Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention

Yacine Gaci, Boualem Benattallah, Fabio Casati, Khalid Benabdeslem

Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention

Yacine Gaci[1], **Boualem Benatallah**[2,3], **Fabio Casati**[4], **Khalid Benabdeslem**[1]
[1] LIRIS - University of Lyon 1, France [2] Dublin City University, Ireland
[3] UNSW Sydney, Australia [4] ServiceNow, USA
{yacine.gaci, khalid.benabdeslem}@univ-lyon1.fr
{boualem.benatallah, fabio.casati}@gmail.com

Abstract

Natural Language Processing (NLP) models are found to exhibit discriminatory stereotypes across many social constructs, e.g. gender and race. In comparison to the progress made in reducing bias from static word embeddings, fairness in sentence-level text encoders received little consideration despite their wider applicability in contemporary NLP tasks. In this paper, we propose a debiasing method for pretrained text encoders that both reduces social stereotypes, and inflicts next to no semantic damage. Unlike previous studies that directly manipulate the embeddings, we suggest to dive deeper into the operation of these encoders, and pay more attention to the way they pay attention to different social groups. We find that stereotypes are also encoded in the attention layer. Then, we work on model debiasing by redistributing the attention scores of a text encoder such that it forgets any *preference* to historically advantaged groups, and attends to all social classes with the same intensity. Our experiments confirm that reducing bias from attention effectively mitigates it from the model's text representations.

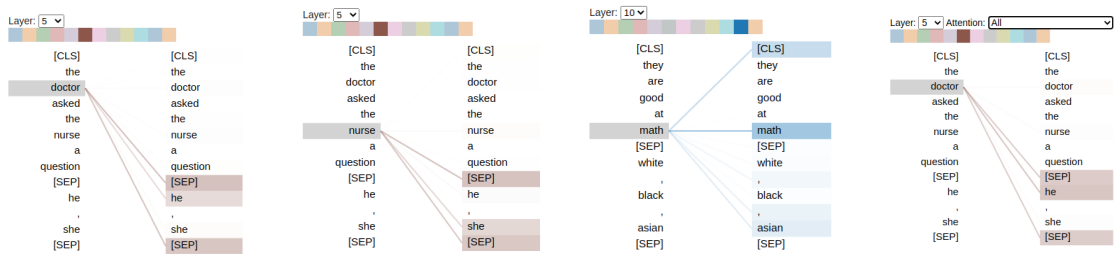
1 Introduction

Natural Language Processing (NLP) is increasingly penetrating real-world processes such as recruitment (Hansen et al., 2015), legal systems (Dale, 2019), healthcare (Velupillai et al., 2018) and Web Search (Nalisnick et al., 2016). Part of this success is attributed to the underlying embedding layer which encodes sophisticated semantic representations of language (Camacho-Collados and Pilehvar, 2018). However, this wide adoption has called into question the fairness aspect of modern text encoders. Recent research exposed them for replicating discriminatory social biases which may cause unintended and undesired model behaviors with respect to social groups (Bolukbasi et al., 2016; Caliskan et al., 2017; May et al., 2019). As an example, a resume filtering system built on top of such biased encoders display overly improper preferences towards male applicants (Dastin, 2018).

Methods to debias text encoders have recently been proposed, ranging from Counterfactual Data Augmentation (CDA) (Webster et al., 2020), projection on bias-free subspaces (Kaneko and Bollegala, 2021), contrastive learning (Cheng et al., 2020), adversarial attacks (Wang et al., 2021) or simply extending existing debiasing techniques from static word embeddings (Liang et al., 2020a; Bolukbasi et al., 2016; Manzini et al., 2019; Gaci et al., 2022a). However, these methods have shown mixed results, often failing to reduce the amount of bias to a satisfactory degree (Gonen and Goldberg, 2019; Blodgett et al., 2020; Meade et al., 2021). We argue that part of this shortcoming owes to the fact that current debiasing techniques, by operating exclusively on the model's embeddings, are not removing bias entirely. In this paper, we propose that some biases can also be encoded in the attention mechanism, and these stay relatively out of reach for methods that do not manipulate attention directly.

To illustrate how biases are reflected in attention, we show some attention heads of BERT (Devlin et al., 2018) in Figure 1[1]. Consider the following sentence *"The doctor asked the nurse a question."* Aiming to analyze how every word representation relates to different demographics, we add a dummy second input consisting of words representing distinct groups (e.g. ***he*** and ***she*** after the [SEP] token[2]). Figure 1(a) illustrates that *doctor* pays much more attention to *he* than to *she*[3], while Figure 1(b) reveals that *nurse* attends to *she*. This finding suggests that gender stereotypes are encoded in attention weights. Likewise, in Figure 1(c), *math* is more related to *asian* than to *white* or *black*,

[1]The figures are produced using bertviz (Vig, 2019a)
[2]BERT uses [SEP] token to separate the sentences in two-sentence inputs
[3]The darker the color, the higher the attention weight is



(a) (b) (c) (d)

Figure 1: Attention patterns in modern text encoders suggest the existence of potential gender and racial biases. Figures (a), (b) and (c) correspond to the base version of BERT, while Figure (d) corresponds to BERT after applying the debiasing method proposed by Kaneko and Bollegala (2021)

conforming to the famous racial stereotype casting asians as good mathematicians (Trytten et al., 2012; Shah, 2019). More intriguing is Figure 1(d), where we find that even after applying the bias reduction method proposed by Kaneko and Bollegala (2021), gender bias is still reflected in the attention of the supposedly debiased text encoder. We experimented with other debiasing methods and found similar results. These examples convey that biases can be hidden in the attention mechanism, and thus pose the risk of being recovered in representations and predictions.

In this paper, we propose a novel bias measure based on attention weights in order to quantify the amount of bias encoded in attention heads. We show that modern text encoders display substantial amounts of bias in their attention components. Also, we quantitatively show that current debiasing methods do very little to mitigate social stereotypes, and merely conceal them in the attention layer. Then, we propose Attention-Debiasing (***AttenD***), an attention-based debiasing approach which works as follows: Given that attention weights conform with undesired biases (e.g., *doctor* attending to ***he***, and *nurse* to ***she*** in Figure 1), we finetune the parameters of the text encoder of interest such that it learns to produce equal attention scores for every word in the input sentence with respect to social groups. Returning to the example of Figure 1, AttenD redistributes attention scores of *doctor* such that it attends to ***he*** and ***she*** with the same intensity, thus eliminating any preference toward one of the groups. However, alterations to the attention of *doctor* on the remaining words of the input sentence must be kept to a minimum in order not to corrupt the semantic understanding of the original text encoder. To do that, we distill the original attentions from an unaltered teacher text encoder (Hinton et al., 2015; Gou et al., 2021). In this setting, we encourage the debiased model to copy the original attention from its teacher to minimize semantic offset.

We suggest that by equalizing attention weights, text encoders forget biased associations between groups and attributes. Thus, for a given input sentence, whether it mentions a man or a woman, a muslim or a christian, the model's attention on the mentioned group is the same, which leads to very close text representations, and hence identical predictions. In our experiments, we address gender, racial and religious biases, and show that AttenD not only reduces social stereotypes from the attention layer itself, but also from representations and predictions when applied to the tasks of textual inference and hate speech detection. We release our code and data on GitHub.[4]

2 Related Work

2.1 Bias Quantification

To date, there are three main approaches to detect stereotypes in text encoders: (1) **representation-based**: where vector relationships between different types of inputs are measured. For example, Caliskan et al. (2017) and May et al. (2019) compared the cosine similarity between representations of groups and attributes, and found that groups have unequal similarities. (2) **likelihood-based**: These approaches examine how often text encoders prefer stereotypes over anti-stereotypes. Preferences in this case are defined in terms of higher likelihoods as produced by language models using embeddings of the text encoders under study (Kurita et al., 2019; Nadeem et al., 2020; Nangia et al., 2020; Gaci et al., 2022b). (3) **inference-**

[4] https://github.com/YacineGACI/AttenD

based: These methods employ text encoders in downstream NLP tasks (Blodgett et al., 2020) such as natural language inference (Dev et al., 2020), sentiment analysis (Díaz et al., 2018) or language generation (Sap et al., 2020; Sheng et al., 2020). Bias in such settings is declared as the difference in outcome when the models are tested with the same input sentence, differing only in social groups. In contrast, we propose a new metric that quantifies bias that is specifically encoded in the attention layer.

2.2 Bias Reduction

The NLP community has produced a wealth of methods to reduce bias from static word embeddings, spanning diverse techniques such as projections on bias-free dimensions (Bolukbasi et al., 2016; Manzini et al., 2019; Kaneko and Bollegala, 2019; Kumar et al., 2020; Ravfogel et al., 2020), adversarial attacks (Xie et al., 2017; Li et al., 2018; Elazar and Goldberg, 2018; Gaci et al., 2022a), or training from scratch with additional fairness constraints (Zhao et al., 2018). However, debiasing large-scale text encoders has started attracting the community's attention only recently. A lot of effort has focused on extending existing techniques to work for large text encoders. For instance, Liang et al. (2020a) contextualize words into sentences by sampling them from existing corpora before applying the methods of Bolukbasi et al. (2016). Kaneko and Bollegala (2021) minimizes the projection of sentence representations on a *learned* bias subspace, while Qian et al. (2019); Bordia and Bowman (2019); Liang et al. (2020b) add bias-reduction objectives to their loss functions. Another line of research uses CDA (Webster et al., 2020) to balance gender correlations in training data, while Lauscher et al. (2021) uses adapters to reduce the large training time that CDA incurs. Other debiasing techniques use contrastive learning (Cheng et al., 2020), zero-shot learning (Schick et al., 2021), dropout (Meade et al., 2022), or regularizing the entropy of attention (Attanasio et al., 2022). This last work differs from ours in that they discourage the model from basing its classification on identity terms while attending to a wider context. On the other hand, our goal is to reduce harmful associations to (dis)advantaged groups by calibrating the attention of the context on identity terms. Besides, our method is applied to text encoders as a general representation layer, while the method of Attanasio et al. (2022) is proposed for hate-speech classification models.

2.3 Effect of Attention

Attention plays a central role in modern NLP systems (Wiegreffe and Pinter, 2019). Several techniques have used attention to dissect and explain the inner functioning of text encoders (Vig, 2019b; Clark et al., 2019; Hoover et al., 2020; Tenney et al., 2020; Bastings and Filippova, 2020). However, recent studies argued that attention cannot be used as a reliable tool to explain the behavior of models (Jain and Wallace, 2019; Pruthi et al., 2020; Bastings and Filippova, 2020). In this work, we disagree with these anti-attention studies, and following the arguments of Wiegreffe and Pinter (2019), we emphasize that their limitations include:

(1) the experimental setup was particularly limited to recurrent architectures (RNNs). We believe that one cannot generalize the findings to all kinds of models that use attention, especially transformer-based models like the ones we use in this paper, that mainly constitute of attention layers (Vaswani et al., 2017; Devlin et al., 2018). (2) Whether attention explains or not depends on the definition of explainability one is looking for (Lipton, 2018; Rudin, 2019). Although Jain and Wallace (2019) casts some doubt on the notion that attention grants one true and faithful interpretation, we believe they do not invalidate the fact that attention does show which features are most meaningful to models (Wiegreffe and Pinter, 2019). (3) Our own experiments confirm that by reducing bias in attention, we find that it is also reduced in embeddings and predictions. This hints that attention *contributes* in the decision-making of text encoders.

3 Bias Quantification Using Attention

Despite the applicability of our work on any text encoder that is built upon self-attention, we focus in this paper on models based on the encoder side of the transformer architecture, such as BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), or ALBERT (Lan et al., 2019). This owes to the decoder side being usually used in auto-regression tasks, and less often to encode text.

In this section, we present our metric to compute the amount of bias that is encoded in attention across an entire corpus $\mathbb{S}$. First, we identify bias types of interest such as gender, race, religion. This is achieved by defining a set of tuples $\mathbb{G}$ for every

religion	gender
muslim, christian, jewish	he, she
quran, bible, torah	man, woman

Table 1: Examples of group tuples per bias

bias type such that $\mathbb{G} = \{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_k\}$ where each $\mathcal{T}_i$ describes social groups, or their attributes. Table 1 shows some possible values for $\mathcal{T}_i$ that we use in our study.[5] Then, for every sentence s in $\mathbb{S}$, we randomly pick a tuple $\mathcal{T}_i$ from $\mathbb{G}$ and construct s_g, an artificial sentence formed by words of $\mathcal{T}_i$. For example, given Table 1, s_g can be *"muslim, christian, jewish"* or *"man, woman"*. Finally, we use both s and s_g to make two-sentence inputs similar to the examples of Figure 1.

After augmenting the corpus with artificial group-related sentences, we feed each augmented input to the text encoder under study, and collect the resulting self-attention weights. Each token in the augmented input distributes its attention on all other tokens according to their importance. Thus, every group in s_g has its own attention allocation, i.e. the vector consisting of attention weights that tokens in s give to the current group token. We declare bias in this case as the difference between attention allocations of groups. In other words, if the sentence distributes its attention on social groups differently (e.g. *doctor* in Figure 1 attends to *he* and not to *she*), then there is bias. Specifically, we measure Pearson correlation between attention allocations of social groups in a given attention head, aggregated over a corpus:

$$Bias(\mathbb{S}, \mathbb{G}) = \frac{1}{|\mathbb{S}||\mathbb{G}|} \sum_{s \in \mathbb{S}} \sum_{s_g \in \mathbb{G}} \frac{1}{|\binom{\mathbb{G}}{2}|} \sum_{i,j \in \binom{\mathbb{G}}{2}} \rho(A_s^{g_i}, A_s^{g_j}) \tag{1}$$

where ρ is Pearson correlation, $\binom{\mathbb{G}}{2}$ produces all possible pairs of social groups given a tuple, $A_s^{g_i}$ is the attention vector that sentence s allocates to group g_i. If this quantity is close to 0, we can say that attention exhibits bias since the average of correlations across sentences and groups is nearly 0. We use the News-commentary-v15 corpus[6] as evaluation data and compute attention bias of BERT. We present the results for each attention head and for each bias type in Figure 2.

We observe that BERT's attention heads encode different stereotypes with different intensities, conforming to the findings of Bhardwaj et al. (2021). Also, the lower layers of attention appear to encode more bias than the top layers since their heads are much darker. We believe this to be the consequence of lower layers being more aware of the input tokens, while top layers are fed transformations of the input as it flows through the attention stack.

We also compute attention bias on BERT heads after applying different existing debiasing approaches, and present the results for gender bias in Figure 3. When we compare the heatmaps in Figure 3 to the heatmap of Figure 2(a), we notice that they are very similar, and attention bias is hardly removed. In some cases, it is even amplified (e.g. head 3 in layer 10 in Figure 3(a)). We stipulate that even though those debiasing methods show acceptable results with embedding-based bias evaluations, they appear to ignore the bias reflected in attention and thus can be recovered at prediction. In the next section, we present our own debiasing methods that aims to reduce bias from the top k most biased attention heads.

4 Debiasing Method

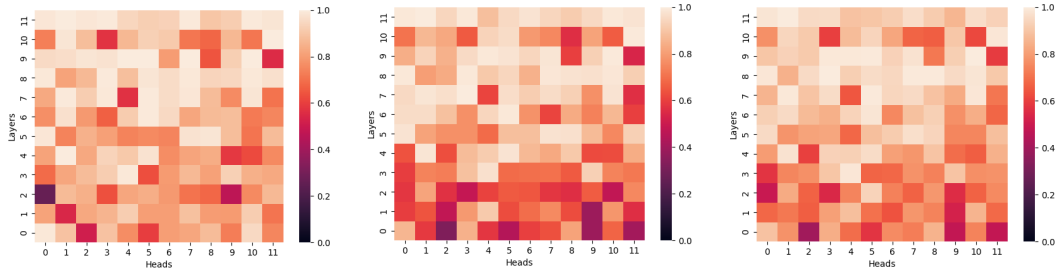
The first step of AttenD is augmenting each sentence s in the training corpus $\mathbb{S}$ with an artificial second input s_g consisting of words related to groups of a given bias type, as explained in Section 3 and in the examples of Figure 1. Then, we finetune the encoder's parameters such that the top k most biased heads produce equalized attentions on groups, i.e. each token in s pays the same amount of attention to tokens of s_g, thus eliminating preferences and stereotypes. We minimize semantic loss by compelling the model to learn the original semantics from an *unaltered* teacher model by copying its internal attention.

We schematize the operation of AttenD in Figure 4. Gr_1, Gr_2 and Gr_n in the figure correspond to the tokens of s_g. Both matrices represent one attention head of the text encoder before (left) and after (right) debiasing. The matrices should be read in rows. Each row depicts the attention weights of the corresponding token on all the other tokens of the input (s + s_g)[7]. The matrices are conceptually split in four blocks: (1) attentions of s on s, (2) attentions of s on s_g, (3) attentions of s_g on s,

[5] The full list can be found in Appendix A.2

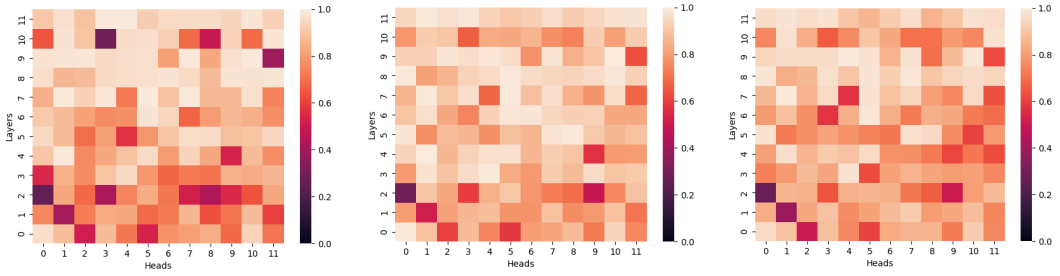
[6] http://www.statmt.org/wmt20/translation-task.html

[7] [CLS] token (vector representation of s) is also included for attention calibration. Details in Appendix A.6



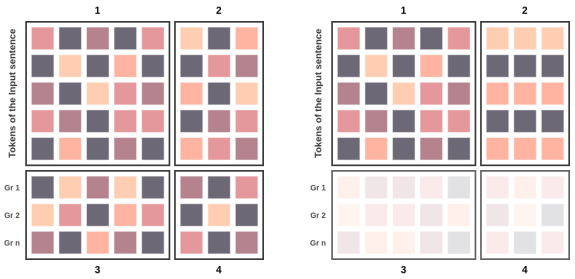
(a) (b) (c)

Figure 2: Attention bias in BERT base broken out by layer and head. (a) gender, (b) race, (c) religion



(a) (b) (c)

Figure 3: Attention bias in BERT base broken out by layer and head after the application of several debiasing methods to reduce gender bias. These debiasing methods are (a) CDA (b) Sent-D (Liang et al., 2020a), (c) Kaneko Kaneko and Bollegala (2021)



(a) (b)

Figure 4: Overview of an attention head before (a) and after (b) debiasing

and (4) attentions of s_g on s_g. Debiasing consists in making the columns of block 2 equal. In other words, each token in s pays the same amount of attention to all the groups as indicated in the right side of Figure 4. We preserve the semantics of the original text encoder by keeping block 1 of Figure 4 unchanged. Both blocks 3 and 4 are irrelevant to the results, since they denote attentions of our artificially inserted second input s_g. So, we do not impose any restrictions on them. In the following, we describe the important details of AttenD.

4.1 Equalizing attentions on social groups

The rationale behind attention equalization is to eliminate any inclination for the encoder to prefer any social group to the detriment of others. Equalizing attention vectors of block 2 (as defined in Figure 4) is equivalent to making them equal to a pivot vector. In our method, we consider the attention vector of s on the first social group as the pivot (first column in block 2 of Figure 4), and minimize the mean square error between the pivot and the attention vectors of s on the other groups, one at a time. Suppose $\mathbf{A}^{l,h,s,s_g} = Attn(s, s_g; l, h)$ is the attention matrix at layer l, head h of the encoder E, computed from the input $s + s_g$. The equalization loss is given by Equation 2.

$$L_{equ} = \sum_{s \in \mathbb{S}} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{i=2}^{||s_g||} ||\mathbf{A}^{l,h,s,s_g}_{:\sigma,\sigma+1} - \mathbf{A}^{l,h,s,s_g}_{:\sigma,\sigma+i}||_2^2 \tag{2}$$

where L is the number of layers of the text encoder, H the number of heads, $||s_g||$ the number of social groups in s_g and σ is the position of the special token [SEP] that marks the end of s and the beginning of s_g. As can be seen, $\mathbf{A}^{l,h,s,s_g}_{:\sigma,\sigma+1}$ is the

pivot vector containing attention scores of s on the first social group token (whose position is directly after [SEP], i.e., $\sigma+1$). Equation 2 forces attention scores on subsequent social groups to be the same as on the first one, thus making them all equal. We also experiment with choosing the last group as pivot, or pick one at random. We find that these alternatives produce comparable results.

4.2 Preserving semantic information

We minimize semantic information loss in a knowledge distillation setting, where we recruit another model to be the *teacher*, and cast the text encoder that we want to debias as the *student* (Hinton et al., 2015; Gou et al., 2021). We initialize the student from the teacher. We do not apply our debiasing strategy on the teacher since it provides a reference to the original unaltered language representations. We compel the student to copy the teacher's attention for every input in the training corpus $\mathbb{S}$.

As in Section 4.1, let $\mathbf{A}^{l,h,s,s_g}$ be the attention of the student model at layer l, head h with s and s_g as input. Likewise, let $\mathbf{O}^{l,h,s,s_g}$ define the teacher's attention matrix. We formalize the preservation of semantic information as a regularizer where we minimize the squared l_2 distance between the student's and the teacher's attention scores.

$$L_{distil} = \sum_{s\in\mathbb{S}}\sum_{l=1}^{L}\sum_{h=1}^{H} ||\mathbf{A}^{l,h,s,s_g}_{:\sigma,:\sigma} - \mathbf{O}^{l,h,s,s_g}_{:\sigma,:\sigma}||_2^2 \quad (3)$$

As can be seen from Equation 3, the student learns only to replicate block 1 (as in Figure 4) of the attention matrices. This is because block 1 contains attention scores of the original input sentence s on itself, thus encoding an important aspect of semantics. We force the student not to reproduce the attention distribution on social groups (block 2) from the teacher since these are supposedly biased, and are left to the care of our debiasing objective. We do not use the MLM loss since the teacher model is already trained using that objective. We describe the overall training objective as a linear combination of the previously defined losses, with λ as a hyperparameter to control the weight of debiasing over semantic preservation.

$$Loss = L_{distil} + \lambda L_{equ} \quad (4)$$

4.3 Negative Sampling

While learning to equalize attention on social groups that constitute the second half of the input, the text encoder bears the risk of distributing its attention uniformly on *any* second half, no matter what it is. This is particularly alarming when the text encoder is subsequently employed in double-sentence tasks (Wang et al., 2018) such as semantic textual similarity or sentence entailment.

To overcome the above obstacle, we introduce negative sampling. Instead of using words related to social groups in order to generate the artificial second input s_g, we randomly sample words (negative examples) from the vocabulary. In this case, we do not equalize the attentions but compel the student to copy its teacher even for blocks 2, 3 and 4. We do this in order to prevent the text encoder from learning to assign the same attention weight to all tokens of the second input when these do not define social groups. We control the ratio of negative examples with a hyperparameter η.

5 Evaluation

In this section, we first describe our experimental setup, then evaluate AttenD from two viewpoints: *fairness* and *representativeness*. Fairness is traditionally evaluated with two types of metrics: *intrinsic* metrics that measure bias in text representations regardless of their application, and *extrinsic* metrics that quantify bias in downstream tasks that text representations enable. We acknowledge that intrinsic metrics have recently been criticized (Goldfarb-Tarrant et al., 2020; Aribandi et al., 2021; Blodgett et al., 2021). However, we believe that a strong evaluation of bias should include both intrinsic, extrinsic and qualitative methods to draw a comprehensive evaluation. Since Aribandi et al. (2021) surmise that StereoSet and Crows-Pairs are more stable than other intrinsic measures of bias (e.g. WEAT (Caliskan et al., 2017) or SEAT (May et al., 2019)), we use them in this work. For extrinsic metrics, we evaluate our method on the tasks of textual inference and hate speech detection. Due to space limitations, we move the second one to Appendix A.3, in addition to other qualitative evaluations and ablation studies[8].

5.1 Debiasing setup

To facilitate comparison, we follow existing literature (Nadeem et al., 2020; Nangia et al., 2020) in defining social groups for each type of bias, although the approach presented here is not restricted

[8]We make our code and data available at https://github.com/YacineGACI/AttenD

Model	gender	race	religion	Overall
BERT	84.52	79.65	82.57	82.25
Sent-D	86.45	79.27	82.98	82.90
Kaneko	82.71	75.34	78.13	78.73
CDA	82.50	73.92	78.39	78.61
AttenD	**93.85**	**93.64**	**93.85**	**93.78**

Table 2: Attention bias on BERT before and after applying debiasing methods, the higher the better

to that, and can be leveraged for both other kinds of biases and for a more inclusive definition of the groups. In the experiments, we show results of debiasing based on **(binary) gender** (*male*, *female*), **race** (*white*, *black*, *asian*, *hispanic*) and **religion** (*muslim*, *christian*, *jewish*, *buddhist*). We leverage the definition words from previous work (Liang et al., 2020a). The full list can be found in Appendix A.2. We apply AttenD on BERT[9] (Devlin et al., 2018), and use the News-commentary-v15 corpus[10] as training data. It contains 223,153 sentences of which we use 80% for training and 20% for development. As for the top k most biased heads, we find that debiasing all heads works best in practice for all text encoders under study. Thus, we set the top k to the maximum number of heads for every model. Appendix A.1 contains details about hyperparameter search.

5.2 Evaluations of Fairness

5.2.1 Intrinsic Evaluation

We start by quantifying the amount of attention bias described in Section 3 in BERT base before and after applying AttenD. For accurate comparisons against previous work, we decided to include the baselines whose final debiased models have been published in order to avoid errors of training and/or tuning hyperparameters. Thus, we compare AttenD against (1) Sent-D (Liang et al., 2020a) which extended the method of *Hard-Debias* (Bolukbasi et al., 2016) to work on transformer-based text encoders, (2) the debiasing procedure proposed by Kaneko and Bollegala (2021) that finetunes the text encoder to minimize the projection of its embeddings on a predefined bias subspace, and (3) CDA. We also conduct a simple ablation study by training without negative examples (*AttenD*$^-$) when necessary. We use the development set of the News-commentary-v15 corpus to compute bias scores, and report the results in Table 2.

Although existing debiasing methods have been shown to reduce bias in embeddings (Liang et al., 2020a; Kaneko and Bollegala, 2021), we observe that they do very little to reduce it in attention. In fact, the method of Kaneko and Bollegala (2019) and CDA induce the model to encode more bias in the attention layer, which might make it to resurface in predictions if not addressed correctly.

In the following, we demonstrate that AttenD is also capable of mitigating bias from text representations and likelihoods. To do that, we finetune text encoders of interest before and after applying debiasing methods on the language modeling task. Then, we use the publicly available subsets of two stereotype benchmarks: StereoSet (Nadeem et al., 2020) and Crows-Pairs (Nangia et al., 2020). Both provide likelihood-based diagnostics to measure how often stereotypes are considered likelier than anti-stereotypes, given the language model's likelihoods. An ideal *unbiased* text encoder should score 50% in these benchmarks, i.e. it prefers neither stereotypes nor anti-stereotypes. Table 3 provides the evaluation results. StereoSet also provides a means to compute a language modeling (LM) score whose purpose is to check whether the encoder is still good at the task of language modeling, and that debiasing didn't hurt semantic performance.

We observe that AttenD shows impressive debiasing performance when evaluated with likelihood-based diagnostics. Improvements go up to 9.53% with a slight decrease in the accuracy of language modeling (-1.43%). We notice that our method yields the best results overall, and illustrates that reducing biases from attention directly helps with mitigating them from the model as a whole. Table 3 shows that using negative examples minimizes semantic information loss.

5.2.2 Extrinsic Evaluation

This approach of measuring bias builds on the intuition of Dev et al. (2020) stating that biased representations lead to invalid inferences, whose ratio quantifies bias. They construct a challenge benchmark for the natural language inference task where every hypothesis should be *neutral* to its premise. For example, suppose that the premise is *The **driver** owns a van* and the hypothesis is *The **man** owns a van*. The hypothesis neither entails nor contradicts the premise. If the predictions of a classifier deviate from neutrality, the underlying text encoder is

[9] In Appendix A.8, we also apply AttenD on ALBERT, RoBERTa, DistilBERT and SqueezeBERT

[10] http://www.statmt.org/wmt20/translation-task.html

Model	Crows-Pairs				StereoSet				LM
	gender	race	religion	Overall	gender	race	religion	Overall	
BERT	58.02	58.14	71.43	60.48	62.75	54.68	56.41	56.04	83.70
Sent-D	**51.53**	55.23	**60.0**	56.90	53.33	55.09	**51.28**	54.71	81.39
Kaneko	57.63	53.68	64.76	57.82	58.82	56.24	57.69	56.04	**85.58**
CDA	54.58	**50.78**	60.95	55.06	55.69	53.01	53.85	54.18	81.38
AttenD$^-$	53.05	53.68	69.52	57.23	**51.37**	54.37	55.13	53.37	80.92
AttenD	**51.53**	**50.78**	61.90	**54.58**	54.51	**52.29**	56.41	**53.18**	82.27

Table 3: Stereotype scores on BERT before and after applying debiasing methods. The closer to 50, the better. However, for the language modeling score (LM), the higher the better.

Model	Bias type	NN	FN	τ:0.5	τ:0.7
BERT	gender	36.38	36.45	36.06	33.96
	race	75.96	76.57	76.51	74.91
	religion	43.47	43.55	43.45	41.77
Sent-D	gender	44.74	45.10	44.54	42.06
	race	59.61	59.28	59.20	56.22
	religion	29.64	29.08	29.02	27.24
Kaneko	gender	53.15	53.33	52.75	49.65
	race	84.24	84.84	84.80	83.26
	religion	69.27†	69.80†	69.72†	67.66†
CDA	gender	64.00	65.38	64.70	61.19
	race	77.33	77.79	77.78	75.93
	religion	49.00	49.03	48.97	46.78
AttenD	gender	**65.91**	**67.10**	**66.69**	**63.59**
	race	92.26	92.77	92.74	91.79
	religion	68.51	69.08	68.95	66.97

Table 4: Inference-based bias measurements. Best scores are highlighted in **bold**, underlined, or marked with † for **gender**, race and religion† respectively

assumed biased. Suppose that the set contains M instances, and let the predictor's probabilities of the i^{th} instance for entail, contradict and neutral be e_i, c_i and n_i. Following Dev et al. (2020), we report three measures of inference-based bias: (1) Net Neutral (**NN**): $NN = \frac{1}{M}\sum_{i=1}^{M} n_i$; (2) Fraction Neutral (**FN**): $FN = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{n_i=max(e_i,c_i,n_i)}$; (3) Threshold τ (**T:**τ): $T : \tau = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}_{n_i>\tau}$.

In this experiment, we finetune text encoders on **MNLI** dataset for natural language inference (Wang et al., 2018). A bias-free model should score 1 (100%) in all three measures. We report our findings in Table 4. Our method outperforms the original model and the baselines. This result shows that AttenD succeeds in mitigating stereotypes in real world inference settings. We also debias hate speech detection models and report the results in Appendix A.3.

Models	Single sentence		Double sentence				
	sst2	cola	stsb	mrpc	mnli (m/mm)	rte	wnli
BERT	92.78	56.05	88.97	92.25	83.54 / 82.68	70.04	45.07
Sent-D	91.63	**59.08**	89.58	90.12	**84.97** / 83.51	68.95	28.17
Kaneko	91.97	56.50	88.44	90.69	84.48 / 83.66	59.93	52.11
CDA	92.32	55.98	88.93	90.60	84.31 / 82.26	64.26	25.35
AttenD$^-$	92.32	56.25	81.12	80.44	84.59 / 83.96	58.12	39.44
AttenD	**92.66**	55.22	**89.62**	**91.22**	84.63 / **84.19**	**70.40**	**53.52**

Table 5: Performance of different models on GLUE tasks. The table shows *accuracy* scores for **sst2**, **rte**, **wnli**, and **mnli** for both matched and mismatched instances; *f1* for **mrpc**; *spearman correlation* for **stsb**; and *matthews correlation* for **cola**

5.3 Evaluations of Semantic Preservation

We use GLUE benchmark (Wang et al., 2018) to verify whether the debiased text encoder still holds enough semantic information to be applicable in downstream NLP tasks. In essence, GLUE assesses the natural language understanding capabilities of NLP models. So, it constitutes a suitable stack to evaluate the semantic preservation of AttenD. In this experiment, we finetune our debiased models on seven different tasks from GLUE and show that per-task accuracy is preserved in Table 5. We also observe that not using negative examples (AttenD$^-$) severely hurts semantics.

6 Conclusion

We proposed in this paper to pay closer attention to the attention mechanism of text encoders. Specifically, we find that social bias also resides in attention weights, in addition to in representations and embeddings. We characterize attention bias in text encoders by looking at which demographics are most relevant given a stereotypical scenario under different attention maps in various layers. Our bias quantification method is a weighted average of Pearson correlations between attention allocations for demographic group words. We also propose a novel debiasing method by modifying

the self-attention weights so as to ensure equal attention activations across all group words for each token in each input sentence. At the same time, we use knowledge distillation from a teacher text encoder to preserve the useful semantics contained within. Finally, we utilize negative sampling with non-demographic word sets as the second sentence, where the teacher objective rather than attention equalization objective is applied, to prevent sentence-pair functionality in text encoders from being destroyed. We find that by mitigating biases from attention, the overall model bias is also reduced. We demonstrate this with various experiments that probe for bias internally, and when text encoders are used in downstream tasks, namely sentence inference and hate speech detection with limited costs to semantic usefulness.

7 Limitations

AttenD introduces many advantages. It is intuitive, simple in implementation, and inexpensive in terms of data resources. Also, the definitions of bias types and social groups in AttenD are extremely easy and flexible. However, we are aware of the following limitations: (1) There are more social divisions in the real world than the three dimensions we studied. Besides, bias types can be correlated in intricate ways, and it is not clear which or how many groups to include. For these reasons, we follow previous work and constrain our experiments to common use-cases. (2) We calibrate attention scores of every word in the input. However, some words are inherently charged with a strong inclination toward one group, e.g., *beard* to ***male*** or *pregnant* to ***female***. Such words need not be debiased, which requires compiling expensive lists of related words for every social group and protecting them from attention equalization. We rely on knowledge distillation to retain as much useful semantic information as possible in order to assuage this concern. (3) We use discrete words for debiasing, and it is not obvious how to extend this to treat implicit bias (e.g. bias existing between *doctor* and *engineer* because they are both stereotyped to be rather occupations for men) or bias toward finer-grained groups, e.g. female muslims, or buddhist Black Americans. (4) The template structure that we use for bias quantification and reduction does not necessarily capture all forms of social bias that are potentially concealed in the attention mechanism. Bias can also be internally encoded in attention weights in templates different from *"sentence from a corpus **[SEP]** demographic 1 demographic 2 demographic 3"*. For example, instead of declaring bias as a difference in attentions of the original sentence on social groups, we do the reverse and focus on the attention of groups on the sentence. Or we can analyze the attention distribution of demographics on attributes such as occupations or polarity adjectives, by using templates such as *"sentence describing a person in a demographic group **[SEP]** occupation 1 occupation 2 occupation 3"*. We believe that calibrating the attention of demographics on occupation terms helps in reducing implicit bias, at least the one related to occupations. Also, different kinds of template structures should be used in order to capture the most complete notion of attention bias possible. These points constitute sound and promising future directions for our research. We plan to address them all in upcoming work.

8 Ethical Considerations

We propose a debiasing method based on equalizing and calibrating the attention of text encoders on mentions of social groups. Although the approach is, in itself, independent from the choice of such groups, or the selection of identity terms and definition words that characterize these groups, we focus in our experiments on bias types and groups commonly used in the debiasing literature; namely binary gender, race and religion. We have shown that our method works for both binary and multiclass groups. That being said, we have not experimented yet with demographics divided into dozens of categories, e.g. nationality, or the full scope of gender. We also did not include analysis for groups who are victims of under-criticized microaggressions such as old people, fat people or people suffering from physical/mental disabilities. We justify our experimental decisions with the following: (1) Current work in the literature focuses primarily on the three major demographic dimensions. So to facilitate comparison, we used that too. (2) Existing benchmarks to quantify the amount of bias in text encoders are often limited to binary gender, race and religion. So even though our approach enables the reduction of bias for minority groups, we have no reliable data and benchmarks to assess whether debiasing is indeed effective for such groups. We encourage researchers and data collectors in the field to produce more inclusive benchmarks in the

future.

We would like to remind all users that our models are not perfect, even after going through debiasing. Although our experiments show that bias is indeed reduced, it is not completely mitigated. Also, there is the possibility that finetuning on the News-commentary-v15 corpus might introduce new biases encoded in the data. More generally, the bias detection experiments used in this paper and in all related work have positive predictive ability, which means that they can only detect the presence of bias, not the absence of it. So it is possible that bias is still hiding under different forms that current experimental lenses fail to detect. We believe that the community needs to include some aspect of human evaluation to faithfully assess the stereotypical propensities of text encoders. We project to do that in future work.

References

Vamsi Aribandi, Yi Tay, and Donald Metzler. 2021. How reliable are model diagnostics? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1778–1785.

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*, pages 1105–1119.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155.

Rishabh Bhardwaj, Navonil Majumder, and Soujanya Poria. 2021. Investigating gender bias in bert. *Cognitive Computation*, pages 1–11.

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

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping norwegian salmon: an inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29:4349–4357.

Shikha Bordia and Samuel Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on twitter across multiple protected characteristics. *EPJ Data science*, 5:1–15.

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

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2020. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286.

Robert Dale. 2019. Law and word order: Nlp in legal tech. *Natural Language Engineering*, 25(1):211–217.

Jeffrey Dastin. 2018. Amazon scraps secret ai recruiting tool that showed bias against women. In *Ethics of Data and Analytics*, pages 296–299. Auerbach Publications.

Sunipa Dev, Tao Li, Jeff M Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7659–7666.

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

Mark Díaz, 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*, pages 1–14.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

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

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022a. Iterative adversarial removal of gender bias in pretrained word embeddings. In *Proceedings of the 37th ACM/SIGAPP Symposium On Applied Computing*, pages 829–836.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022b. Masked language models as stereotype detectors? In *EDBT 2022*.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sanchez, Mugdha Pandya, and Adam Lopez. 2020. Intrinsic bias metrics do not correlate with application bias. *arXiv preprint arXiv:2012.15859*.

Hila Gonen 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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

C Hansen, M Tosik, G Goossen, C Li, L Bayeva, F Berbain, and M Rotaru. 2015. How to get the best word vectors for resume parsing. In *SNN Adaptive Intelligence/Symposium: Machine Learning*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Benjamin Hoover, Hendrik Strobelt, and Sebastian Gehrmann. 2020. exbert: A visual analysis tool to explore learned representations in transformer models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 187–196.

Forrest Iandola, Albert Shaw, Ravi Krishna, and Kurt Keutzer. 2020. Squeezebert: What can computer vision teach nlp about efficient neural networks? In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 124–135.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. 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 3543–3556.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Vaibhav Kumar, Tenzin Singhay Bhotia, and Tanmoy Chakraborty. 2020. Nurse is closer to woman than surgeon? mitigating gender-biased proximities in word embeddings. *Transactions of the Association for Computational Linguistics*, 8:486–503.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Quantifying social biases in contextual word representations. In *1st ACL Workshop on Gender Bias for Natural Language Processing*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Anne Lauscher, Tobias Lüken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. *arXiv preprint arXiv:2109.03646*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020a. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515.

Sheng Liang, Philipp Dufter, and Hinrich Schütze. 2020b. Monolingual and multilingual reduction of gender bias in contextualized representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5082–5093.

Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57.

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

Thomas Manzini, Yao Chong Lim, Yulia Tsvetkov, and Alan W Black. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047*.

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

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *NAACL-HLT (1)*.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2021. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527*.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Eric Nalisnick, Bhaskar Mitra, Nick Craswell, and Rich Caruana. 2016. Improving document ranking with dual word embeddings. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 83–84.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804.

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

Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2020. Learning to deceive with attention-based explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4782–4793.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228.

Shauli Ravfogel, 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*.

Manoel Horta Ribeiro, Pedro H Calais, Yuri A Santos, Virgílio AF Almeida, and Wagner Meira Jr. 2018. Characterizing and detecting hateful users on twitter. In *Twelfth international AAAI conference on web and social media*.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.

Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF, and Hugging Face. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Association for Computational Linguistics*.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Niral Shah. 2019. "asians are good at math" is not a compliment: Stem success as a threat to personhood. *Harvard Educational Review*, 89(4):661–686.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3239–3254.

Ian Tenney, James Wexler, Jasmijn Bastings, Tolga Bolukbasi, Andy Coenen, Sebastian Gehrmann, Ellen Jiang, Mahima Pushkarna, Carey Radebaugh, Emily Reif, et al. 2020. The language interpretability tool: Extensible, interactive visualizations and analysis for nlp models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 107–118.

Deborah A Trytten, Anna Wong Lowe, and Susan E Walden. 2012. "asians are good at math. what an awful stereotype" the model minority stereotype's impact on asian american engineering students. *Journal of Engineering Education*, 101(3):439–468.

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.

Sumithra Velupillai, Hanna Suominen, Maria Liakata, Angus Roberts, Anoop D Shah, Katherine Morley, David Osborn, Joseph Hayes, Robert Stewart, Johnny Downs, et al. 2018. Using clinical natural language processing for health outcomes research: overview and actionable suggestions for future advances. *Journal of biomedical informatics*, 88:11–19.

Jesse Vig. 2019a. Bertviz: A tool for visualizing multihead self-attention in the bert model. In *Proceedings of the 2019 ICLR Workshop on Debugging Machine Learning Models*.

Jesse Vig. 2019b. Visualizing attention in transformer-based language representation models. *arXiv preprint arXiv:1904.02679*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.

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

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. *Advances in Neural Information Processing Systems*, 30:585–596.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853.

A Appendix

A.1 Training Hyperparameters

We used Adam optimizer (Kingma and Ba, 2014) with a learning rate of $5e^{-6}$ for 3 epochs. We keep the betas to their default values (0.9, 0.999) as in PyTorch implementation (Paszke et al., 2017). We set the loss coefficient λ to 2.0 and the negative ratio η to 0.8 meaning that in 80% of the iterations, we use negative examples whose number we set to 5 in each negative iteration. For the number of heads k having the top scores of attention bias, we experimented with different values for k: $\{10, 20, 30, 50, 70, 100, all\}$. We found that debiasing all heads works best for all the text encoders that we experimented with. We only tuned the values of λ, η, k, the learning rate, and the number of epochs. We conducted the hyperparameter search manually on the development set of the News-commentary-v15 corpus, and selected the hyperparameter configuration that maximized the attention-based bias metric. As for GLUE experiments, we follow the experimental setup of Devlin et al. (2018) and train each task for 3 epochs with a learning rate of $2e^{-5}$ on their respective training data. We ran all of our training and experiments on a NVIDIA Tesla V100 GPU.

A.2 Definition of bias types and social groups used in this paper

While the approach is independent of the definition of social groups and categories (it could work for any kind of grouping, e.g., cuisine styles or sports), in the experiment we focus on groups commonly used in the debiasing literature: binary gender, religion and race. This is to facilitate comparison, but nothing in the approach prevent it from being used with broader and more inclusive groups. This being said, we have not experimented yet with debiasing where a dimension is divided in dozens of categories.

We list the definition tuples that we used in Table 6. We show that AttenD does not incur strict rules for defining social groups, unlike previous

work (Bolukbasi et al., 2016; Kaneko and Bollegala, 2019, 2021) that require the definition words to be organized in a predefined format (pairs of words or bag of words for every group), and provided in relatively large quantities. We can see from Table 6 that it is sufficient to define one tuple per bias type (e.g., race) if the tuples are hard to come by. Also, the tuples need not be of the same size (e.g., in religion there is a missing word for *buddhist* group since it is not clear which word to use in that tuple). This desired property owes to the fact that AttenD does not learn subspaces or directions for every bias type as previous works do (Bolukbasi et al., 2016; Kaneko and Bollegala, 2019; Kumar et al., 2020; Kaneko and Bollegala, 2021). In contrast, AttenD uses the tuples in order to equalize the attentions of the input sentence, and make the words therein attend to the groups with the same intensity. These example categories used in experiments are neither complete nor exhaustive, and in some experiments also include terms possibly considered inappropriate but that appear in the corpus and we may still want to debias from.

A.3 Extrinsic bias evaluation on the task of hate-speech detection

Recent studies show that intrinsic metrics of bias do not necessarily correlate with bias measures on concrete real-world applications (Goldfarb-Tarrant et al., 2020). In the body of this paper, we already conducted intrinsic and extrinsic bias evaluations. In this experiment, we validate the efficacy of our debiasing method on a concrete real-world hate speech detection application where an input snippet of text is classified as either offensive (*toxic*, *harmful*, *disrespectful*, etc.) or not. We use hate speech detection because it is well studied in the literature (Burnap and Williams, 2016; Ribeiro et al., 2018; Zhang et al., 2018), and high-quality datasets which are tagged with social groups already exist (Borkan et al., 2019; Mathew et al., 2021).

Admittedly, common social biases have also been shown to exist in hate speech detection models, for example in associating toxicity to frequently attacked groups (such as "muslim" or "gay") even if the text itself is not toxic (Dixon et al., 2018; Park et al., 2018). In this experiment, we adopt the bias definition of Borkan et al. (2019) which casts bias as a skewing in the hate speech detector scores based solely on the social groups mentioned in the text. In other words, we consider a model to exhibit unintended social stereotypes if the model's performance varies across groups. We use the bias measures proposed by Borkan et al. (2019) which are based on the Area Under the Receiver Operating Characteristic Curve (ROC-AUC, or AUC) metric. AUC measures the probability that a randomly chosen negative example (not offensive) receives a lower toxicity score than a randomly chosen positive example (offensive), meaning that a perfect model should always have an AUC score of 1.0. Stated differently, all negative examples have lower toxicity scores than positive examples. While AUC is used to measured the general performance of classifiers, Borkan et al. (2019) propose three extensions of AUC to measure bias. We summarize them in the following:

Subgroup (Sub) AUC: where AUC is computed only on the group under consideration and not on all the examples of the test benchmark, i.e. only positive and negative examples of the target group are considered. This metric represents the model's performance on a given group. A higher value means that the model is good at distinguishing between toxic and non-toxic texts specific to the group.

Background Positive Subgroup Negative (BPSN) AUC: where AUC is calculated on the negative examples of the target group, and the positive examples of the background (all other groups except the group under consideration). This metric computes whether the model *discriminates* against the target group with respect to the others. This value is reduced when non-toxic examples of the group have *higher* toxicity scores than actually toxic examples of the background.

Background Negative Subgroup Positive (BNSP) AUC: where AUC is calculated on the positive examples of the target group, and the negative examples of the background. This metric computes whether the model *favors* the target group with respect to the others. This value is reduced when toxic examples of the group have *lower* toxicity scores than non-toxic examples of the background.

In this experiment, we finetune the text encoder under study on hate speech detection task using the training set of HateXplain dataset (Mathew et al., 2021). We also use the test portion of HateXplain for the evaluation, which contains posts from Twitter[11] and Gab[12] annotated with their ground-truth

[11] https://twitter.com
[12] https://gab.com

gender		religion			
male	*female*	*muslim*	*christian*	*jewish*	*buddhist*
man	woman	muslim	christian	jewish	buddhist
boy	girl	muslims	christians	jews	buddhists
father	mother	islam	christianity	judaism	buddhism
brother	sister	mosque	church	synagogue	temple
grandfather	grandmother	quran	bible	torah	
son	daughter	imam	priest	rabbi	monk
gentleman	lady	mohammad	jesus	moses	buddha
he	she				
his	her				
himself	herself				

race			
white	*black*	*asian*	*hispanic*
white	black	asian	hispanic

Table 6: Full list of definition tuples for bias types and social groups used in this work

toxicity scores and the social groups and communities they target. Fundamentally, the three metrics described above give bias scores per group. In order to combine the per group scores in one overall measure, we apply the Generalized Mean of Bias (GMB) introduced by the Google Conversation AI Team as part of their Kaggle competition[13], and later used by Mathew et al. (2021) in their own evaluations. The formula of GMB is as the following:

$$GMB(b) = (\frac{1}{|b|}\sum_{g=1}^{|b|} b_g^p)^{1/p} \quad (5)$$

where b is an array of AUC scores per group, and b_g is the AUC score of group g. We follow Mathew et al. (2021) and set p to *-5*. We compute the GMB of all three metrics: Subgroup, BPSN and BNSP. As for Subgroup, we also add the standard deviation as it gives valuable information about how much the performance of the hate speech detection model varies across groups. We report our results in Table 7, in addition to classic performance measures.

We observe that AttenD provides competitive results across the four bias metrics, and largely outperforms the baselines. Especially with *GMB-BNSP*, where bias scores of the original model are very low (i.e. it is throttled by social biases), we observe the best improvements overall, and by a large margin compared to existing debiasing methods. Also, the variance in model performance is lowest with AttenD, which confirms that the corresponding hate speech detection model has less stereotypes about different social groups. Finally, the general performance (Accuracy, F1 score and AUC) of the hate speech detection model after debiasing is not hurt.

A.4 Visualizing debiasing results

In this experiment, we aim to visualize the effects of debiasing on attention weights. We only focus on binary gender bias for two reasons: First, it is easier to visualize binary variables on a 2D plane than multiclass variables (such as race, religion...). Second, gender is the most well studied bias type (Bolukbasi et al., 2016; Caliskan et al., 2017; May et al., 2019), so linguistic resources and vocabularies for gender exist and are well documented. We use the vocabulary words compiled by (Kaneko and Bollegala, 2019) and categorized into three non-overlapping subsets: (1) **Male-definition** Ω^M whose corresponding words are exclusively male-gendered such as *father*, *king* or *uncle*. (2) **Female-definition** Ω^F which is a set of inherently female words (*mother*, *queen*, *aunt*...). (3) **Gender-stereotype** Ω^S which is constituted of words that are not gendered by definition, but that carry a strong gender stereotype such as *doctor* being attributed to ***male*** or *nurse* to ***female***.

For every word $w \in \Omega^M \cup \Omega^F \cup \Omega^S$, we extract sentences from the News-commentary-v15 corpus where w is mentioned. We denote this set as S^w. Then, for every sentence $s \in S^w$, we append the dummy input *"man, woman"* as explained in Section 3 and the example of Figure 1. The augmented input s' is then fed to the text encoder of

[13] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview/evaluation

Models	Performance			Bias			
	Acc↑	F1↑	AUC↑	STD-Sub↓	GMB-Sub↑	GMB-BPSN↑	GMB-BNSP↑
BERT	0.783	0.823	0.870	0.119	0.698	**0.800**	0.379
Sent-D	0.791	0.825	0.870	0.121	0.689	0.725	0.583
Kaneko	0.797	0.833	0.872	0.112	0.705	0.789	0.512
AttenD	0.789	0.829	0.866	**0.085**	**0.808**	0.793	**0.726**

Table 7: AUC-based bias measures on hate speech detection task

interest (BERT base in this experiment), and we collect the attention scores of w on the second-half tokens ***man*** and ***woman***. Finally, for every word $w \in \Omega^M \cup \Omega^F \cup \Omega^S$, we take the mean of its attention scores in S^w. By the end of this procedure, we have for every word w its attention score on the words ***man*** (a_m^w) and ***woman*** (a_f^w) as computed on the News-commentary-v15 corpus which includes overall 223,153 sentences. We take the difference $a_m^w - a_f^w$ which indicates the preference of the text encoder to consider w as male (positive difference) or female (negative difference). The absence of gender bias is reflected in difference scores near zero.

We plot the results in Figure 5 where the x-axis represents the differences $a_m^w - a_f^w$, and the y-axis random values to separate the words vertically. Stereotype words (green dots) should have values near 0, which is not the case in Figure 5(b). This means that BERT has a strong preference for one of the genders, and is thus heavily biased. In contrast, our method brings the attention of stereotype words near 0, meaning that they prefer neither male nor female connotations. Moreover, the spread of stereotype words in Figure 5(d) is narrower than male- or female-oriented words, which is desired since these are inherently gendered and must pick a side. This result strengthens the claim that AttenD preserves semantic information, and is less severe in reducing bias from gendered words as it is on gender-neutral words. The difference in spread is less apparent in the original BERT model. We also note that debiasing the embeddings of BERT rather than the attention mechanism as in (Kaneko and Bollegala, 2021) (Figure 5(c)) is not enough since bias information is still lurking (and perhaps made worse for some words) in the attention component. Thus, we conclude that working on attention directly constitutes our best option for debiasing to date.

Models	Single sentence tasks		Double sentence tasks				
	sst2	cola	stsb	mrpc	mnli (m/mm)	rte	wnli
BERT	**92.78**	56.05	88.97	**92.25**	83.54 / 82.68	70.04	45.07
AttenD	92.66	55.22	**89.62**	91.22	**84.63 / 84.19**	**70.40**	**53.52**
AttenD$^-$	92.32	**56.25**	89.12	80.44	84.59 / 83.96	58.12	39.44

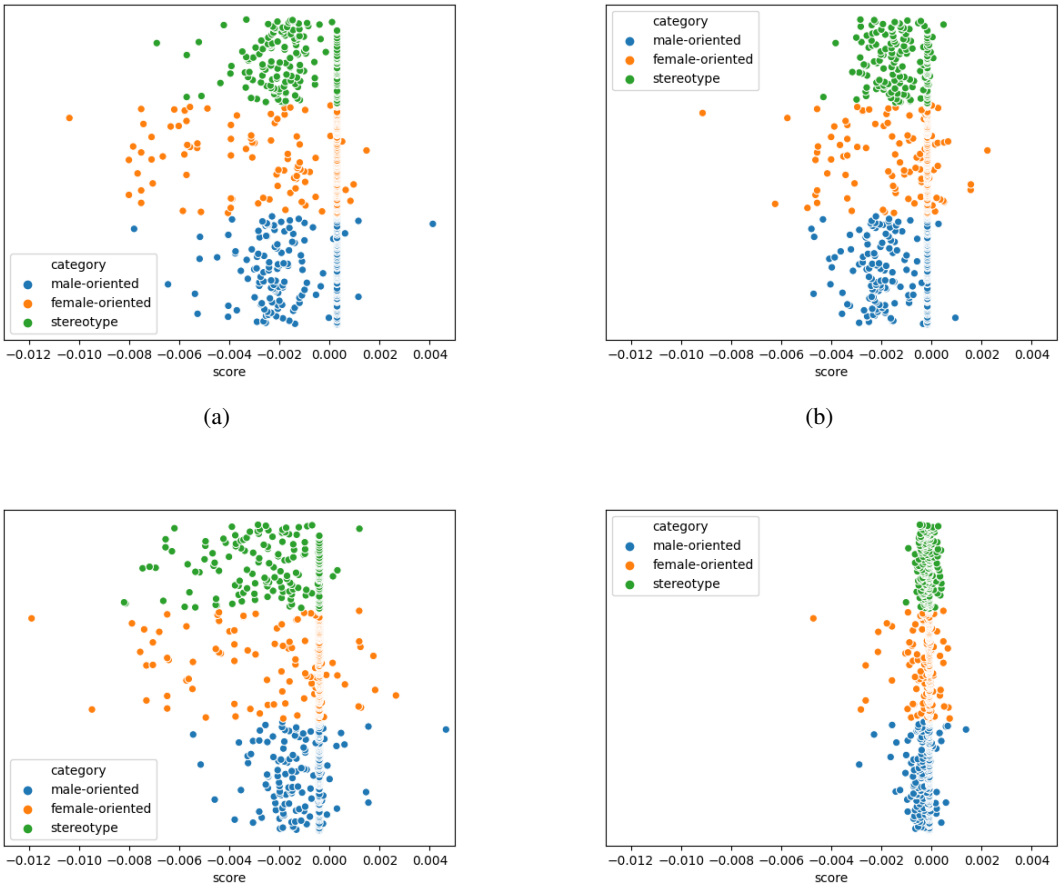
Table 8: Effect of negative examples on GLUE tasks. The table shows *accuracy* scores for **sst2**, **rte**, **wnli**, and **mnli** for both matched and mismatched instances; *f1* for **mrpc**; *spearman correlation* for **stsb**; and *matthews correlation* for **cola**

A.5 Effect of negative examples on representativeness

We remind that the introduction of negative examples to training serves in forcing the text encoder not to rely on a dangerous shortcut which is distributing its attention uniformly on all the tokens constituting the second half of the input, no matter what the input is. This is particularly important in double-sentence tasks where the text encoder is given two input sentences. In addition to Tables 3 and 5 which highlighted the effect of negative sampling on the final stereotype scores, the primary goal of using negative examples remains the preservation of the text encoder's representativeness. In Table 8, we report the performance of *AttenD* and *AttenD*$^-$ with and without negative examples respectively on GLUE tasks. Unsurprisingly, the lack of negative examples does not damage the performance of single-sentence tasks since these ignore the second half of the input altogether. However, in double-sentence tasks where both halves are used for prediction, Table 8 shows that negative sampling plays a pivotal role in preserving the semantics of text encoders, and bypassing the side effects inflicted by attention equalization.

A.6 Word-Level vs Sentence-Level Debiasing

As previously explained in the paper, AttenD calibrates the attention weighs of all tokens of the input sentence on group-related words. Since we used BERT-based models in our experiments, the



(a) (b)

(c) (d)

Figure 5: Scatter plots of attention scores on male - female direction. (a) Original BERT, (b) BERT debiased by Sent-D (c) BERT debiased by (Kaneko and Bollegala, 2021), (d) BERT debiased by AttenD

first token in the input is the special [CLS] token, which is considered by the NLP community as a vector representation for the entire input sentence. In the current version of AttenD, we also calibrate the attention weighs of the special [CLS] token on groups, in addition to calibrating the other tokens of the sentence. One can see this notion as a combined word-level and sentence-level debiasing. In this experiment, we motivate this design choice by comparing it to word-level and sentence-level debiasing separately. For word-level, we exclude the [CLS] token from the attention equalization process, whereas in sentence-level we only calibrate the attention of [CLS]. We use all the bias evaluations run so far to understand the difference in performance. Tables 9, 10, 11, 12 and 13 report the results of StereoSet, Crows-Pairs, inference, hate speech and GLUE experiments respectively. We denote word-level debiasing by **No [CLS]**, and sentence-level debiasing by **Only [CLS]** in the tables. The combination of both is referred to as **AttenD**, and is the variant that we promote in this paper. We observe that while the three settings are good at reducing bias from text encoders, AttenD is superior than word-level and sentence-level debiasing since it capitalizes on the benefits of both. It enjoys the fine granularity of reducing bias from every word, while it also mitigates biases that manifest at sentence-level.

A.7 Static vs Random ordering of group-related words

In the preprocessing step of our method (as explained in Section 3), we use a preset ordering of group-related words of a given bias type to form the second input. For example, if we have the groups *Muslim*, *Christian*, *Jew* and *Buddhist* defining the religion bias type, AttenD constructs the second input using the same preset ordering of groups across all samples of the training data. Continuing the example above, AttenD appends the following artificial sentence "muslim, christian, jew, buddhist". In this experiment, we change the ordering of groups in a random way. Tables 9, 10, 11, 12 and 13 also report the bias scores of AttenD (static ordering) and AttenD with random ordering.

Although the semantic performance of AttenD with random ordering is better, we notice that it suffers from a stronger presence of bias than in its static counterpart. In Table 12, AttenD with random ordering has an AUC score of 0 in one of the groups, which made the GMB extremely small. We suspect that the relatively poor fairness of random ordering owes to the fact that the model might be confused by different orderings throughout the iterations. A more serious analysis of the impact of group order on the overall performance (fairness and semantics) of AttenD motivates the direction of future work.

Models	AttenD		No [CLS]		Only [CLS]		Random Order	
Overall (lm/ss)	83.34	**53.04**	80.37	53.71	81.70	55.51	82.91	54.75
gender (lm/ss)	78.24	53.73	76.86	**52.94**	75.88	54.51	79.02	55.69
race (lm/ss)	86.28	**51.87**	84.10	53.01	85.24	55.09	86.75	54.57
religion (lm/ss)	88.46	**53.85**	84.62	60.26	85.26	56.41	87.18	56.41
profession (lm/ss)	80.96	**54.14**	76.63	**54.14**	78.99	56.24	79.17	54.51

Table 9: Language modeling (lm) and Stereotype scores (ss) on StereoSet of different variants of AttenD

Models	AttenD	No [CLS]	Only [CLS]	Random Order
Overall	55.7	56.1	**55.5**	58.36
gender	57.36	50.76	**50.0**	53.82
race	**51.15**	54.84	53.1	57.75
religion	**64.76**	69.52	65.71	67.62
age	43.68	56.32	44.83	**54.02**
sexual orientation	**58.33**	71.43	63.1	64.29
nationality	57.86	**53.46**	65.41	62.28
disability	60.0	61.67	**58.33**	65.0

Table 10: Bias measurements of different variants of AttenD on Crows-Pairs

A.8 Effect of AttenD on other transformer-based text encoders

We evaluate five widely used sentence-level text encoders: BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019), RoBERTa (Liu et al., 2019), DistilBERT (SANH et al.) and SqueezeBERT (Iandola et al., 2020). For each model, we evaluate both its base and large variants (except for DistlBERT and SqueezeBERT since these are not available in HuggingFace's transformers library[14]), original and debiased; which gives a total of sixteen evaluated models. We use Crows-Pairs dataset (Nangia et al., 2020) to quantify the intensity of undesired stereotypes encoded therein. As a reminder, ideal stereotype scores according to Crows-Pairs benchmark should be close to 50, i.e. models preferring neither stereotypes nor anti-stereotypes. Tables 14, 15, 16 and 17 show the bias results for BERT, ALBERT, RoBERTa and DistillBERT/SqueezeBERT respectively.

All five models exhibit substantial levels of bias, and in each of the bias types with differing intensities (religion, sexual orientation and disability be-

[14] https://huggingface.co/transformers/index.html

Model	Bias type	NN	FN	τ:0.5	τ:0.7
AttenD	gender	01.31	00.43	00.35	00.21
	race	93.31	93.94	93.90	93.04
	religion	68.51†	69.08†	68.95†	66.97†
No [CLS]	gender	00.85	00.36	00.30	00.20
	race	76.14	76.24	76.19	74.26
	religion	40.80	40.04	39.98	37.78
Only [CLS]	gender	**02.35**	**01.60**	**01.38**	**00.90**
	race	81.63	81.52	81.50	80.37
	religion	44.40	44.01	43.95	42.76
Random Order	gender	01.54	00.51	00.39	00.23
	race	54.71	54.92	54.89	52.49
	religion	26.94	26.67	26.59	24.58

Table 11: Inference-based bias measurements on different variants of AttenD. Best scores are highlighted with **bold character**, underlined, or marked with † for **gender**, race and religion† respectively

ing the bias categories with the most severe stereotyping). Also, we find that the large variants are more biased than their base counterparts mainly because large models, with their larger capacity and greater number of parameters, can capture more intricate and more sophisticated aspects of training data, exposing them to learn more bias. This finding corresponds well to results of previous work (Nangia et al., 2020; Nadeem et al., 2020). The tables also show that AttenD is effective in mitigating bias from BERT, ALBERT, RoBERTa, DistilBERT and SqueezeBERT, and produces a reduction of up to 25%. We note that AttenD succeeds in debiasing all models, with varying effectiveness across bias types. We also note that AttenD meets the best success with ALBERT as reductions are greater on this particular text encoder. We believe this is because ALBERT is composed of a single transformer layer (Lan et al., 2019) with substantially less parameters than BERT or RoBERTa; which makes debiasing easier since there is no interference between different attention layers. Finally, we see from the tables that AttenD sometimes contributes to adding a bit of bias. We observe that this phenomenon is rare, and happens especially with bias types we did not include in our design[15]. We assume that not explicitly compelling the text encoder to equalize attention heads corresponding to these overlooked bias types gave it green light to adjust these attentions in a way to facilitate solving the optimization problem; even if it entails adding bias. We plan to include all bias types present in Crows-Pairs dataset to our debiasing design as a future work.

[15]In the current version of this work, we remind that we only consider three bias types: gender, race and religion

Models	Performance			Bias			
	Acc↑	F1↑	AUC↑	STD-Sub↓	GMB-Sub↑	GMB-BPSN↑	GMB-BNSP↑
AttenD	0.789	0.829	0.866	**0.085**	**0.808**	0.793	**0.726**
No [CLS]	0.791	0.830	0.871	0.114	0.710	**0.797**	0.530
Only [CLS]	0.765	0.805	0.838	0.142	0.660	0.766	0.636
Random Order	0.784	0.822	0.861	/	/	0.764	/

Table 12: AUC-based bias measures on hate speech detection task on different variants of AttenD

Models	Single sentence tasks		Double sentence tasks				
	sst2	cola	stsb	mrpc	mnli (m/mm)	rte	wnli
AttenD	92.66	55.22	**89.62**	91.22	**84.63** / 84.19	70.40	**53.52**
No [CLS]	91.51	40.85	88.94	91.62	84.49 / 84.02	68.95	40.85
Only [CLS]	92.43	55.23	89.43	90.04	84.42 / 84.67	**71.84**	23.94
Random Order	**93.23**	**59.07**	88.85	**91.94**	83.75 / **84.86**	**71.84**	30.99

Table 13: GLUE performance of different variants of AttenD. The table shows *accuracy* scores for **sst2**, **rte**, **wnli**, and **mnli** for both matched and mismatched instances; *f1* for **mrpc**; *spearman correlation* for **stsb**; and *matthews correlation* for **cola**

Models	BERT base		BERT large	
Overall	60.48 → 55.70	-04.78	59.68 → 56.96	-02.72
race	58.14 → 51.15	-06.99	60.08 → 53.49	-06.59
gender	58.02 → 57.36	-00.66	55.34 → 53.05	-02.29
socioeconomic	59.88 → 51.16	-08.72	56.40 → 57.56	+01.16
nationality	62.89 → 57.86	-05.03	52.20 → 57.23	+05.03
religion	71.43 → 64.76	-06.67	68.57 → 66.67	-01.90
age	55.17 → 43.68	+01.15	55.17 → 54.02	-01.15
sexual orientation	67.86 → 58.33	-09.53	65.48 → 67.86	+02.41
physical appearance	63.49 → 61.90	-01.89	69.84 → 65.08	-04.76
disability	61.67 → 60.00	-01.67	76.67 → 65.00	-11.67

Table 14: Bias reduction in BERT base and large measured on Crows-Pairs dataset. Each cell is organized as follows: o → d +/-diff where o is the stereotype score of the original model, d is that of the debiased model using attention-based debiasing, and $diff$ is the difference in stereotype score. Negative values correspond to reduction in bias (desired) where positive values mean addition of bias (undesired).

Models	ALBERT base		ALBERT large	
Overall	56.76 → 51.99	-04.77	60.48 → 53.58	-06.90
race	51.36 → 48.84	-00.20	59.11 → 50.97	-08.14
gender	54.20 → 53.44	-00.76	56.11 → 48.47	-04.58
socioeconomic	60.47 → 61.05	+00.58	54.07 → 50.00	-01.16
nationality	51.57 → 57.86	+06.29	62.26 → 60.38	-04.07
religion	59.05 → 60.00	+00.95	76.19 → 61.90	-14.29
age	65.52 → 42.53	-08.05	54.02 → 54.02	-00.00
sexual orientation	75.00 → 38.10	-13.10	71.43 → 63.10	-08.33
physical appearance	46.03 → 41.27	+04.76	58.73 → 57.14	-01.59
disability	86.67 → 61.67	-25.00	73.33 → 58.33	-15.00

Table 15: Bias reduction in ALBERT base and large measured on Crows-Pairs dataset. Each cell is organized as follows: o → d +/-diff where o is the stereotype score of the original model, d is that of the debiased model using attention-based debiasing, and $diff$ is the difference in stereotype score. Negative values correspond to reduction in bias (desired) where positive values mean addition of bias (undesired).

Models	RoBERTa base		RoBERTa large	
Overall	53.98 → 51.39	-02.59	61.27 → 56.83	-04.44
race	47.09 → 50.39	-02.52	61.43 → 53.49	-07.94
gender	54.96 → 45.80	-00.76	51.91 → 51.91	-00.00
socioeconomic	56.40 → 55.81	-00.59	66.28 → 59.88	-06.40
nationality	45.28 → 43.40	+01.88	56.60 → 55.35	-01.25
religion	56.19 → 60.00	+03.81	59.05 → 62.86	+03.81
age	64.37 → 56.32	-08.05	71.26 → 62.07	-09.19
sexual orientation	69.05 → 48.81	-17.86	71.43 → 59.52	-11.91
physical appearance	66.67 → 60.32	-06.35	68.25 → 66.67	-01.58
disability	71.67 → 65.00	-06.67	66.67 → 70.00	+03.33

Table 16: Bias reduction in RoBERTa base and large measured on Crows-Pairs dataset. Each cell is organized as follows: o → d +/-diff where o is the stereotype score of the original model, d is that of the debiased model using attention-based debiasing, and $diff$ is the difference in stereotype score. Negative values correspond to reduction in bias (desired) where positive values mean addition of bias (undesired).

Models	DistilBERT		SqueezeBERT	
Overall	56.83 → 51.26	-05.57	57.43 → 54.71	-02.72
race	53.29 → 47.87	-01.16	55.04 → 56.01	+00.97
gender	54.58 → 46.56	-01.14	52.67 → 48.47	-01.14
socioeconomic	55.81 → 58.14	+02.33	57.56 → 51.16	-06.40
nationality	54.09 → 50.94	-03.15	53.46 → 61.01	+07.55
religion	70.48 → 57.14	-13.34	74.29 → 60.95	-13.34
age	59.77 → 48.28	-08.05	55.17 → 48.28	-03.45
sexual orientation	70.24 → 55.95	-14.29	70.24 → 57.14	-13.10
physical appearance	55.56 → 63.49	+07.93	52.38 → 52.38	-00.00
disability	61.67 → 56.67	-05.00	70.00 → 61.67	-08.33

Table 17: Bias reduction in DistilBERT and SqueezeBERT measured on Crows-Pairs dataset. Each cell is organized as follows: o → d +/-diff where o is the stereotype score of the original model, d is that of the debiased model using attention-based debiasing, and $diff$ is the difference in stereotype score. Negative values correspond to reduction in bias (desired) where positive values mean addition of bias (undesired).