ABLE: Agency-BeLiefs Embedding to Address Stereotypical Bias Through Awareness Instead of Obliviousness

Michelle YoungJin Kim*, **Junghwan Kim***, **Kristen Marie Johnson**
Michigan State University, University of Michigan, Michigan State University
East Lansing, MI, USA; Ann Arbor, MI, USA; East Lansing, MI, USA
kimmic16@msu.edu, kimjhj@umich.edu, kristenj@msu.edu

Abstract

Natural Language Processing (NLP) models tend to inherit and amplify stereotypical biases present in their training data, leading to harmful societal consequences. Current efforts to rectify these biases typically revolve around making models oblivious to bias, which is at odds with the idea that humans require increased awareness to tackle these biases better. This prompts a fundamental research question: are bias-oblivious models the only viable solution to combat stereotypical biases? This paper answers this question by proposing the Agency-BeLiefs Embedding (ABLE) model, a novel approach that actively encodes stereotypical biases into the embedding space. ABLE draws upon social psychological theory to acquire and represent stereotypical biases in the form of agency and belief scores rather than directly representing stereotyped groups. Our experimental results showcase ABLE's effectiveness in learning agency and belief stereotypes while preserving the language model's proficiency. Furthermore, we underscore the practical significance of incorporating stereotypes within the ABLE model by demonstrating its utility in various downstream tasks. Our approach exemplifies the potential benefits of addressing bias through awareness, as opposed to the prevailing approach of mitigating bias through obliviousness.

Keywords: Bias Detection, AI Fairness, Stereotyping

1. Introduction

Recent studies in Natural Language Processing (NLP) have unveiled a concerning issue: NLP models frequently exhibit stereotypical biases associated with demographic groups (Bolukbasi et al., 2016; Caliskan et al., 2017; May et al., 2019). Given the widespread deployment of these models across diverse domains, the escalating potential for risks and harms stemming from these biases demands our immediate attention.

In response to this formidable challenge, a range of strategies has surfaced to address and alleviate these biases within NLP models. A prevailing objective of these strategies is to promote a state of bias obliviousness within the models. For instance, counterfactual data augmentation methods (Zhao et al., 2018; Zmigrod et al., 2019; Webster et al., 2020; Lauscher et al., 2021; Qian et al., 2022; Xie and Lukasiewicz, 2023; Fatemi et al., 2023) strive to balance training data by replicating each training sentence for every demographic group, ensuring that every group appears in identical contextual settings. As a result, under counterfactual data augmentation, models cannot form any biased associations. Alternatively, equalizing loss techniques are explicitly designed to minimize disparities in embeddings (Cheng et al., 2021; He et al., 2022; Li et al., 2023), prediction probabilities (Guo et al., 2022; Zhou et al., 2023) or attention weights (Gaci et al., 2022) across different demographic groups. Subspace removal methods (Bolukbasi et al., 2016; Dev et al., 2020; Liang et al., 2020; Ravfogel et al., 2020; Kaneko and Bollegala, 2021; Ravfogel et al., 2022; Kumar et al., 2023) identify bias directions within the embedding space and subsequently eliminate or penalize embedding components along these bias directions. Through the targeted removal of components exhibiting correlations with bias, these equalizing loss and subspace removal methods prevent models from encoding bias.

While these bias-oblivious approaches have successfully reduced stereotypical biases in certain scenarios, they still remain vulnerable to bias when subjected to further fine-tuning. The conventional practice of fine-tuning a single, large language model across a spectrum of tasks necessitates separate bias mitigation in each fine-tuning instance. This not only poses a logistical challenge but also imposes a substantial financial burden. Recent empirical findings have revealed that bias mitigation measures do not consistently carry over to downstream tasks (Jin et al., 2021; Cao et al., 2022a; Kaneko et al., 2022; Shen et al., 2022; Cabello et al., 2023) [1], adding complexity to the situation. Moreover, bias-oblivious models lose access to bias information, rendering them incapable of analyzing stereotypes in text data, diagnosing biased model outputs, or identifying distribution shifts from

[*] These authors contributed equally to this work.

[1] It is worth mentioning that there is also conflicting evidence supporting the transferability of bias measures across different tasks (Orgad et al., 2022; Ladhak et al., 2023).

evolving stereotypes.

Bias-oblivious approaches stand in stark contrast to the way humans address stereotypical biases. Extensive research in social science confirms that awareness of one's own biases plays a pivotal role in reducing bias, as opposed to embracing obliviousness (Lee, 2017; Pope et al., 2018; Boring and Philippe, 2021). Stereotyping is a natural human tendency to simplify our understanding of society within the limits of our cognitive resources. To counter stereotypical biases effectively, it requires active awareness and proactive interventions, rather than adopting an oblivious stance. The question then arises: can NLP models also leverage bias awareness to effectively address bias?

In response to this fundamental question, our paper introduces the Agency-BeLiefs Embedding (ABLE) model, a novel approach designed to proactively incorporate stereotypical biases into the embedding space. By leveraging insights from social psychological theories (Koch et al., 2016), our model learns to predict these biases in the form of *agency* and *belief* scores, endowing the model with a profound awareness of bias (§3.1). Additionally, we employ contrastive learning loss to ensure the consistent representation of each stereotyped group by clustering texts containing the same group (§3.2).

Our experimental results illustrate the remarkable effectiveness of ABLE in learning agency and belief stereotypes while preserving the language model's proficiency (§4). Our model formulates stereotypical biases as agency and belief scores rather than focusing on specific demographic groups, making it generalizable to unseen demographic groups. We emphasize the practical importance of embedding this bias awareness within the ABLE model and demonstrate its utility in various downstream tasks such as toxicity and hate speech detection (§5). Our research challenges the conventional practice of avoiding bias through obliviousness, advocating instead for the active recognition and intervention of bias as a more effective approach.

2. Background

2.1. Social Psychological Theories

Stereotyping is a cognitive process characterized by the tendency to generalize specific attributes to entire social groups. The manifestation of these stereotypical biases in society leads to adverse consequences, including the marginalization of certain groups from an equitable place in society, the exacerbation of social inequalities in resource allocation, and the psychological impact on individuals due to the awareness and internalization of these biases (Timmer, 2011).

Modern social psychology theories take a multifaceted approach to characterizing stereotypes associated with social groups, moving beyond the dichotomous categorization of these stereotypes as strictly positive or negative. The Stereotype Content Model (SCM) (Fiske et al., 2002; Fiske, 2018) introduces two fundamental dimensions in social perception: *warmth* and *competence*. The SCM is based on the premise that individuals aim to assess both the intentions directed towards them (*warmth*) and the abilities to fulfill those intentions (*competence*) within their social context. A notable insight from the SCM is that the presence of positive stereotypes along one dimension does not necessarily negate the presence of negative stereotypes along the other dimension. The interplay between *warmth* and *competence* gives rise to distinct stereotypical emotions, such as pity for groups perceived as warm but incompetent and envy for groups perceived as cold but competent.

The Agency-Belief-Communion (ABC) theory (Koch et al., 2016, 2020) employs a data-driven approach to identify traits that better explain stereotypes. The resulting list of traits reveals two dimensions strongly correlated with stereotypes: *agency* (competence or socioeconomic success) and *beliefs* (polarity along the conservative-progressive spectrum). Although communion (warmth) does not exhibit a direct correlation with stereotypes, it is associated with the proximity to the center along the agency and beliefs dimensions. In line with this ABC theory, our ABLE method models stereotypes along the *agency* and *beliefs* dimensions.

Recent developments in NLP research have increasingly integrated insights from the aforementioned social psychological theories. A series of studies have explored biases within NLP models, specifically in relation to the stereotype dimensions identified by these theories. For instance, Fraser et al. (2021) identified the SCM subspace within word embedding space and analyzed benchmark datasets on stereotypical bias. Herold et al. (2022) scrutinized how NLP models associate the SCM dimensions with disabled people. Cao et al. (2022b) introduced a novel association test metric, applying it to study the stereotypical biases along the ABC dimensions that NLP models encode for various demographic groups, including intersectional groups. In a related vein, Davani et al. (2023) examined the impact of the SCM dimensions on labeling and model performance for hate speech detection. These studies collectively highlight that both SCM and ABC theories offer valuable frameworks for analyzing stereotypical biases in NLP models.

In studies closely aligned with our research, the focus has been on mitigating the influence of stereotype dimensions within NLP models. Ungless et al. (2022) employed a method that identifies and re-

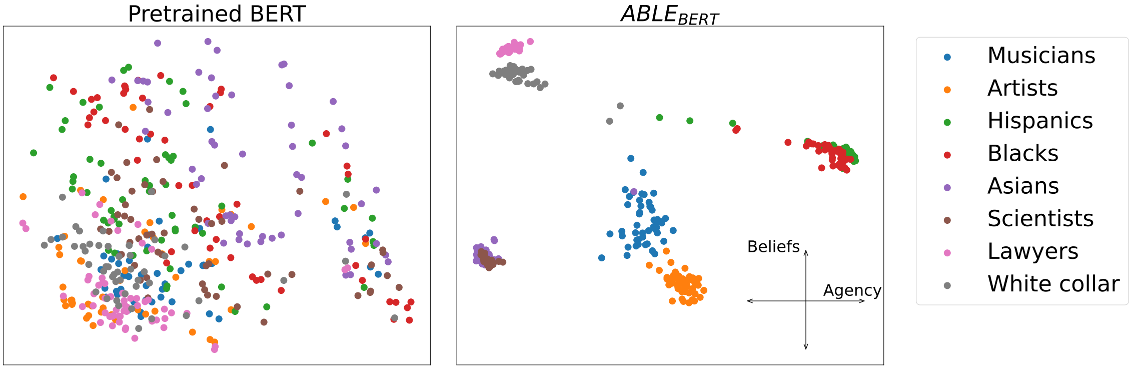

Figure 1: **Embedding Space Comparison.** We plot the first two principal components by applying PCA on the embedding spaces for pretrained BERT (left) and $\text{ABLE}_{\text{BERT}}$ (right). Each point corresponds to a single text sample color-coded by the demographic group. While the pretrained BERT embedding space does not exhibit a meaningful pattern, $\text{ABLE}_{\text{BERT}}$ puts embeddings for texts with the same demographic groups close together. Moreover, two directions correspond to agency and beliefs dimensions: from conservative on top to progressive on bottom and from competent on left to incompetent on right.

moves the SCM dimensions from the embedding space of language models to address bias. Building upon this work, Omrani et al. (2023) proposed two similar debiasing methods. However, all of the aforementioned studies adopt a bias-oblivious approach, in contrast to the bias-aware approach of our ABLE method.

2.2. Contrastive Learning

Contrastive learning aims to encode sample similarity in the embedding space of deep learning with minimal supervision. Given an anchor sample, the first step involves selecting similar samples as positives and dissimilar samples as negatives. Subsequently, the contrastive loss penalizes large anchor-positive distance and small anchor-negative distance. This pushes similar samples closer together and keeps dissimilar samples far apart.

Following Reimers and Gurevych (2019), we employ the contrastive loss based on an (anchor, positive, negative) triplet (a, p, n):

$$\mathcal{L}_{(a,p,n)} = \max(\|e_a - e_p\| - \|e_a - e_n\| + \epsilon, 0) \quad (1)$$

where e_a, e_p, and e_n denote the embedding vectors for the anchor, positive, and negative, respectively. The triplet loss in Equation 1 penalizes when the anchor is not closer to positive by ϵ than to negative in Euclidean distance.

In the ABLE method, we encourage texts with the same demographic group to embed closer together using contrastive learning. As shown in Figure 1, the embeddings that are originally scattered in the left Figure are clustered according to the demographic groups in the ABLE embedding space shown in the right Figure. For example, texts with the “Asians” group, denoted in purple, are embedded closer together into a cluster while farther away from different groups such as “Blacks”.

3. Methods

We describe the Agency-BeLiefs Embedding (ABLE) model in two key stages. In Section 3.1, we detail the creation of the Agency-Beliefs (AB) Dataset, wherein agency and beliefs scores are assigned to stereotyped sentences. In Section 3.2, we outline the multi-task learning objective function that the ABLE model uses to train.

3.1. AB Dataset

We generate the AB Dataset that consists of stereotypical sentences with corresponding agency and beliefs scores. A total of 4,229 stereotypical sentences containing 79 targets are sourced from the publicly available portion of StereoSet (Nadeem et al., 2021). StereoSet is a benchmark dataset designed to evaluate the stereotypical associations with targets in NLP models through fill-in-the-blank style tasks. A sample from the dataset has a stereotypical sentence, “The football player was pretty stupid” and a target, “football player”. These sentences are drawn from both intrasentence and intersentence tasks within StereoSet.

We assign agency and beliefs scores to each sentence based on the demographic group and the presence of stereotypical words in the sentence. To develop the theory-grounded dataset, the demographic groups are defined as the social groups from Study 1 in the ABC theory (Koch et al., 2016). For the AB dataset, the collected sentences are labeled with the demographic groups in the ABC theory using the target information in the StereoSet. The example sentence, “The football player was pretty stupid” has the target “football player” and, hence, is mapped to the “Athletes” demographic group. See Appendix A for the details of the mapping.

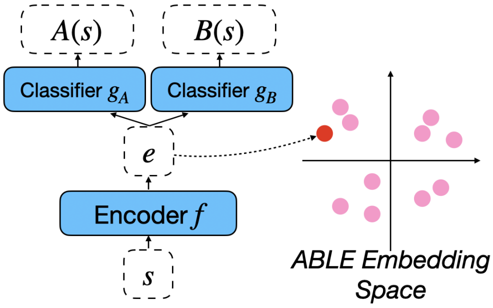

Figure 2: **ABLE Architecture.** Given a stereotypical sentence s, the encoder f generates its embedding $e = f(s)$. The embedding e predicts the agency and beliefs scores $A(s)$ and $B(s)$ using the classifiers g_A and g_B as $g_A(e)$ and $g_B(e)$, respectively.

To identify the stereotypical words, we employ a dictionary developed by Nicolas et al. (2019). The dictionary is constructed with an initial lexicon sourced from social psychology papers, expanded using WordNet. It contains 152 words for agency and 34 words for beliefs. Each word in the dictionary has a high or low direction, designated as +1 or -1, respectively. For example, under the agency dimension, “smart” has a high direction, denoted as +1, while “stupid” has a low direction, denoted as -1.

Given a stereotypical sentence s, we use the "target" feature in StereoSet as the demographic group d. Then, we identify sets W_A, W_B of agency and beliefs words appearing in s. The agency and belief scores $A(s)$ and $B(s)$ are defined as:

$$A(s) = \begin{cases} A(d) & \text{if } A(d) \neq 0 \\ \frac{1}{|W_A|} \sum_{w \in W_A} A(w) & \text{otherwise} \end{cases}$$

and

$$B(s) = \begin{cases} B(d) & \text{if } B(d) \neq 0 \\ \frac{1}{|W_B|} \sum_{w \in W_B} B(w) & \text{otherwise} \end{cases}$$

where $A(d)$ and $B(d)$ are agency and beliefs scores of the demographic group d derived from Study 1 in Koch et al. (2016). $A(w)$ and $B(w)$ are agency and beliefs scores of the word derived from Nicolas et al. (2019). We note that both A and B take values in $\{-1, 0, 1\}$. Table 1 presents examples in the resulting AB Dataset, and Table 2 shows the distribution of the agency and beliefs scores.

3.2. ABLE Training

We train the ABLE model using a multi-task learning objective, jointly optimizing for agency/beliefs score prediction and contrastive loss. Our model consists of an encoder f and two classifiers g_A and g_B as depicted in Figure 2.

As described in Section 2.1, the agency and beliefs dimensions for a given demographic group are effective predictors of stereotypes associated with the group. Accordingly, our approach involves encoding stereotypical bias into the model by training it to predict the agency and beliefs scores. We use two classifiers g_A and g_B to predict agency and beliefs scores, respectively. We formulate these prediction tasks as classification problems and employ cross-entropy losses, denoted as $\mathcal{L}_A$ and $\mathcal{L}_B$. Formally, $\mathcal{L}_A = -\sum_{s \in S} A(s) \cdot \log(g_A(s))$ and $\mathcal{L}_B = -\sum_{s \in S} B(s) \cdot \log(g_B(s))$, where s is a sentence in the dataset S.

However, our objective extends beyond accurate score prediction; it encompasses ensuring the coherence of the ABLE embedding space. In addition to accurate score prediction, we aim to map similar stereotypes to similar embedding vectors. Taking inspiration from Kim and Johnson (2022), we leverage triplet-based contrastive learning to improve the clustering of sentences related to the same demographic group.

For each stereotypical sentence s with demographic group d, we construct a single triplet (a, p, n) where the anchor a set to be s. The positive sample p is randomly selected from the sentences with the same demographic group, while the negative sample n is drawn from sentences with a different demographic group. The set of triplets is denoted as T. Our contrastive loss is:

$$\begin{aligned} \mathcal{L}_c &= \sum_{(a,p,n) \in T} \mathcal{L}_{(a,p,n)} \\ &= \sum_{(a,p,n) \in T} \max\left(\|e_a - e_p\| - \|e_a - e_n\|, 0\right). \end{aligned}$$

The final loss of the model is the sum of the prediction losses and the contrastive loss:

$$\mathcal{L} = \mathcal{L}_A + \mathcal{L}_B + \lambda \mathcal{L}_C$$

where λ is a hyperparameter that controls the relative importance of the prediction losses and the contrastive loss.

4. Model Validation

In this section, we conduct a comprehensive evaluation to ascertain the successful training of the ABLE model. Our assessment focuses on several key aspects: the model’s capacity to effectively cluster sentences with the same demographic group, the performance of agency and beliefs score predictions, and the model’s retained proficiency in language modeling as demonstrated through its performance on the GLUE benchmark.

Stereotypical Sentence	Demographic Group	Score$_{\text{Agency}}$	Score$_{\text{Beliefs}}$
The Muslim community is known for being strongly religious.	Muslims	0	1
She used to be a performing artist. She votes for liberal parties.	Artists	-1	-1
The smart chess player looked at the pieces on the board.	Educated	1	0
The football player was pretty stupid.	Athletes	-1	1

Table 1: **Examples in the AB Dataset.** We show examples with the corresponding demographic groups and scores. Words from the agency dictionary are colored in magenta, and those from the beliefs dictionary in teal.

		Score			
		-1	**0**	**1**	**Avg**
Agency	Word	5.6	86.0	8.3	0.029
	Demo	34.2	40.5	25.3	-0.089
	Final	35.3	36.6	28.2	-0.071
Beliefs	Word	0.6	97.2	2.2	0.018
	Demo	17.7	34.2	48.1	0.304
	Final	17.8	33.7	48.6	0.306

Table 2: **Score Distribution in the AB Dataset.** We show the percentage of each agency/beliefs score in stereotypical words from the dictionary, the demographic groups, and the aggregated final. The right-most column is the average score.

4.1. Experimental Settings

We explain the datasets and models used for the training and validation of the ABLE model. Implementation details are provided for reproducibility. The data and code for the experiments are available at `https://github.com/MSU-NLP-CSS/ABLE`.

Datasets. The AB Dataset comprises 4,229 stereotypical sentences representing 79 distinct demographic groups. To rigorously evaluate the ABLE model, we randomly select 10 of these demographic groups, which collectively account for 526 sentences, and designate them as the test set. The remaining 3,706 sentences constitute the training set for ABLE training.

Models. We use BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2020) in our experiments. BERT and RoBERTa are transformer-based bidirectional language models. For the classifiers used for agency/beliefs score prediction, we use a linear layer on top of the encoder. We use ABLE$_{\text{BERT}}$ and ABLE$_{\text{RoBERTa}}$ to denote the models after training BERT and RoBERTa on the ABLE objective, respectively. ABLE training starts from the pretrained BERT and RoBERTa models.

Model	Agency	Beliefs
ABLE$_{\text{BERT}}$	0.993 / 0.813	0.995 / 0.870
ABLE$_{\text{RoBERTa}}$	0.986 / 0.778	0.995 / 0.754

Table 3: **Agency/Beliefs Score Prediction Performance.** We verify that the ABLE models achieve strong performance for agency/beliefs score prediction. Each entry reflects: *training*/*testing* accuracy.

Implementation. All models are implemented with PyTorch (Paszke et al., 2019) and Huggingface's Transformers (Wolf et al., 2020). For ABLE training, we use the Adam optimizer (Kingma and Ba, 2015) and set the learning rate from $\{2e-05, 5e-05\}$, an epoch from $\{1, 3, 5\}$, and a dropout rate from $\{0.2, 0.5\}$. All experiments are conducted on an Nvidia Quatro RTX 5000, 16 GB memory GPU in a machine with Intel(R) Xeon(R) Silver 4214 CPU @ 2.20GHz.

4.2. Agency-Beliefs Predictions

This section validate the performance of the ABLE model in predicting agency/beliefs scores. The accuracy of agency/beliefs score prediction is a critical indicator of the ABLE model's ability to effectively encode stereotypical bias. Ensuring the model's proficiency in this prediction task is, therefore, a fundamental step in the validation process.

To enhance the robustness of our findings, each experiment is conducted three times, and the accuracy for agency/beliefs score prediction is averaged. Subsequently, this accuracy is averaged for each demographic group. Finally, we aggregate these accuracy once more to obtain the final accuracy, which we report in Table 3. The results demonstrate that the ABLE model achieves a reasonable performance in both agency and beliefs score prediction tasks.

For a more comprehensive analysis, we provide visual representations of the agency and beliefs scores for each demographic group. These visualizations are derived by calculating the average

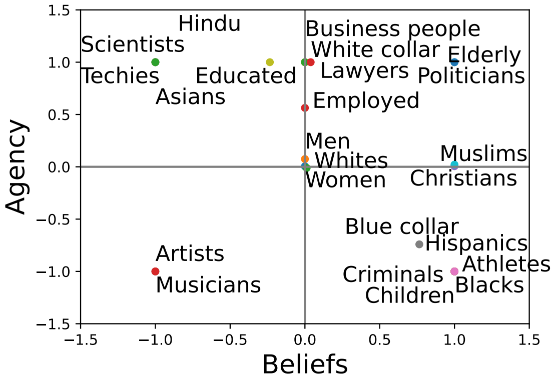

Figure 3: **Stereotypical Bias in the Agency-Beliefs Dimension.** The x-axis is the *beliefs* dimension, where the positive direction indicates conservative beliefs and the negative direction implies progressive beliefs. The *agency* dimension is along the y-axis, where higher agency means a higher chance of socio-economic success.

agency/beliefs scores from the ABLE model's output for each demographic group, as depicted in Figure 3.

Figure 3 supports the result of the ABC model study done on U.S. participants and demographic groups (We refer the readers to Figure 1 in Koch et al. (2016)). Along the agency dimension, groups "Business people," "Lawyers," and "White collar" are observed as the ones with the highest agency, namely, more likely to have socio-economic success. On the other hand, groups "Criminals," "Blacks," and "Hispanics" display low agency. As for the beliefs dimension, the positive axis means having conservative beliefs, while the negative suggests progressive beliefs. The model predicted high beliefs scores for "Muslims," while low scores for "Musicians" and "Artists." Just as in the ABC model, groups "Men," "Women," and "Whites" are located in the origin. One difference is that the predicted agency score of "Hindu" is lower than that in the ABC model study, where it ranked the highest score among demographic groups. A possible reason is the lack of seed words in the StereoSet data that belong to the group "Hindu."

4.3. Demographic Groups after ABLE Training

The ABLE model trains to render consistent embeddings through contrastive learning. The contrastive loss in the ABLE training encourages the sentences within the same demographic group to stay close to each other. To validate that the ABLE model clusters the embedding vectors well, we measure the isotropy following Arora et al. (2016); Mu and Viswanath (2018).

A set of vectors are called *isotropic* if they are uniformly distributed in all directions. If the embedding vectors are more isotropic, then the embedding vectors are less clustered. Therefore, we expect our embedding vectors will have smaller isotropy measure.

Model	Train Isotropy	Test Isotropy
BERT	1.30e-06	6.73e-08
ABLE$_{\text{BERT}}$	↓ 1.08e-06 2.23e-07	7.35e-07
RoBERTa	3.58e-06	2.33e-06
ABLE$_{\text{RoBERTa}}$	↓ 3.50e-06 7.93e-08	↓ 2.28e-06 4.53e-08

Table 4: **Isotropy Measures.** The isotropies of the embedding space before and after the training of the ABLE model are compared. Higher values indicate strong isotropy.

Drawing upon Mu and Viswanath (2018), we first define the partition function

$$Z(u) = \sum_{i=1}^{N} e^{u^T w_i},$$

for each unit vector u. Then, the isotropy measure of the embedding matrix $W = [w_1, ..., w_N]$ is defined as

$$\mathcal{I}(W) = \frac{\min_{\|u\|=1} Z(u)}{\max_{\|u\|=1} Z(u)}.$$

We compare the isotropy of the ABLE models to the pretrained models. The results are shown in Table 4. As expected, the ABLE space has lower isotropy measures in the stereotype embedding spaces compared to the pretrained models. That is, after learning stereotypes, the embedding space becomes more anisotropic. The difference in isotropy is more significant in RoBERTa than in BERT. As for RoBERTa, after fine-tuning, the isotropy drops 10^{-2} times when computed with the embeddings of training data groups. The isotropy declines in a similar ratio when computed with the embeddings of test data groups. Based on the decrease of isotropy after fine-tuning, we infer that the ABLE model pushes the same demographic groups closer in the ABLE space.

The visualization of embedding space also supports our inference. Figure 1 displays the projection of embeddings of texts that mention stereotypes of the chosen demographic groups: Asians, Blacks, Hispanics, Artists, Lawyers, Musicians, Scientists, and White Collar. The left figure reveals that embeddings of the pretrained model are spread out across the space. On the other hand, in the right figure, the embeddings within the groups are closer. The agency and beliefs directions can also be observed based on the positions of demographic groups. On the lower side, the groups Musicians and Artists with liberal beliefs are positioned; on the upper

	BERT	$\text{ABLE}_{\text{BERT}}$	RoBERTa	$\text{ABLE}_{\text{RoBERTa}}$
CoLA	0.544	0.550	0.559	0.555
MRPC	0.833	0.817	0.872	0.877
RTE	0.656	0.651	0.692	0.684
SST	0.924	0.926	0.940	0.942
STS-B	0.888	0.886	0.893	0.895
WNLI	0.563	0.563	0.563	0.563

Table 5: **GLUE Benchmark.** We report the Spearman correlation for STS-B, Matthew's correlation for CoLA, and the accuracy for all other tasks. Reported results are averaged over three runs.

side, the groups with conservative beliefs, such as Lawyers and Hispanics. The groups with high agency are placed on the left side, e.g., Asians and Scientists, while those with low agency, for example, Blacks, are on the right.

4.4. GLUE Benchmark

Finally, we check if the ABLE model retains the proficiency of language models by testing on the GLUE benchmark (Wang et al., 2018). Following Omrani et al. (2023) and Kaneko and Bollegala (2021), the GLUE benchmark tasks with small-scale training data are chosen to demonstrate that the debiased models have minimal effects due to task-specific fine-tuning. We report the performance of the ABLE model on the six tasks in Table 5. On all six tasks, the ABLE model performs competitively. The average accuracy when using ABLE drops slightly on the RTE task for both $\text{ABLE}_{\text{BERT}}$ and $\text{ABLE}_{\text{RoBERTa}}$. Yet we also observe improved performance in other tasks such as COLA and STS-B for $\text{ABLE}_{\text{BERT}}$. Based on these experimental results from GLUE, we conclude that the ABLE model maintains the language models' proficiency and does not lose their generalization ability.

5. Model Applications

In this section, we delve into the utilization of ABLE models for tasks that can leverage stereotype information. Specifically, we focus on toxicity detection and hate speech detection as our chosen tasks for analysis.

Datasets. For toxicity, we use the Jigsaw Unintended Bias in Toxicity Classification Dataset [2]. Jigsaw is a crowd-sourced toxicity dataset of over 2 million public comments. Each comment is assigned a toxicity score and labeled as toxic if the score is greater than or equal to 0.5. For hate speech, the Measuring Hate Speech Dataset (Kennedy et al., 2020; Sachdeva et al., 2022) is utilized. The Measuring Hate Speech corpus (MHS) consists of 50,070 social media comments spanning YouTube, Reddit, and Twitter, labeled with a hate speech score by 11,143 annotators. Similar to Jigsaw, a comment is labeled as hate speech if its hate speech score is greater than 0.5.

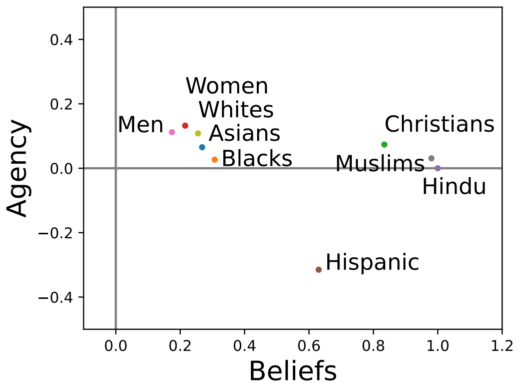

Figure 4: **Toxicity in the Agency-Beliefs Dimension.** Agency and belief dimensions of target groups in the toxicity dataset Jigsaw are shown.

5.1. Stereotyping and Toxicity

We discuss the relationship between stereotyping and toxicity. Using the trained ABLE model, we measure the agency and beliefs scores of the groups in Jigsaw that overlap with those in the ABC study. Similar to the mapping done on the AB dataset for Figure 1, the target groups in Jigsaw are mapped to the demographic groups in the ABC model for better analysis. The details of the mapping are provided in Appendix A.

The average agency and beliefs scores of each group are plotted and shown in Figure 4. Interestingly, we observe many overlaps in the groups' positions in Figures 3 and 4. These overlaps suggest that stereotypical biases are reflected in the toxicity data. Namely, when a comment mentions a stereotype of a particular group, the comment is likely to be considered toxic. Although most demographic groups showed overlaps, a few groups, such as "Asians" and "Hindu," are not plotted in a similar location as in Figure 3. The ABLE model's prediction of agency and beliefs scores of these groups do not align with stereotypical biases of the groups. We surmise that the reason for this misalignment may come from the fact that the majority of toxic comments on those groups have low agency and high beliefs scores. That is, comments with low agency and high beliefs scores had high toxicity scores.

Next, we take a more detailed look into the agency and beliefs predictions on toxicity data. Ta-

[2] https://www.kaggle.com/competitions/jigsaw-unintended-bias-in-toxicity-classification/data

		BERT				ABLE$_{\text{BERT}}$			
		10k	20k	40k	80k	10k	20k	40k	80k
Gender	Men	0.716	0.814	0.825	0.896	0.698	0.768	0.857	0.866
	Women	0.750	0.826	0.851	0.922	0.713	0.789	0.864	0.890
	Trans	0.716	0.791	0.806	0.881	0.612	0.761	0.821	0.851
Religion	Chrisitan	0.752	0.845	0.870	0.951	0.702	0.814	0.882	0.916
	Jewish	0.788	0.827	0.863	0.906	0.663	0.820	0.827	0.886
	Muslim	0.725	0.748	0.770	0.824	0.633	0.750	0.760	0.810
	Hindu	0.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
	Buddhist	1.0	0.667	0.667	1.0	1.0	1.0	0.667	1.0
	Atheist	0.736	0.868	0.934	0.983	0.777	0.810	0.942	0.893
Race/ Enthic.	Black	0.659	0.714	0.730	0.778	0.604	0.680	0.735	0.754
	White	0.659	0.722	0.751	0.790	0.615	0.684	0.756	0.776
	Asian	0.794	0.856	0.875	0.975	0.669	0.806	0.869	0.95
	Latino	0.7	0.843	0.843	0.929	0.671	0.814	0.829	0.914
	Other	0.625	0.875	0.625	0.875	1.0	0.875	0.75	0.875

Table 6: **Toxicity Prediction Among Demographic Groups.** On gender, religion, and race/ethnicity biases, we compare the distribution of accuracy of BERT and ABLE$_{\text{BERT}}$ according to the training data size (10k, 20k, 40k, and 80k).

ble 6 reports the accuracies of toxicity prediction on each demographic group. BERT and ABLE$_{\text{BERT}}$ are trained on different sizes of training data: 10,000, 20,000, 40,000, and 80,000. As expected, we observe an improvement in performance as the training data size increases for both models. The performance of the models is also compared according to the stereotypical bias types: gender, religion, and race/ethnicity. In all bias cases, ABLE$_{\text{BERT}}$ shows a more even performance across demographic groups than BERT. For instance, for religion predictions of the models trained on 10k, the maximum difference across groups for BERT is 1.0, while the maximum difference for ABLE$_{\text{BERT}}$ is 0.367. Yet in other training settings, we acknowledge that the difference between BERT and ABLE$_{\text{BERT}}$ is not as significant.

Our proposed method, ABLE, displays competitive performance across experiments on toxicity detection. When toxic comments are projected onto the ABLE space, they display overlaps with the stereotyping plot, allowing an understanding and analysis of the spectrum of stereotypical biases. Also, ABLE$_{\text{BERT}}$ performs competitively across different stereotypical biases in different training settings. These results suggest that there are high correlations between toxicity and stereotyping and that the ABLE architecture can be expanded to tasks pertinent to stereotypical biases.

5.2. Stereotyping and Hate Speech

To further examine the application of the ABLE model on tasks related to stereotypical biases, we compare the performance of ABLE across demographic groups on hate speech detection. BERT and ABLE$_{\text{BERT}}$ are fine-tuned with the training data of size 20. The models' prediction accuracies on test data are measured to compare the performance.

As shown in Table 7, the proposed ABLE model improves hate speech detection throughout all demographic groups except “Jewish”. For the bias type gender, accuracy improves the most in “Men”. The performance of "Atheist" shows the sharpest increase in the context of religion. For race/ethnicity, the greatest improvement occurs in “White”.

We observe that the ABLE model, enriched with stereotype information, demonstrates outstanding performance on the Measuring Hate Speech Dataset. Our hypothesis posits that comments containing or amplifying stereotypes are more likely to be classified as hate speech.

6. Conclusion

Our proposed Agency-BeLiefs Embedding (ABLE) model represents a proactive approach to learning stereotypical biases within the model's embedding space. ABLE stands in stark contrast to the conventional bias-oblivious methods used to address stereotypical biases. Motivated by the latest social psychology research on stereotyping, the ABLE model acquires stereotypical biases in the dimensions of agency and beliefs. To maintain consistency within the embedding space, we employ contrastive learning, encouraging sentences sharing the same target groups to be closer to each other.

Through extensive empirical evaluations, we validate that the ABLE model effectively learns to predict agency/beliefs scores while preserving its lan-

		BERT	$\text{ABLE}_{\text{BERT}}$
Gender	Men	0.809	0.956
	Women	0.739	0.844
	Trans	0.852	0.985
Religion	Chrisitan	0.847	0.991
	Jewish	0.418	0.355
	Muslim	0.865	0.998
	Hindu	0.792	0.896
	Buddhist	0.767	0.933
	Atheist	0.696	0.957
Race/ Enthic.	Black	0.623	0.664
	White	0.819	0.947
	Asian	0.617	0.678
	Latino	0.691	0.773
	Other	0.802	0.924

Table 7: **Hate Speech Prediction Among Demographic Groups.** On gender, religion, and race/ethnicity biases, we compare the distribution of accuracy of BERT and $\text{ABLE}_{\text{BERT}}$.

guage modeling proficiency. Furthermore, we illustrate the practical significance of incorporating stereotypes with experiments on downstream tasks: toxicity and hate speech detection. The experimental results hint at the broader potential applications of the ABLE model, encompassing the curation of stereotype data, the analysis of stereotypes within texts and language models, and more. This evidence underscores the notion that addressing bias through awareness may indeed yield more substantial benefits than attempting to mitigate it blindly.

7. Limitations

Our work focuses on datasets and models that are entirely in the English language. Moreover, the stereotypes in our study are mostly U.S.-based and are expressed in English. We call for the replication of our work on multi-lingual datasets with diverse cultural backgrounds.

Moreover, our AB dataset is constructed based on various hand-designed rules and the authors' judgments. We leave the automatic procedure to encode stereotypical biases into language models as future work.

8. Ethical Considerations

The data and code for the ABLE model are open to the public and thus can be used to study stereotypical bias. The AB dataset, used for both training and testing of the ABLE model, assumes a particular framework for coding stereotypes. However, this framework may not encompass the full range of stereotypes, limiting the dataset's scope. Therefore, while valuable, the AB dataset offers a narrow perspective on stereotypes. Researchers should approach its findings with caution, supplementing them with other methodologies to achieve a more comprehensive understanding of stereotypes.

Finally, we advise not to use this research for malicious intentions, such as amplifying and spreading stereotypical biases.

9. Bibliographical References

Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. 2016. A Latent Variable Model Approach to PMI-based Word Embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399.

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. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Anne Boring and Arnaud Philippe. 2021. Reducing discrimination in the field: Evidence from an awareness raising intervention targeting gender biases in student evaluations of teaching. *Journal of Public Economics*, 193:104323.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 370–378, New York, NY, USA. Association for Computing Machinery.

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

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022a. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.

Yang Trista Cao, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022b. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North Amer-*

ican Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Pengyu Cheng, Weituo Hao, Siyang Yuan, Shijing Si, and Lawrence Carin. 2021. Fairfil: Contrastive neural debiasing method for pretrained text encoders. In *International Conference on Learning Representations*.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Zahra Fatemi, Chen Xing, Wenhao Liu, and Caimming Xiong. 2023. Improving gender fairness of pre-trained language models without catastrophic forgetting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1249–1262, Toronto, Canada. Association for Computational Linguistics.

Susan T. Fiske. 2018. Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science*, 27(2):67–73. PMID: 29755213.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. 2002. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of personality and social psychology*, 82(6):878.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. 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 600–616, Online. Association for Computational Linguistics.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022. Debiasing pretrained text encoders by paying attention to paying attention. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9582–9602, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

Jacqueline He, Mengzhou Xia, Christiane Fellbaum, and Danqi Chen. 2022. MABEL: Attenuating gender bias using textual entailment data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9681–9702, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brienna Herold, James Waller, and Raja Kushalnagar. 2022. Applying the stereotype content model to assess disability bias in popular pre-trained NLP models underlying AI-based assistive technologies. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 58–65, Dublin, Ireland. Association for Computational Linguistics.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On transferability of bias mitigation effects in language model fine-tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough!

– on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Michelle YoungJin Kim and Kristen Marie Johnson. 2022. CLoSE: Contrastive learning of subframe embeddings for political bias classification of news media. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2780–2793, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The abc of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of personality and social psychology*, 110(5):675.

Alex Koch, Roland Imhoff, Christian Unkelbach, Gandalf Nicolas, Susan Fiske, Julie Terache, Antonin Carrier, and Vincent Yzerbyt. 2020. Groups' warmth is a personal matter: Understanding consensus on stereotype dimensions reconciles adversarial models of social evaluation. *Journal of Experimental Social Psychology*, 89:103995.

Deepak Kumar, Oleg Lesota, George Zerveas, Daniel Cohen, Carsten Eickhoff, Markus Schedl, and Navid Rekabsaz. 2023. Parameter-efficient modularised bias mitigation via AdapterFusion. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2738–2751, Dubrovnik, Croatia. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen McKeown, and Tatsunori Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3206–3219, Dubrovnik, Croatia. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Cynthia Lee. 2017. Awareness as a first step toward overcoming implicit bias. *Enhancing justice: Reducing bias; GWU Law School Public Law Research Paper No. 2017-56; GWU Legal Studies Research Paper No. 2017-56*, 289.

Yingji Li, Mengnan Du, Xin Wang, and Ying Wang. 2023. Prompt tuning pushes farther, contrastive learning pulls closer: A two-stage approach to mitigate social biases. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14254–14267, Toronto, Canada. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Ro{bert}a: A robustly optimized {bert} pretraining approach.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Jiaqi Mu and Pramod Viswanath. 2018. All-but-the-top: Simple and effective postprocessing for word representations. In *International Conference on Learning Representations*.

Gandalf Nicolas, Xuechunzi Bai, and Susan T Fiske. 2019. Automated dictionary creation for analyzing text: An illustration from stereotype content.

Ali Omrani, Alireza Salkhordeh Ziabari, Charles Yu, Preni Golazizian, Brendan Kennedy, Mohammad Atari, Heng Ji, and Morteza Dehghani. 2023. Social-group-agnostic bias mitigation via the stereotype content model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4123–4139, Toronto, Canada. Association for Computational Linguistics.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Devin G. Pope, Joseph Price, and Justin Wolfers. 2018. Awareness reduces racial bias. *Management Science*, 64(11):4988–4995.

Rebecca Qian, Candace Ross, Jude Fernandes, Eric Michael Smith, Douwe Kiela, and Adina Williams. 2022. Perturbation augmentation for fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9496–9521, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan D Cotterell. 2022. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Aili Shen, Xudong Han, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2022. Does representational fairness imply empirical fairness? In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 81–95, Online only. Association for Computational Linguistics.

Alexandra Timmer. 2011. Toward an Anti-Stereotyping Approach for the European Court of Human Rights. *Human Rights Law Review*, 11(4):707–738.

Eddie Ungless, Amy Rafferty, Hrichika Nag, and Björn Ross. 2022. A robust bias mitigation procedure based on the stereotype content model. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 207–217, Abu Dhabi, UAE. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. 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 1667–1682, Online. Association for Computational Linguistics.

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, Brussels, Belgium. Association for Computational Linguistics.

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*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi 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 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*, pages 38–45, Online. Association for Computational Linguistics.

Zhongbin Xie and Thomas Lukasiewicz. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15730–15745, Toronto, Canada. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Fan Zhou, Yuzhou Mao, Liu Yu, Yi Yang, and Ting Zhong. 2023. Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4227–4241, Toronto, Canada. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

10. Language Resource References

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, Online. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

A. Mapping of the StereoSet and the ABC Model

The target groups of the StereoSet, the Jigsaw, and the MHS are mapped to one of the 80 demographic groups in the ABC model for the visualization of Figures 3 and the experiments in Section 5. The mappings of the Steresoet are shown in Table 10. The mappings of the Jigsaw are shown in Table 8. Finally, the mappings of the MHS are shown in Table 9.

Asians	asian
Atheist	atheist
Blacks	black
Buddhist	buddhist
Christians	christian
Women	female
Hindu	Hindu
Hispanic	latino
Jewish	jewish
Men	male
Muslims	muslim
Trans	transgender
Whites	white

Table 8: **Mappings of the Target Groups of the Jigsaw and the Demographic Groups in the ABC Model.** The first column lists the demographic groups in the ABC model. The second column is the target groups of the Jigsaw that are assigned to the corresponding demographic group.

Asians	asian
Atheist	atheist
Blacks	black
Buddhist	buddhist
Christians	christian
Women	female
Hindu	Hindu
Hispanic	latinx
Jewish	jewish
Men	male
Muslims	muslim
Trans	transgender men transgender women transgender unspecified
Whites	white

Table 9: **Mappings of the Target Groups of the MHS and the Demographic Groups in the ABC Model.** The first column lists the demographic groups in the ABC model. The second column is the target groups of the MHS that are assigned to the corresponding demographic group.

Artists	performing artist
Asians	Japanese, Vietnam
Blacks	African, Ethiopian, Somalia, Sierra Leon, Ethiopia, Eriteria, Ghanaian, Eritrean, Cameroon, Cape Verde
Blue collar	plumber, policeman, butcher, mover, delivery man, tailor
Business people	entrepreneur
Children	schoolgirl, schoolboy
Christians	Bible
Criminals	prisoner
Educated	chess player, psychologist, historian, researcher, mathematician
Elderly	grandfather
Employed	manager, assistant, commander, producer
Hindu	Bengali, Brahmin, Bangladesh
Lawyers	prosecutor
Hispanics	Columbian, Ecuador, Hispanic
Men	gentlemen, male, himself
Musicians	guitarist, musician
Muslims	Lebanon, Saudi Arabian Afghanistan, Syria, Muslim, Iranian, Morocco, Yemen, Persian people, Iraq, Arab
Scientists	chemist, physicist
Politicians	politician
Techies	software developer, engineer
White collar	civil servant
Whites	Russian, Italy, Britain, Spain, Crimean, Ukrainian, Norway, Norweigan
Women	mommy, sister, mother, herself
Athletes	football player

Table 10: **Mappings of the Target Groups of the StereoSet and the Demographic Groups in the ABC Model.** The first column lists the demographic groups in the ABC model. The second column is the target groups of the StereoSet that are assigned to the corresponding demographic group.