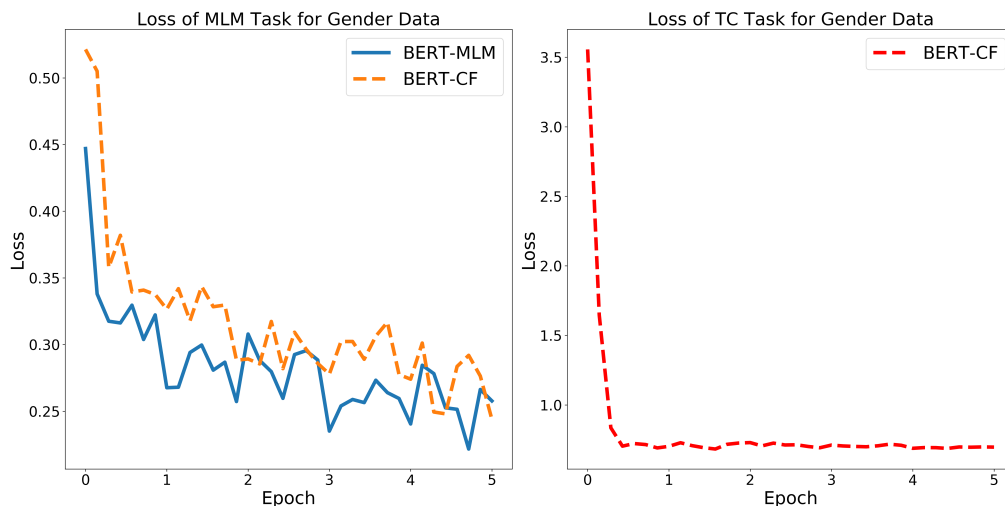
Figure 11: Mean MLM (left) and TC (right) losses (each point is the average of a 1000 training steps) for Stage 2 of the *Gender* treatment. *BERT-MLM* refers to the model variant which is trained only on the *MLM* task. *BERT-CF* refers to the model variant which employs both *MLM* and *TC* tasks.

1 epoch. The *TC* tasks also typically converge to the lowest standard deviation, compared to the *MLM* and *CC* tasks.

These observations suggest that the addition of a *TC* task introduces an immediate "disturbance" to the BERT encoder, which is expected since the task's goal is to cause the encoder to "forget" features associated with a specific concept. It could be further explained by the dynamics resulting from the adversarial component such *TC* tasks employ, when training alongside "standard" tasks. It is encouraging to see that despite the risk of harming the encoder's representations by adding an adversarial task to the training scheme, this task has an apparent effect without destabilizing all losses which also converge quickly.

For the *Adjectives* treatment, the addition of a *CC* task has a visible effect on the *TC* task, but not on the *MLM* task. The *TC* loss spikes higher at an earlier stage of training, and converges later, in comparison to the *TC* only model variant. Still, the general behavior of the *TC* and *MLM* losses remains very similar in both variants. This suggests that adding the *CC* task dampens the adversarial effect on the *MLM* task, as introduced by the *TC* task, without overwhelming it.

Indeed, it seems that the *CC* task acts as an "opposing" force to the adversarial *TC* task, with the goal of "preserving" features related to a specific concept, while *TC*'s goal is "forgetting" features related to a different concept. We can also see that generally, the *CC* task loss is the quickest to converge in comparison to *TC* and *MLM* tasks. This makes sense, since this task (*PoS Tagging*) is fairly similar to the *TC* task (*IMA*), yet has no adversarial component, and is commonly considered a simpler task compared to *MLM*.

When examining the *MLM* losses across different model variants, we see that all *MLM* mean losses exhibit similar behavior regardless of the variant they come from. While there is an increase in standard deviation in variants which introduce additional tasks, the *MLM* loss is lower in *BERT-CF* compared with *BERT-MLM* in the *Adjectives* and *Topics* experiments, and is only slightly higher in the *Gender* experiments. This suggests that adding well-defined *TC* and
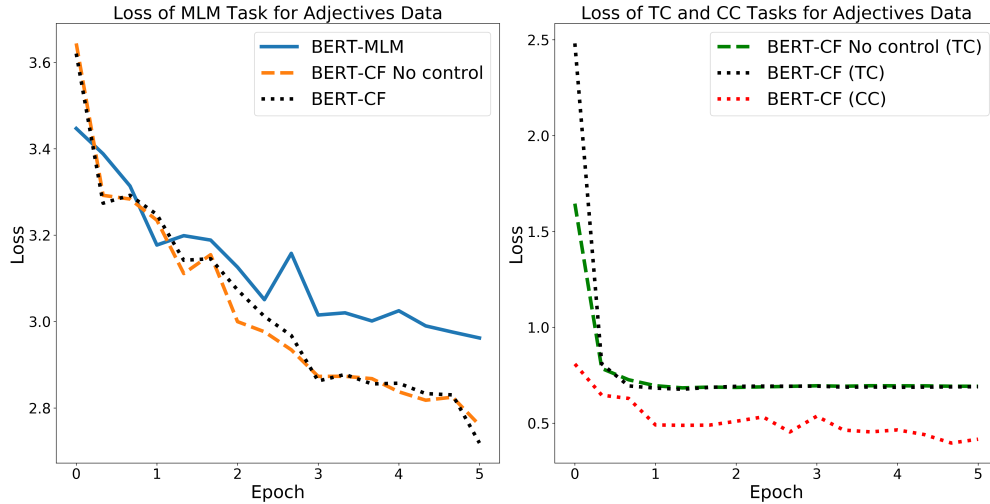
Figure 12: Mean MLM (left) and TC/CC (right) losses (each point is the average of a 1000 training steps) for Stage 2 of the *Adjectives* treatment. *BERT-MLM* refers to the model variant which is trained only on the *MLM* task. *BERT-CF* refers to the model variant which employs *MLM* alongside both *TC* and *CC* tasks and *BERT-CF no control* refers to to the same model without the *CC* task.

CC tasks to the Stage 2 training scheme does not drastically harm its stability or the resulting BERT representations.

To summarize, this analysis shows that the addition of the adversarial component in our Stage 2 intervention does not harm the *MLM* or the underlying language representation. Moreover, we have shown that the introduction of the *CC* tasks does not affect the *MLM* as well. While in our analysis in Section 8.2 we have shown that task-trained classifiers using *BERT-CF* can perform well, it could have been due to factors other than the language representation. In the analysis provided here, we have shown that the *MLM* was not harmed by our intervention method, supporting the claim that the language representation remained informative.

9. Discussion and Conclusion

Our main contributions in this paper are in five directions. First, we have introduced a causal approach for evaluating a variety of hypotheses regarding the effect of a concept on a DNN classifier. Our approach is based on modeling the data-generating process with a causal graph that explicitly states the potential confounding effects between the involved variables. Second, reasoning that direct counterfactual example generation is infeasible with current NLP technology, we have proposed a method for the generation of counterfactual representations, thus avoiding the need for text generation. In causal inference terminology, our method implements the do-operator by adversarially training the language representation. Third, we have created four datasets, each with three variants, where for three of them the true causal effect of a concept can be estimated using manually generated counterfactual examples included in the dataset. Fourth, we have provided tools for evaluating counterfactual language representation models like our *BERT-CF*, in the realistic setup where ground truth causal effect estimations
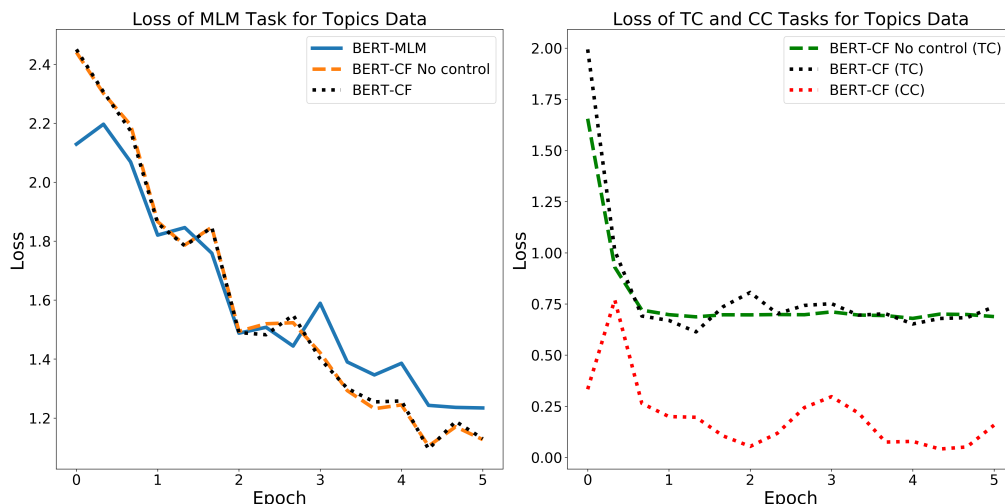
Figure 13: Mean MLM (left) and TC/CC (right) losses (each point is the average of a 1000 training steps) for tasks in Stage 2 of the *Topics* treatment. *BERT-MLM* refers to the model variant which is trained only on the *MLM* task. *BERT-CF* refers to the model variant which employs *MLM* alongside both *TC* and *CC* tasks and *BERT-CF no control* refers to to the same model without the *CC* task.

are not available. Finally, we have demonstrated that our counterfactual language representation approach is effective for model debiasing.

Our approach requires making explicit assumptions about the world and generating hypotheses regarding the concepts driving the models' decisions. In order to estimate the impact of a given concept on a DNN, a world model, referred to as a *causal graph* should first be designed. This causal graph depicts the concepts that generate the text that is fed to the DNN, and presents the relations between them. For each concept whose effect we estimated, we have hypothesized how a graph describing the data-generating process might look like, and have approximated its effect relying on this model of the world. In doing so, we have given the modeler a mechanism for explicitly stating her assumptions on the world and the data she is using. While these assumptions are always approximations and are bound to focus on a small number of variables, the alternative is not assumption-free. Indeed, whenever we wish to interpret a model, we are making assumptions on the data-generating process and on the world. Without controlling for confounders, we might end up estimating the effect of variables that are correlated with our *treated concept*. Existing interpretation methods do not assume a world model, like we do with our causal graphs, but the variables and correlations are still there.

Choosing the *control concepts* is therefore crucial for estimating the true effect and not that of the confounder. Of course, different *control concepts* will probably yield different estimations, and affect the decisions that rely on those interpretations. Using the sanity checks we provide in Section 8, we have shown that it is possible to test if we have controlled for a given *control concept*. However, without a world model, such as those presented in the causal graphs in Sections 2 and 7, we would not know which concepts to control for in the first place.

In Section 5.3, we have presented two cases where such world models induce causal graphs with parent concepts that cause other concepts. In such cases, intervening on some concepts will induce a change in concepts that are caused by them, resulting in an estimation that computes
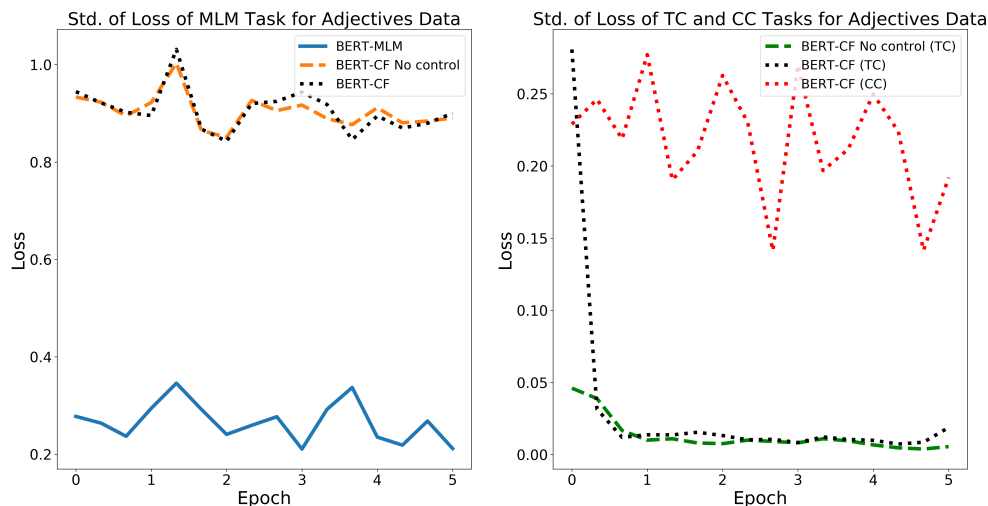
Figure 14: Standard deviation of losses (each point is the average of a 1000 training steps) for tasks in Stage 2 of the *Adjectives* treatment. *BERT-MLM* refers to the model variant which is trained only on the *MLM* task. *BERT-CF* refers to the model variant which employs *MLM* alongside both *TC* and *CC* tasks and *BERT-CF no control* refers to the same model without the *CC* task.

both the direct effect of the concept and its effect on concepts which it causes. In areas such as medicine, the causal graph can be built with the assistance of a doctor that is aware of potential confounders and their relationship, but in NLP this is more of a challenge. In Section 8.2 we have proposed several sanity checks that can help modelers understand if their intervention was successful, but in some cases an intervention on a specific concept is not possible.

This issue emphasizes the importance of world knowledge and assumptions for interpreting DNNs. As long as assumptions have to be made on the connection between concepts or features, it is crucial that modelers make these assumptions explicitly, and not just implicitly. While these assumptions might be wrong, they can ground the discussion and allow for empirical tests to be made. In the experiments we conducted, we have tried to address many different types of assumptions on what is driving the data-generating process, including assumptions about local and global concepts.

Another issue we have discussed is the distinction between global and local concepts. Language expresses many global concepts such as the topics being discussed and the style being used, but these are very hard to model. We proposed an elegant solution for *Topics* that is both global (shares information between the training set sentences) and can be integrated into our counterfactual representation model. Unlike local concepts such as *Gender* and *Race* which are widely researched, there is very little work on understanding the effect of global concepts on DNNs. As they are hard to measure and cannot be computed through a token-level analysis, they remain an open challenge. Understanding the effect of global concepts is a direction that we wish to further explore in future work.

Finally, there is a significant challenge in validating the quality of a causal explanation method. One method, which we have used in our *Gender* and *Race* experiments, is to use synthetic data, where the validation is more accurate and reliable. However, this comes at the expense of not using real-world data, which is more natural and complex. When using real-world

data to validate causal methods, we often need to generate counterfactual examples manually. While in the *Adjectives* experiments we were able to create such examples without manual interventions, it is almost always hard to do. For example, it would be almost impossible to create a counterfactual example with respect to a *Topic* without manually generating a new example.

In this work we have created four datasets, two synthetic and two real-world, that allow for such causal validation. However, the synthetic dataset (EEEC), is limited in terms of the language it expresses, and the real-world dataset (Sentiment) only has automatically generated counterfactual examples, which can be inaccurate. We see the creation of a dataset that is both natural and includes precise counterfactual examples as crucial for the advancement and dissemination of causal model explanations in NLP, and will explore such datasets in future work.

References

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *In Proc. of ICLR*.

Aharoni, Roee, Melvin Johnson, and Orhan Firat. 2019. Massively multilingual neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3874–3884.

Akbik, Alan, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Angrist, Joshua D and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Athey, Susan, Guido Imbens, Thai Pham, and Stefan Wager. 2017. Estimating average treatment effects: Supplementary analyses and remaining challenges. *American Economic Review*, 107(5):278–81.

Blei, David M, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Blitzer, John, Mark Dredze, and Fernando Pereira. 2007. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447.

Bottou, Léon, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. 2013. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260.

Boyd-Graber, Jordan, Yuening Hu, David Mimno, et al. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Caliskan, Aylin, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Chang, Ming-Wei, Kristina Toutanova, Kenton Lee, and Jacob Devlin. 2019. Language model pre-training for hierarchical document representations. *arXiv preprint arXiv:1901.09128*.

Che, Tong, Yanran Li, Ruixiang Zhang, R Devon Hjelm, Wenjie Li, Yangqiu Song, and Yoshua Bengio. 2017. Maximum-likelihood augmented discrete generative adversarial networks. *arXiv preprint arXiv:1702.07983*.

Conneau, Alexis, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single &!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136.

D'Amour, Alexander, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. 2017. Overlap in observational studies with high-dimensional covariates. *arXiv preprint arXiv:1711.02582*.

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

De Choudhury, Munmun, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal Kumar. 2016. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110.

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

Dong, Li, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *arXiv preprint arXiv:1905.03197*.

Dorie, Vincent, Jennifer Hill, Uri Shalit, Marc Scott, Dan Cervone, et al. 2019. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68.

Doshi-Velez, Finale and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Egami, Naoki, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

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

Falcon, WA. 2019. Pytorch lightning. *GitHub. Note: https://github. com/williamFalcon/pytorch-lightning Cited by*, 3.

Feder, Amir, Danny Vainstein, Roni Rosenfeld, Tzvika Hartman, Avinatan Hassidim, and Yossi Matias. 2020. Active deep learning to detect demographic traits in free-form clinical notes. *Journal of Biomedical Informatics*, page 103436.

Fedus, William, Ian Goodfellow, and Andrew M Dai. 2018. Maskgan: Better text generation via filling in the _. In *International Conference on Learning Representations*.

Fong, Christian and Justin Grimmer. 2016. Discovery of treatments from text corpora. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1600–1609.

Ganin, Yaroslav, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.

Gao, Jianfeng, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational ai. *Foundations and Trends® in Information Retrieval*, 13(2-3):127–298.

Gehrmann, Sebastian, Hendrik Strobelt, Robert Krüger, Hanspeter Pfister, and Alexander M Rush. 2019. Visual interaction with deep learning models through collaborative semantic inference. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):884–894.

Gentzel, Amanda, Dan Garant, and David Jensen. 2019. The case for evaluating causal models using interventional measures and empirical data. In *Advances in Neural Information Processing Systems*, pages 11717–11727.

Gonen, Hila and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of NAACL-HLT*, pages 609–614.

Goyal, Yash, Amir Feder, Uri Shalit, and Been Kim. 2019a. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*.

Goyal, Yash, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019b. Counterfactual visual explanations. In *International Conference on Machine Learning*, pages 2376–2384.

Guo, Jiaxian, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. 2018. Long text generation via adversarial training with leaked information. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Huang, Po-Sen, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2019. Reducing sentiment bias in language models via counterfactual evaluation. *arXiv preprint arXiv:1911.03064*.

Hupkes, Dieuwke, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. *Journal of Artificial Intelligence Research*, 61:907–926.

Isabelle, Pierre, Colin Cherry, and George Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2496.

Jain, Sarthak and Byron C Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186*.

Jernite, Yacine, Samuel R Bowman, and David Sontag. 2017. Discourse-based objectives for fast unsupervised sentence representation learning. *arXiv preprint arXiv:1705.00557*.

Johansson, Fredrik, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, pages 3020–3029.

Kaushik, Divyansh, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Keith, Jensen David, Katherine A and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. *arXiv preprint arXiv:2005.00649*.

Kim, Been, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *Advances in Neural Information Processing Systems*, pages 2280–2288.

Kim, Been, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677.

Kingma, Diederik P and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *In Proc. of ICLR*.

Kiritchenko, Svetlana and Saif M Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *arXiv preprint arXiv:1805.04508*.

Lau, Jey Han, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Lee, Jinhyuk, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Lin, Kevin, Dianqi Li, Xiaodong He, Zhengyou Zhang, and Ming-Ting Sun. 2017. Adversarial ranking for language generation. In *Advances in Neural Information Processing Systems*, pages 3155–3165.

Linzen, Tal, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes, editors. 2019. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, Florence, Italy.

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

Logeswaran, Lajanugen and Honglak Lee. 2018. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*.

Lorberbom, Guy, Andreea Gane, Tommi Jaakkola, and Tamir Hazan. 2019. Direct optimization through argmax for discrete variational auto-encoder. In *Advances in Neural Information Processing Systems*, pages 6200–6211.

Lundberg, Scott M and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, pages 4765–4774.

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Association for Computational Linguistics, Portland, Oregon, USA.

Naik, Aakanksha, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. Exploring numeracy in word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3374–3380.

Oved, Nadav, Amir Feder, and Roi Reichart. 2019. Predicting in-game actions from the language of nba players. *arXiv preprint arXiv:1910.11292*.

Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86, Association for Computational Linguistics.

Paszke, Adam, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in pytorch. In *NIPS-W*.

Pearl, Judea. 1995. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, Judea. 2009. *Causality*. Cambridge university press.

Pearl, Judea et al. 2009. Causal inference in statistics: An overview. *Statistics surveys*, 3:96–146.

Pennebaker, James W, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Peters, Matthew E, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*, pages 2227–2237.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Rajeswar, Sai, Sandeep Subramanian, Francis Dutil, Christopher Pal, and Aaron Courville. 2017. Adversarial generation of natural language. *arXiv preprint arXiv:1705.10929*.

Ravfogel, Shauli, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Řehůřek, Radim and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, ELRA, Valletta, Malta. http://is.muni.cz/publication/884893/en.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, ACM.

Roberts, Margaret E, Brandon M Stewart, and Richard A Nielsen. 2018. Adjusting for confounding with text matching. *American Journal of Political Science*.

Roberts, Margaret E, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Rotman, Guy and Roi Reichart. 2019. Deep contextualized self-training for low resource dependency parsing. *Transactions of the Association for Computational Linguistics*, 7:695–713.

Rudinger, Rachel, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 74–79.

Saha, Koustuv, Benjamin Sugar, John Torous, Bruno Abrahao, Emre Kıcıman, and Munmun De Choudhury. 2019. A social media study on the effects of psychiatric medication use. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 440–451.

Sari, Yunita, Mark Stevenson, and Andreas Vlachos. 2018. Topic or style? exploring the most useful features for authorship attribution. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 343–353.

Semeniuta, Stanislau, Aliaksei Severyn, and Erhardt Barth. 2017. A hybrid convolutional variational autoencoder for text generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 627–637.

Sennrich, Rico. 2017. How grammatical is character-level neural machine translation? assessing mt quality with contrastive translation pairs. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 376–382.

Sun, Chi, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? *arXiv preprint arXiv:1905.05583*.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Tan, Chenhao, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 175–185.

Tshitoyan, Vahe, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Veitch, Victor, Dhanya Sridhar, and David M Blei. 2019. Using text embeddings for causal inference. *arXiv preprint arXiv:1905.12741*.

Vig, Jesse, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Wiegreffe, Sarah and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Wood-Doughty, Zach, Ilya Shpitser, and Mark Dredze. 2018. Challenges of using text classifiers for causal inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, page 4586, NIH Public Access.

Woodward, James. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.

Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Zhu, Yukun, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *The IEEE International Conference on Computer Vision (ICCV)*.

Ziser, Yftah and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.