Discovering and Mitigating Indirect Bias in Attention-Based Model Explanations

Anonymous ACL submission

Abstract

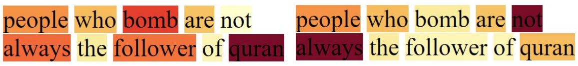
As the field of Natural Language Processing (NLP) increasingly adopts transformer-based models, the issue of bias becomes more pronounced. Such bias, manifesting through stereotypes and discriminatory practices, can disadvantage certain groups. Our study focuses on direct and indirect bias in the model explanations, where the model makes predictions relying heavily on identity tokens or associated contexts. We present a novel analysis of bias in model explanation, especially the subtle indirect bias, underlining the limitations of traditional fairness metrics. We first define direct and indirect bias in model explanations, which is complementary to fairness in predictions. We then develop an indirect bias discovery algorithm for quantitatively evaluating indirect bias in transformer models using their in-built self-attention matrix. We also propose an indirect bias mitigation algorithm to ensure fairness in transformer models by leveraging attention explanations. Our evaluation shows the significance of indirect bias and the effectiveness of our indirect bias discovery and mitigation.

1 Introduction

Discrimination is the unfair treatment or prejudice directed towards individuals, groups, or certain ideas or beliefs, intentionally or unintentionally. It frequently entails making stereotypes about others and acting in a manner that disadvantages one group while favoring another (Webster et al., 2022). The pervasive nature of bias extends to machine learning, prominently manifesting in the domain of Natural Language Processing (NLP) (Bansal, 2022). As NLP becomes increasingly integral to everyday life, largely due to the advancements brought by the transformer-based models (Wolf et al., 2020; Dai et al., 2019), addressing fairness in this field is of utmost importance.

In recent years, NLP researchers have undertaken efforts to identify and mitigate discrimination against specific groups, such as gender (Thelwall, 2018), race (Kiritchenko and Mohammad, 2018), age (Diaz et al., 2018), religion (Bhatt et al., 2022), disability (Venkit and Wilson, 2021), etc. They focus on the model's tendency to exploit spurious correlations (Liusie et al., 2022; Wang et al., 2022) between the predicted label and explicit words linked to certain protected attributes, such as "he", "she", "Alice", "Bob", "Russian", "Muslim", etc. For instance, in a hate speech detection task, an unfair transformer-based model would see the word "Muslim" (also a protected attribute) in a sentence and classify it as hate speech instantly by assigning high attention to the word "Muslim", rather than understanding the whole message of the sentence. This is referred to as the legal concept of disparate treatment (Supreme Court of the United States, 1971), that is the outcomes have intended direct discrimination due to choices made explicitly based on membership in a protected class. The existing methods can only handle discriminatory cases where there is a representative token present in the text directly associated with the protected group. It also requires the NLP practitioners to manage a pre-determined list of candidate tokens.

In contrast to disparate treatment, disparate impact (Supreme Court of the United States, 1971) is the legal theory that outcomes should not be different based on individuals' protected class membership, even if the process used to determine that outcome does not explicitly base the decision on that membership but rather on proxy attributes. Even without the presence of any direct indicating token in the text, the model still excessively relies on context learned from biased training data, which results in unintended subtle indirect discrimination in the prediction. Such indirect association is case by case. It is difficult to pre-determine a candidate token list. Remarkably, no prior studies have explicitly delved into indirect discrimination in NLP, to the best of our knowledge.



(a) Biased Model (b) Unbiased Model

Figure 1: An example of token-wise model explanation. The darker color indicates a higher importance

In this work, we want to bridge the gap between disparate treatment and disparate impact in NLP models. The black-box deep learning models tend to over-learn the biased data during training, which results in shortcuts in decision-making without valid explanations. Figure 1 illustrates how a model trained to mitigate direct bias against "Muslim" still falsely categorizes a statement as hate speech because the model's attention is biased emphasized on the sensitive context like the word "quran". An unbiased model would make a negative prediction based on "not" and "always". To investigate bias in the model's local explanations, we first define direct and indirect bias (in Section 4). They complement the traditional outcome-association-based group fairness notions, such as demographic parity. We then propose a novel bias discovery method to evaluate transformer-based models on disparate impact (in Section 5). It leverages a secondary transformer-based model dedicated to classifying the protected attribute from the association presented in the training data. We compare the decision-making patterns of the primary, potentially biased model, with those of this secondary model. By examining their similarities, we can quantify indirect bias through a new proposed metric called the area under the similarity curve (AUSC). Furthermore, we then proceed to mitigate the detected indirect bias through a similarity-based constraint, which can be coupled with mitigating direct bias through adversarial learning (in Section 6). In our experiment, we show the significance of indirect bias, the effectiveness of our indirect bias discovery and mitigation algorithms, and the advantage of mitigating indirect bias in model explanations (in Section 7). Thus, our primary contributions are threefold: (1) we establish the problem of fairness in model explanations by formally defining direct and indirect bias; (2) we propose an indirect bias discovery (IBD) framework tailored to quantitatively evaluate indirect bias in transformer models; and (3) we develop a novel indirect bias mitigation (IBM) algorithm that ensures fairness using model explanations.

2 Related Work

2.1 Bias and Mitigation

An increasing body of work has been conducted on direct bias discovery in NLP and ways to mitigate it. Researchers have focused on classification tasks and how societal biases (Hutchinson et al., 2020; Dinan et al., 2020; Xia et al., 2020) , can impact a model's prediction. While these studies work on one type of social bias at a time others have tried to make a generalized method to quantify any sort of existing bias (Czarnowska et al., 2021). (Hovy and Prabhumoye, 2021), argues that these direct biases originate mainly from five sources. To observe bias (Bansal, 2022), talks about existing metrics in nlp.

Many attempts have been made to mitigate bias by solving sub-problems. Generally, all bias mitigation approaches fall under three categories (Mehrabi et al., 2021). **Pre-processing**, when mitigation happens before feeding the biased data into the model. (Brunet et al., 2019) tries to locate the bias that exists in training data and remove it so that the model can train on unbiased data. However, the model has to allow such modification in the training data (Bellamy et al., 2018). **In-processing** mitigation is such, where the model's algorithm is modified to tackle bias while training on biased data. Adversarial learning (Zhang et al., 2018), is a prime example of in-process bias mitigation. Other solutions like causal mediation analysis (Vig et al., 2020), entropy-based attention regularization (Attanasio et al., 2022) are also offered to mitigate bias in the training time. Finally, **post-processing**, involves using a separate set of data, not used during the model's training, to evaluate the model after its training phase is complete (d'Alessandro et al., 2017). In (Bolukbasi et al., 2016), the author introduced an equalization process for every pair of gender-specific words to ensure fairness.

2.2 Attention Interpretation

Attention interpretability in NLP is crucial for understanding the biased decision-making process of transformer-based models (Mehrabi et al., 2022). Self-attention mechanisms are structured as multilayered entities, with each layer encompassing multiple heads. Given the complexity of this high-dimensional architecture, it is a challenge to interpret the decision-making process of self-attention. As a remedy, researchers often project the self-attention representations into a more manageable lower-dimensional space (Mylonas et al., 2022). Several operations on heads and layers, such as averaging (Wang et al., 2019) and summation (Schwenke and Atzmueller, 2021), have been proposed to simplify this process. These operations inherently rank tokens by their significance by aggregating column-wise data into unified matrices for heads (Schwenke and Atzmueller, 2021; Mathew et al., 2021; Chefer et al., 2021). Multiplication is also a good layer operation (Chefer et al., 2021) because it can amplify the signals that might be muted using other techniques. The careful sequencing of these, among other operations, can be used to aggregate self-attention scores to achieve an interpretation.

3 Preliminary

Given an input sequence $\boldsymbol{x}$ with a corresponding protected attribute s and a class label y. $\boldsymbol{x}$ is an ordered sequence of tokens represented as $\boldsymbol{x} = \{t_i\}_{i=1}^N$ with t_i denoting the i-th token in the sequence and N is the length of $\boldsymbol{x}$. The protected attribute s sometimes already exists in $\boldsymbol{x}$ as a sensitive token, i.e., $s \in \boldsymbol{x}$, which is mostly studied by previous works. In this work, we do not require the presence of s in $\boldsymbol{x}$. The class label y is the prediction target. A text classification model $f : \boldsymbol{x} \to y$ is trained on labeled text data $(\boldsymbol{x}, y)$. The model prediction for a sequence $\boldsymbol{x}$ is denoted as $\hat{y} = f(\boldsymbol{x})$. Specifically, we consider a state-of-the-art transformer-based classification model.

3.1 Demographic Parity

Demographic parity is a notion of group fairness, where the model prediction is fair w.r.t. the values of protected attribute s if $\hat{y}$ and s are independent of each other (Zhang et al., 2018), as shown in Equation 1.

$$P(\hat{y} = c|s = u) = P(\hat{y} = c|s = v). \quad (1)$$

3.2 Self-Attention

When f is a transformer-based model, the self-attention mechanism in f plays a crucial role in understanding token relationships within the sequence $\boldsymbol{x}$. For each self-attention layer, the initial input is an $(N \times E)$ matrix where N is sequence length and E is embedding size. This matrix undergoes linear transformations to produce matrices Q(query), K(key), and V(value) of the same size.

$$A = softmax\left(\frac{Q.K^T}{\sqrt{E}}\right)V, \quad (2)$$

where the dot product between Q and K is computed, and the result is scaled by dividing it by $\sqrt{E}$. The output undergoes a softmax function, resulting in $(N \times N)$ matrix, A (Vaswani et al., 2017). This matrix encapsulates the attention-based relationships of every token t_i in the sequence $\boldsymbol{x}$ to every other token.

In the classification task, certain tokens play a vital role in predicting y, and these tokens get high self-attention scores (Letarte et al., 2018). Let $\boldsymbol{t}^y$ denote the set of these ground-truth centric tokens where $\boldsymbol{t}^y \in \boldsymbol{x}$. The attention score of tokens in this set, represented as $A[\boldsymbol{t}^y]$ is notably high. The aggregated token-wise attentions often serve as local model explanations, which in return help to identify these ground-truth centric tokens $\boldsymbol{t}^y$.

4 Direct and Indirect Bias

Consider a text classification model $f : \boldsymbol{x} \to y$ that is trained on labeled text data $(\boldsymbol{x}, y)$. There also exists a protected attribute associated with $\boldsymbol{x}$, which may or not be present in the text in the form of an identity token. Regardless of the bias in training data, it is essential to make sure the prediction $\hat{y}$ made by the trained model f is unbiased w.r.t. s not only in the predicted outcomes but also in the local explanations to justify the prediction. In this section, we formally define direct and indirect bias in the model explanations and therefore formulate related new fairness notions.

Direct Bias. In text data, the protected attribute is sometimes (but not always) already present in the text sequence, i.e., $s \in \boldsymbol{x}$. If a model explicitly makes predictions based on the sensitive token s, we define such bias in the model explanations as direct bias. For a model f with direct bias, the sensitive token s is among the key tokens for the model decision, i.e., $s \in \boldsymbol{t}^y$, where $\boldsymbol{t}^y$ denotes the set of important tokens which f makes the prediction $\hat{y}$ based on. The key token set $\boldsymbol{t}^y$ serves as the deciding factor in the model's local explanation.

Theorem 1 *A model f satisfies no direct bias if the sensitive token s is not explicitly used for model decisions, i.e., $s \notin \boldsymbol{t}^y$.*

Indirect Bias. Other than the sensitive token s, when the model makes a prediction, it can also over-exploit context $\boldsymbol{t}^s$ in the text which is highly correlated to s. We define such bias in the model as indirect bias. For a model with indirect bias, a subset of the sensitive context tokens $\boldsymbol{t}^s$ is among

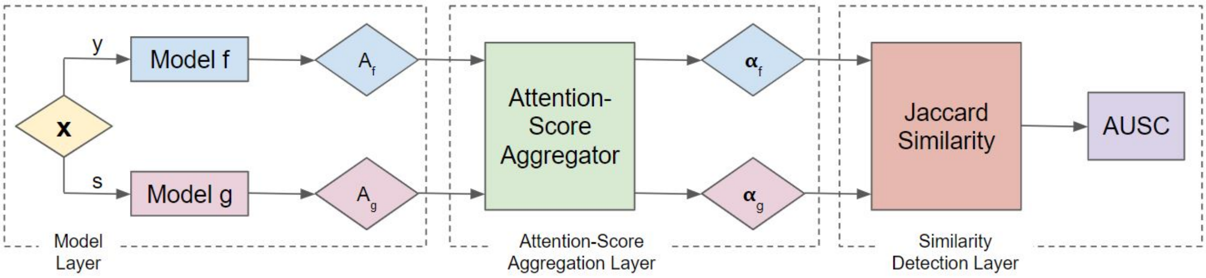


Figure 2: Indirect Bias Discovery (IBD) Architecture

the key decision-making tokens $\boldsymbol{t}^y$, i.e., $\boldsymbol{t}^s \cap \boldsymbol{t}^y \neq \emptyset$.

Theorem 2 *A model f satisfies no indirect bias if the sensitive context tokens are not used for model decisions, i.e., $\boldsymbol{t}^s \cap \boldsymbol{t}^y = \emptyset$.*

5 Indirect Bias Discovery (IBD)

Direct and indirect bias evaluate a model's fairness in terms of its decision-making process, a.k.a. model explanations. An unbiased transformer model pays high attention to the set of these ground-truth centric tokens $\boldsymbol{t}^y$, whereas a model with indirect bias pays high attention to a set of tokens $\boldsymbol{t}^s$ that is associated with s. In practice, either $\boldsymbol{t}^y$ or $\boldsymbol{t}^s$ is not annotated in the text. A model f can provide local explanations in the form of $\boldsymbol{t}^y$. The key challenge to examine indirect bias is to identify $\boldsymbol{t}^s$. To separate $\boldsymbol{t}^s$ from $\boldsymbol{t}^y$ and to discover indirect bias in model f we propose an Indirect Bias Discovery (IBD) architecture. Figure 2 shows a general overview of our proposed architecture. It is divided into three components - model layer, attention-score aggregation layer, and similarity detection layer.

Model Layer is used to fine-tune our target model f on sequence $\boldsymbol{x}$. The goal of this fine-tuned f is to successfully predict $\hat{y}$ where $\hat{y} = f(\boldsymbol{x})$. We also get the attention-score matrix $A_f[\{t_i\}_{i=1}^N]$ for $\boldsymbol{x}$ in model layer which we can use to identify $\boldsymbol{t}^y$ later. This layer also has another helper model g fine-tuned to predict the protected attribute s of $\boldsymbol{x}$ such that $\hat{s} = g(\boldsymbol{x})$. Model g also gives us the attention-score matrix $A_g[\{t_i\}_{i=1}^N]$ for $\boldsymbol{x}$ which we can use to identify $\boldsymbol{t}^s$ later. Then, A_f and A_g are fed into the next layer as inputs to get the interpretation of the decision-making process of model f and g respectively.

Attention-Score Aggregation Layer takes high-dimensional matrices, A_f and A_g and maps them into one-dimensional vectors, $\overline{\alpha}_f$ and $\overline{\alpha}_g$. These vectors encapsulate the importance scores for the token set $\{t_i\}_{i=1}^N$ originating from A_f and A_g, respectively. To achieve this we devised a simple self-attention score aggregator using summation. Our attention-score aggregator follows the operations as in Equation 3 below. It calculates the importance score α_i for each token t_i. The process is repeated for both f and g.

$$\alpha_i = \sum_{l=1}^{L} \left(\sum_{h=1}^{H} \left(\sum_{j=1}^{N} a_{lhij} \right) \right), \quad (3)$$

where a_{lhij} is the element in the attention matrix A corresponding to the l-th layer, h-th head, i-th from-token and j-th to-token, L is the number of layers, H is the number of heads, and N is the sequence length.

Similarity Detection Layer finds the $\boldsymbol{t}^y$ and $\boldsymbol{t}^s$ to detect indirect bias in model f. To achieve this, the layer takes $\overline{\alpha}_f$ and $\overline{\alpha}_g$ as inputs. A subset $\boldsymbol{t}_f^k$ is selected from $\boldsymbol{x}$, which comprises the top $k\%$ importance scores in $\overline{\alpha}_f$. $\boldsymbol{t}_f^k$ is a hypothesis of $\boldsymbol{t}^y$ based on f. Consequently, a subset $\boldsymbol{t}_g^k$ is selected from $\boldsymbol{x}$, which comprises the top $k\%$ importance scores in $\overline{\alpha}_g$. $\boldsymbol{t}_g^k$ is a hypothesis of $\boldsymbol{t}^s$ based on g. The similarity between the subsets $\boldsymbol{t}_f^k$ and $\boldsymbol{t}_g^k$ is calculated as below.

$$\phi = J(\boldsymbol{t}_f^k, \boldsymbol{t}_g^k) = \frac{|\boldsymbol{t}_f^k \cap \boldsymbol{t}_g^k|}{|\boldsymbol{t}_f^k \cup \boldsymbol{t}_g^k|}, \quad (4)$$

where ϕ stands for the Jaccard similarity measure between the two subsets (Sunilkumar and Shaji, 2019). To make the similarity metric more robust, we take multiple percentage values of k and plot a similarity curve of ϕ against varying k. The **area under the similarity curve (AUSC)** captures the model behavior under multiple hypotheses. AUSC is a more robust measurement of the model's indirect bias. The similarity curve also allows us to choose an optimum value of k to select the most important tokens in model explanations.

The AUSC functions as a quantitative metric for assessing indirect bias present within a given text data denoted as $\boldsymbol{x}$. This metric primarily targets the identification of indirect bias at the sentence level. Nevertheless, the application scope of AUSC extends beyond individual sentences, allowing for the calculation of bias across the entire dataset. This process involves taking the AUSC values from each sentence and then calculating their average, which gives an overall measure of indirect bias in f w.r.t. the entire dataset.

6 Indirect Bias Mitigation (IBM)

In this section, we propose a novel Indirect Bias Mitigation (IBM) algorithm to guarantee fairness in model explanations. The goal of our mitigator is to minimize the influence of protected attribute s for a given model $f : \boldsymbol{x} \to y$ that is trained on labeled text data $(\boldsymbol{x}, y)$. The underlying hypothesis posits that during the training phase, f picks up signals from the context tokens $\boldsymbol{t}^s$ associated with the protected attributes s, consequently leading to biased predictions $\hat{y}$. To mitigate such indirect bias in model explanations, we design a similarity-based regularization term R to constrain the model to only rely on the key prediction centric tokens $\boldsymbol{t}^y$ but not the sensitive context tokens $\boldsymbol{t}^s$.

To obtain this similarity regularization term R, first, we need a pre-trained helper model $g : \boldsymbol{x} \to s$ (same as the one from IBD). During the training of our f model, we take the attention matrix A_f from model f and the attention matrix A_g from g model corresponding to the same samples to calculate the cosine similarity between these two matrices using Equation 5.

$$R = (cos(A_f, A_g))^2 . \tag{5}$$

A greater term R indicates the model f relies on the sensitive context tokens $\boldsymbol{t}^s$ similarly to g. The preference for cosine similarity over Jaccard similarity is attributed to its differentiable nature, which is conducive to gradient-based optimization.

To achieve no indirect bias in model explanation, the model f is trained with the total loss function $\mathcal{L}$ in Equation 6, where we add the similarity regularization term R to the cross-entropy $CE(f(\boldsymbol{x}), y)$.

$$\mathcal{L} = CE(f(\boldsymbol{x}), y) + \lambda R, \tag{6}$$

where λ is a hyper-parameter that controls the trade-off for fair explanations.

Our similarity regularization only aims to remove indirect bias in model explanations. It cannot guarantee the prediction outcome fairness, because the layers after self-attention in the transformers may still exploit the bias in the training data. In practice, it is better to complement direct bias mitigation for traditional outcome fairness with indirect bias mitigation in model explanation. In our evaluation, we show that our indirect bias mitigation is compatible with the most popular in-process mitigation for demographic parity - adversarial debiasing (AD) (Zhang et al., 2018), thus simultaneously achieving both demographic parity in predictions and no indirect bias in model explanations.

7 Experiment

In this section, we evaluate our proposed Indirect Bias Discovery (**IBD**) and Indirect Bias Mitigation (**IBM**) algorithms on sentiment analysis and toxicity detection datasets. Through case studies, we also demonstrate the significance of indirect bias in model explanations and the advantage of mitigating indirect bias.

7.1 Metrics

We use **Accuracy** to evaluate the classification utility performance, as our datasets are relatively balanced. There is a trade-off between utility and fairness. When the same level of fairness is met, the higher utility indicates a better trade-off in the mitigation model.

For classification fairness, we evaluate both on the predicted outcome and the model's local explanations. We use **Risk Difference (RD)** to evaluate the demographic parity in model predictions, where $RD = P(\hat{y} = c|s = u) - P(\hat{y} = c|s = v)$. A low-risk difference indicates fairness in terms of demographic parity in the model predictions.

We use aggregated attention for model explanations and evaluate the indirect bias in model explanations using our proposed metric - **Area Under Similarity Curve (AUSC)**, which is based on the Jaccard similarity defined in Section 5. A higher value of AUSC indicates high indirect bias in the model's local explanations, where the model over-exploits sensitive context tokens in its decision-making process. In addition, we further examine the model explanations with the similarity curve (also defined in Section 5). A curve below the diagonal line indicates no indirect bias in model explanations.

7.2 Datasets

The **Amazon Books Review Dataset**[1], contains feedback from 3 million users on 212,404 unique books. Using a gender inferencing model, a subset of 16,927 users (9,105 male users and 7,822 female users) was identified with high confidence based on common male and female names. This results in a subset of 33,600 reviews (16,965 positive reviews and 16,635 negative reviews), where those rated with 4 or 5 stars were classified as positive and 1-star reviews as negative. The dataset has a risk difference of ~20%, where female users make more positive reviews. The protected attribute in this dataset is the review author's (inferred) gender. Most reviews do not include a gender self-identification token in them.

The **Jigsaw Unintended Bias in Toxicity Dataset** (cjadams et al., 2019) is an archive of approximately 2 million public comments, was released at the end of 2017 following the shutdown of the Civil Comments platform. It was labeled for both the toxicity of the comments and the presence of several protected attributes. A targeted subset of this dataset, labeled specifically for toxicity towards male and female identities, comprised 21,000 records. Within this subset, 13,000 records were associated with male identities and 8,000 with male identities. The comments were classified based on toxicity levels, with 10,490 identified as toxic and 10,510 as non-toxic. The dataset has a risk difference of ~20%, where the ratio of toxic comments towards females is higher.

Both the datasets are split into 82.8% training, 7.2% validation, and 10% testing.

7.3 Models

There is no previous work on indirect bias mitigation on model explanations. We compare our indirect bias mitigation method with some mitigation methods that focus on achieving demographic parity in predictions.

The **Vanilla Model** (Devlin et al., 2018) is a Bert Model with no fairness mechanism built in. We fine-tune the uncased BERT-base model from HuggingFace. It is highly likely to inherit the bias in the training data. It should have a higher accuracy along with a high-risk difference.

Resampling (Kamiran and Calders, 2011) is pre-processing mitigation, which resamples the biased dataset to get an unbiased dataset with a close to 0 risk difference. The sampled unbiased dataset is then used for model training (for a vanilla Bert model) instead of the original training data. However, such a pre-processing method cannot achieve fairness in model predictions when it is evaluated in the original test data.

Adversarial Debiasing (AD) (Zhang et al., 2018) is an in-processing mitigation, which uses adversarial learning to remove the correlation between the predicted outcome and the protected attribute, i.e., achieving demographic parity in predictions. The adversary network is a standard feed-forward network containing two hidden layers with 512 and 128 units with *ReLU* activation function. The output layer of the adversary has a sigmoid activation. The hyperparameter to control the adversary strength is 20. We evaluate whether mitigation for demographic parity also leads to fairness in model explanations.

Our proposed method is to add similarity regularization for indirect bias mitigation on top of adversarial debiasing (**AD + IBM**). The helper model g is a vanilla Bert model trained on the same training data. The hyperparameter λ in Equation 6 to control the regularization strength is 200. **Ours** aims to achieve both demographic parity and no indirect bias. It trades off utility to satisfy both metrics.

7.4 Performance Comparison

Table 1 shows the main result of our evaluation. The four models (Vanilla, Resampling, AD, and Ours) are evaluated on the two datasets.

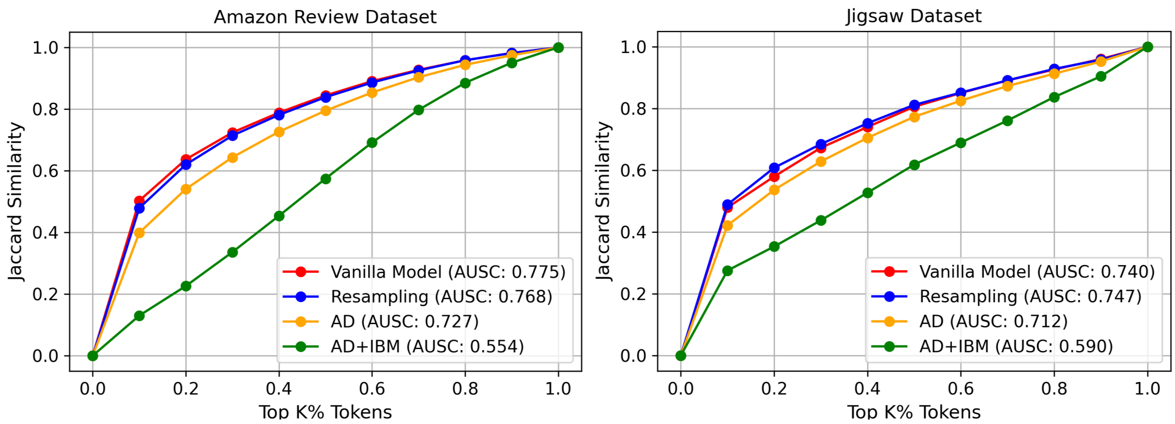
Demographic Parity. For both datasets, as expected, neither the vanilla model nor resampling can achieve low-risk difference in the prediction on testing data. Both AD and Ours achieve low-risk differences through adversarial learning.

Indirect Bias Discovery and Mitigation. The result on AUSC shows that our proposed Indirect Bias Discovery (IBD) algorithm is effective in quantifying the indirect bias in model explanations. For both datasets, the vanilla model, resampling and AD all have high AUSC scores (above 0.7), which means their explanations have indirect bias w.r.t. the protected attribute. There is a slight correlation between RD and AUSC for these models with unconstrained model attention. For our Indirect Bias Mitigation (IBM) algorithm, the similarity regularization makes sure the model learns different patterns from the gender inference (helper) model. Our model explanation has a close to 0.5

[1] Amazon Books Reviews Dataset

Model	Amazon Review Dataset			Jigsaw Dataset		
	Accuracy	RD	AUSC	Accuracy	RD	AUSC
Vanilla Model	0.936	0.194	0.775	0.843	0.192	0.740
Resampling	0.929	0.184	0.768	0.848	0.163	0.747
AD	0.762	0.074	0.727	0.792	0.030	0.712
AD + IBM (Ours)	0.724	0.082	0.554	0.761	0.033	0.590

Table 1: Model Performance on Different Datasets



(a) Amazon Review Dataset

(b) Jigsaw Dataset

Figure 3: Similarity Curve Comparison

AUSC, indicating low indirect bias, i.e., the model only focuses on ground-truth-centric tokens.

We can further compare the model explanation using the similarity curve. Figure 3a and 3b shows the similarity curve for each model on the Amazon review dataset and Jigsaw dataset, respectively. For both datasets, the Vanilla Model curve (red) and the resampling curve (blue) are close to each other. The AD curve (yellow) is slightly under the other two. However, all three of them have a clear arch, which indicates high similarity and high indirect bias. The curve for our proposed IBM model (green) is close to a diagonal line, which is expected for the goal of no indirect bias in model explanations.

Utility Trade-off. We know there is a utility trade-off for fairness in machine learning. The accuracy difference between the vanilla-biased model and the AD unbiased one indicates the trade-off for demographic parity through AD. The trade-off is 0.174 for the Amazon review dataset and 0.051 for the Jigsaw dataset. This means bias mitigation is more difficult for the Amazon review dataset because the sensitive token is not available to the model. This confirms our motivation to mitigate NLP bias beyond direct bias. For indirect bias, a small additional trade-off for no indirect bias is required. The trade-off is 0.038 and 0.031 for the Amazon review dataset and the Jigsaw dataset, respectively. The trade-off is relatively small.

7.5 Case Analysis

To further showcase the significance of indirect bias and the advantage in its mitigation, we also conduct case analysis to directly compare different model explanations on individual examples. Figure 4 shows the explanations provided by different models on selected examples. Due to limited space, full model explanations on long texts are included in the Appendix.

Case (a) is a toxic comment towards males from the Jigsaw dataset. All models except for AD correctly predicted the toxicity. The explanations from vanilla and resampling are "men", "jealous", and "fertility". The explanation from our AD+IBM model relies on "dominance", "because", and "jealous", which is a gender-neutral toxicity logic. AD has a similar explanation but the model failed the prediction. We can also discover the indirect bias from these individual explanations. The vanilla model, resampling, and AD have AUSC 0.628, 0.646, and 0.544, respectively. Our AD+IBM only has 0.503 AUSC, which indicates the lowest indirect bias.

Case (b) is a toxic comment towards females generated by ChatGPT 4. The toxicity context is too subtle that the vanilla model, resampling, and AD model cannot make the correct prediction for it. They all heavily focus on "men". They

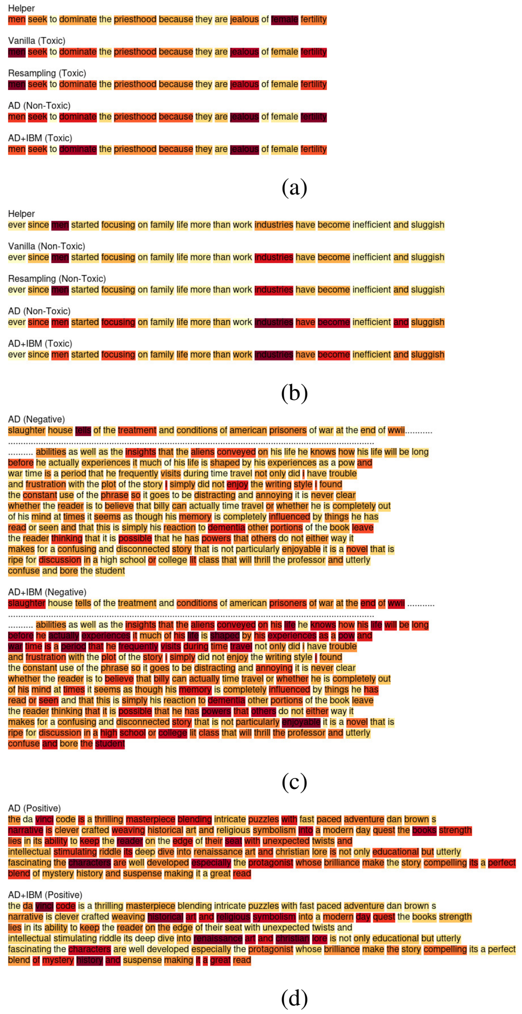


Figure 4: Model explanations on the example cases

associate "men" with non-toxicity, therefore failing the detection. Only our AD+IBM model correctly identified the toxicity. It focuses less on "men", "focusing" and "family life". The toxicity is on the absent female group, where female is "inefficient and sluggish" in "industries". We can further verify our observation on model explanation with AUSC scores. For this case, the explanations from Vanilla model, Resampling, and AD have AUSC scores of 0.816, 0.754, and 0.609, respectively. Ours has 0.543 AUSC, indicating low indirect bias.

Case (c) is a negative review by a female author from the Amazon review dataset. All models correctly predicted the negative sentiment. The explanations from Vanilla, resampling, and AD put more emphasis on topic words (e.g., "story", "style", "dementia", etc.), which are the topics more likely from a female review as suggested by the helper model. For our AD+IBM model, the explanation focuses more on the sentiment-related content (e.g., "not particularly enjoyable", "thrill the professor", "confuse and bore the student", etc.). This means our mitigator avoids potential sensitive context and focuses only on ground-truth-centric tokens. The indirect bias discovered in the AUSC score for Vanilla, resampling, and AD are 0.778, 0.787, and 0.724, respectively. Ours only has 0.575.

Case (d) is a positive review by ChatGPT 4, which is instructed to write a review from a female perspective without revealing they are female. The generated review contains subtle bias inherited from historical data. ChatGPT also provides its justification that the review focuses more on the female characters, including the main protagonist - Sophie Neveu. The helper model suggests that "narrative" and "characters" are associated with female reviewers. In comparison to the other models, the explanation from our AD+IBM model focuses more on the sentiment words (e.g., "keep the reader on the edge", "great", etc). However, the model still suffers from spurious correlations outside of gender bias, such as "historical", "religious", "renaissance", "christian", etc. This is because the model is not trained to mitigate these spurious correlations. For the AUSC scores, Vanilla, resampling, and AD are 0.744, 0.715, and 0.640, respectively. Ours has a low AUSC score of 0.445.

Overall, indirect bias is difficult for AD to mitigate, especially in subtle, complex, and long-text cases. IBD can quantify the indirect bias in the form of AUSC score. Our AD+IBM mitigation is effective in providing neutral unbiased local explanations for all cases.

8 Conclusion

In this work, we study indirect bias in NLP models, a phenomenon less explored but as significant as direct bias. Our contributions include defining direct versus indirect bias, introducing a new framework for quantitatively evaluating indirect bias in transformer models using their in-built self-attention matrix and proposing a mitigation algorithm to ensure fairness in transformer models by leveraging attention explanations. Our evaluation shows the significance and challenging nature of indirect bias in model explanations, and the effectiveness of our proposed discovery and mitigation algorithms. These efforts represent a critical step towards achieving fairness and equity in NLP applications, addressing current research gaps, and guiding future ethical AI development.

9 Limitations

There is no publicly available dataset designed to study indirect bias. For the experiment evaluation, it is challenging to identify the ground truth-sensitive context. The current evaluation of the data we have is not enough to showcase the full spectrum of indirect bias. Our methodology heavily relies on a helper model to infer sensitive attributes. The quality of the helper model hinders the performance of our bias discovery and mitigation algorithm. The need for a helper model also slows down the runtime efficiency. In future work, we will develop a method only utilizing the target model's explanations.

10 Ethical Considerations

This study aims to improve NLP technology to achieve equity for all under-served communities. We want to broaden the scope of NLP fairness. Developing fair and explainable NLP models can free technology from inheriting historical bias in real-world data. Due to the limited options on datasets, we conducted the experiment with a simplified binary setting. The proposed technology is designed to comply with non-binary identities and multi-ethnicity. We hope this project raises awareness of the influence of unintentional bias from NLP models. It is a community effort to develop and advocate open-source, transparent, fair, accountable, and explainable NLP models.

References

Giuseppe Attanasio, Debora Nozza, Dirk Hovy, and Elena Baralis. 2022. Entropy-based attention regularization frees unintended bias mitigation from lists. In *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 1105–1119. Association for Computational Linguistics.

Rajas Bansal. 2022. A survey on bias and fairness in natural language processing. *CoRR*, abs/2204.09591.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in nlp: The case of india.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard S. Zemel. 2019. Understanding the origins of bias in word embeddings. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 803–811. PMLR.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

cjadams, Daniel Borkan, inversion, Jeffrey Sorensen, Lucas Dixon, Lucy Vasserman, and nithum. 2019. Jigsaw unintended bias in toxicity classification.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Trans. Assoc. Comput. Linguistics*, 9:1249–1267.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc Viet Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2978–2988. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–14, New York, NY, USA. Association for Computing Machinery.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online,* *November 16-20, 2020*, pages 314–331. Association for Computational Linguistics.

Brian d'Alessandro, Cathy O'Neil, and Tom LaGatta. 2017. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2):120–134.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5491–5501. Association for Computational Linguistics.

Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, 33(1):1–33.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, *SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018*, pages 43–53. Association for Computational Linguistics.

Gaël Letarte, Frédérik Paradis, Philippe Giguère, and François Laviolette. 2018. Importance of self-attention for sentiment analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 267–275, Brussels, Belgium. Association for Computational Linguistics.

Adian Liusie, Vatsal Raina, Vyas Raina, and Mark J. F. Gales. 2022. Analyzing biases to spurious correlations in text classification tasks. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 2: Short Papers, Online only, November 20-23, 2022*, pages 78–84. Association for Computational Linguistics.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875.

Ninareh Mehrabi, Umang Gupta, Fred Morstatter, Greg Ver Steeg, and Aram Galstyan. 2022. Attributing fair decisions with attention interventions. In *Proceedings of the 2nd Workshop on Trustworthy Natural Language Processing (TrustNLP 2022)*. Association for Computational Linguistics.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35.

Nikolaos Mylonas, Ioannis Mollas, and Grigorios Tsoumakas. 2022. An attention matrix for every decision: Faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classification.

Leonid Schwenke and Martin Atzmueller. 2021. Show me what you're looking for visualizing abstracted transformer attention for enhancing their local interpretability on time series data. *The International FLAIRS Conference Proceedings*, 34.

P Sunilkumar and Athira P Shaji. 2019. A survey on semantic similarity. In *2019 International Conference on Advances in Computing, Communication and Control (ICAC3)*, pages 1–8. IEEE.

Supreme Court of the United States. 1971. Griggs v. duke power co. 401 U.S. 424. March 8.

Mike Thelwall. 2018. Gender bias in sentiment analysis. *Online Inf. Rev.*, 42(1):45–57.

Ashish Vaswani, 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.

Pranav Narayanan Venkit and Shomir Wilson. 2021. Identification of bias against people with disabilities in sentiment analysis and toxicity detection models.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. 2022. Identifying and mitigating spurious correlations for improving robustness in NLP models. In *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 1719–1729. Association for Computational Linguistics.

Yaushian Wang, Hung-Yi Lee, and Yun-Nung Chen. 2019. Tree transformer: Integrating tree structures into self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1061–1070, Hong Kong, China. Association for Computational Linguistics.

CS Webster, S Taylor, C Thomas, and JM Weller. 2022. Social bias, discrimination and inequity in healthcare: mechanisms, implications and recommendations. *BJA Educ*, 22(4):131–137.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media, SocialNLP@ACL 2020, Online, July 10, 2020*, pages 7–14. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 335–340. ACM.

A Appendix

A.1 More model explanations on case analysis

Due to limited space, we only included the explanations from AD and AD+IBM for the long review cases in Section 7.5. Figure 5 and 6 show the full explanations from all evaluated models.

Figure 5: All model explanations on Case (c)

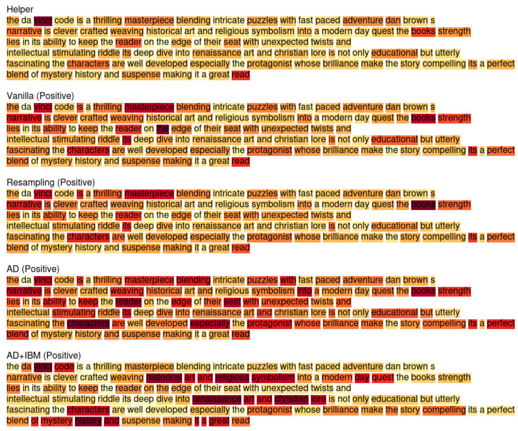


Figure 6: All model explanations on Case (d)