A Robust Bias Mitigation Procedure Based on the Stereotype Content Model

Eddie L. Ungless and **Amy Rafferty** and **Hrichika Nag** and **Björn Ross**
School of Informatics
University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB, United Kingdom
e.l.ungless@sms.ed.ac.uk a.rafferty@live.com
naghrichika@gmail.com b.ross@ed.ac.uk

Abstract

The Stereotype Content model (SCM) states that we tend to perceive minority groups as cold, incompetent or both. In this paper we adapt existing work to demonstrate that the Stereotype Content model holds for contextualised word embeddings, then use these results to evaluate a fine-tuning process designed to drive a language model away from stereotyped portrayals of minority groups. We find the SCM terms are better able to capture bias than demographic agnostic terms related to pleasantness. Further, we were able to reduce the presence of stereotypes in the model through a simple fine-tuning procedure that required minimal human and computer resources, without harming downstream performance. We present this work as a prototype of a debiasing procedure that aims to remove the need for *a priori* knowledge of the specifics of bias in the model.

1 Introduction

It is well established that large language models (LLMs) such as BERT (Devlin et al., 2019), GPT2 (Radford et al., 2019) and related contextualised word embeddings such as ELMo (Peters et al., 2018) are biased against different demographic groups (Guo and Caliskan, 2021; Webster et al., 2020; Kaneko and Bollegala, 2021), in that they often reflect stereotypes in their output. For example, given the prompt "naturally, the nurse is a", these systems will typically output "woman" (Schick et al., 2021). Given the common practice of adapting pre-trained language models for a range of tasks through fine-tuning, upstream bias mitigation may prove to be the most efficient solution (Jin et al., 2021) (though cf. Steed et al. (2022). In this paper, we demonstrate the success of modifying an existing debiasing algorithm to be grounded in a psychological theory of stereotypes - the SCM (Cuddy et al., 2008), to efficiently reduce biases in LLMs across a range of identities. Our proposed debiasing pipeline has the benefit of minimising the time spent researching identity terms and associated stereotypes. Being a fine-tuning procedure, this also reduces the amount of computational resources needed compared to training an unbiased model from scratch. This renders our approach efficient and widely applicable. We demonstrate using BERT, but this same procedure could easily be adapted to other LLMs.

We adapt the fine-tuning procedure from Kaneko and Bollegala (2021). They reduce gender bias in a range of LLMs by fine-tuning using a data set of sentences containing (binary) gendered terms (like "he, man" or "she, lady") (which they call attributes) or stereotypes associated with different genders ("assertive, secretary") (which they call targets). The training objective is to remove associations with gender in the contextualised embeddings of the targets whilst maintaining these associations for the gendered attributes.

Crucially, rather than relying on stereotypes specific to a particular demographic such as men and women (as in Kaneko and Bollegala (2021)) we plan to use the SCM to inform our production of fine-tuning data, inspired by work by Fraser et al. (2021). The SCM states that our stereotyped perception of different demographics can be conceptualised as lying in a vector space with axes of warmth/coldness and competence/incompetence (Cuddy et al., 2008). We tend to consider our own identity group to be warm and competent, and stereotype disfavoured groups such as people experiencing homelessness as cold and/or incompetent (Cuddy et al., 2008).

In the terminology of Kaneko and Bollegala (2021), our attributes are terms relating to warmth and competence taken from Nicolas et al. (2021) (as in Fraser et al. (2021), a paper on stereotypes in static embeddings), our targets are demographic identity terms. Because the SCM is designed to encompass many different minority groups, this avoids the need to generate lists of stereotypes

Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS), pages 207 - 217

unique to each minority group, reducing work load and making the tool easy to adapt to different targets. Therefore, the procedure should be effective for all identity terms we use. We demonstrate this technique for Black/white ethnicity and also the intersectional power dynamic between white men and Mexican American women, but this could easily be expanded to other aspects of identity such as disability and sexuality. Further, whilst we focus on English language and American identities, there is evidence that the SCM may hold relatively well cross-culturally (Cuddy et al., 2009), so this approach may be transferable to other LLMs.

We adapt the Contextualised Embedding Association Test (CEAT) (Guo and Caliskan, 2021) using the vocabulary from Nicolas et al. (2021) in order to measure stereotypes in contextualised word embeddings. The CEAT provides a robust measure of bias in contextualised word embeddings for target words, and is suited for use with the SCM terms.

In addition to using the CEAT to test for bias, we also measure the performance of the model on the language modeling benchmark GLUE (Wang et al., 2018), to ensure the fine-tuning procedure does not adversely impact the quality of the model, an issue Meade et al. (2022) identify as affecting several debiasing techniques.

The main contributions of this paper are to demonstrate:

- that the SCM can be used to detect bias in contextualised word embeddings
- a debiasing procedure that is demographic agnostic and resource efficient[1]

2 Related work

Several contributions have been made towards measuring and mitigating bias in NLU models with minimal *a priori* knowledge. Fraser and colleagues (2021) demonstrated the validity of the SCM for static word embeddings, in that the embeddings of words associated with traditionally oppressed minority groups such as Mexican Americans or Africans tend to lie in the cold, incompetent space, as determined by cosine similarity. Note that, unlike Fraser et al. (2021), we focus on the embeddings of the identity terms themselves, not of words associated with those identities, as we explicitly want to identify whether there is bias in the embeddings. Fraser et al. (2021) looked to establish if the embeddings of associated terms followed the SCM's predictions, not whether the word embeddings were biased in a way as to reflect these stereotypes.

Utama et al. (2020) propose a strategy for debiasing "unknown biases". They train a shallow model which picks up superficial patterns in data that are likely to indicate bias. This is then used to train the main model, which works by downweighting the potentially biased examples, paired with an annealing mechanism which prevents the loss of useful training signals caused by this approach. The models obtained from this self-debiasing framework were shown to perform just as well as models debiased using prior knowledge. In our work we do not train our model from scratch and only focus on social bias, whereas Utama et al. (2020) do not target specific bias types. We chose to prioritise socially relevant biases with the hopes of minimising harm done to minority communities. Further, our method requires far less compute.

Webster et al. (2020) take gendered correlations in pretrained language representations as a case study for measuring and mitigating bias. They build an evaluation framework for detecting and quantifying gendered correlations in models. They find that both dropout regularization and counterfactual data augmentation minimize gendered correlations while maintaining strong model accuracy. Their techniques are applicable when training a model from scratch, whilst ours is a fine-tuning procedure, meaning it requires fewer computational resources.

Schick et al. (2021) explore whether language models can self-diagnose undesirable outputs for self-debiasing purposes. Their approach encourages the model to output biased text, and uses the resulting distribution to tune the model's original output. We argue that our model is more demographic agnostic, as their approach depends heavily on biases captured by Perspective API. Their approach may miss less salient forms of bias as it relies on the model having some representation of the bias category beforehand. Using the SCM, we can work "backwards" from the fact that these communities are harmed to then assume they will be represented as cold and/or incompetent, making our approach more universally applicable.

Cao et al. (2022b) focuses on identifying stereotyped group-trait associations in language models, by introducing a sensitivity test for measur-

[1]Code available at https://github.com/MxEddie/Demagnosticdebias

ing stereotypical associations. They compare US-based human judgements to language model stereotypes, and discover moderate correlations. They also extend their framework to measure language model stereotyping of intersectional identities, finding problems with identifying emergent intersectional stereotypes. Our work is unique from this in that we have additionally performed debiasing informed by the SCM.

Overall, our methodology and approach differs from most other contributions in this field as it focuses on targeting social bias specifically, and we propose a fine-tuning debiasing approach which requires little in the way of human or computer resources and is not limited to a small number of demographics.

3 Data sets and tasks

3.1 Data for Debiasing Procedure

3.1.1 Identity terms (targets)

We established two sets of identity terms (targets) for use with the context debiasing algorithm. The first set relates to racial bias (bias against people of colour based on their (perceived) race). BERT has been shown to demonstrate racial bias in both intrinsic (Guo and Caliskan, 2021) and extrinsic measures (Nadeem et al., 2021; Sheng et al., 2019). To reduce bias against Black people compared to white, we created a list of 20 African American (AA) and 20 European American (EA), 10 male and 10 female names for each, to use in the debiasing procedure. We used names from Guo and Caliskan (2021) (excluding any included in the CEAT tests we deploy, see Section 3.2) and supplemented these lists with common names from a database of US first names (Tzioumis, 2018). Excluding names from the CEAT tests was crucial to ensure a reduction in bias was due to a restructuring of the embedding space and an overall change in how Black individuals were represented, and not due to bias reduction for the specific names we ran the debiasing procedure with.

The second set relates to intersectional bias against Mexican American (MA) women, that is bias against women based on both patriarchal beliefs about their gender and prejudice against their ethnicity. This intersectional bias is evident in the contextualised embeddings BERT produces (Guo and Caliskan, 2021). To reduce bias against MA women compared to white men, we additionally took 10 common Hispanic female names (and manually confirmed that each was used by the Mexican American community through a Google search) from Tzioumis (2018).

The validity of using names to represent demographic groups has been questioned (Blodgett et al., 2021). However, we assume that reducing bias present in the representations of these names will go some way to reducing racial bias in the model.

3.1.2 Stereotype Content terms (attributes)

As with Fraser et al. (2021), we use the Stereotype Content terms from Nicolas et al. (2021), whereby the high morality, high sociability terms are taken to indicate warmth; low morality, low sociability to indicate coldness; high ability, high agency to indicate competence; and low ability, low agency to indicate incompetence. We selected the top 32 most frequent terms from each list (as measured using the Brown Corpus and the NLTK toolkit), to increase the likelihood we would find a large number of example sentences for each. During finetuning, we wish for these terms to maintain their projection in the warmth/coldness or competence/incompetence space, respectively, whilst removing projection in these directions for the target terms (see Section 4 and Figure 1).

Whilst the exact "position" of demographic groups in this conceptual space would vary depending on who is describing them, in this work we always assume the minority group will be represented in the original model as cold and incompetent, in other words the most disfavoured and most likely to experience harm (Cuddy et al., 2008). This minimises workload (no need to establish likely predictions for every demographic considered, beyond identifying the more marginalised group) and centers our approach around improving results for the most negatively represented identity terms. Note, there is no harm in running our debiasing procedure on identities that are already equally associated with one concept i.e. warmth, whilst also reducing stereotyped associations with the other concept i.e. competence.

3.1.3 Fine-tuning data

Having established the list of attribute and target terms, we follow an adapted version of Kaneko and Bollegala (2021)'s procedure for generating fine-tuning development data. During early analyses, we found the AA names occurred very infrequently in their provided news commentary data set, likely a reflection of the lack of AA represen-

tation in mainstream news (Diuguid and Rivers, 2000). We therefore opted to use data from Reddit, from 2018[2], (a separate data set to that used for the CEAT, see below), as this contained many example sentences across all names. We sampled from this data set sentences which contained either one of the attribute or one of the target terms, and no more, of 128 tokens or less. We extracted at least 24,000 sentences for each attribute and target dimension. This was stored as a dictionary that was passed to the debiasing script. We took a random sub-sample of 1000 of each to use as development data.

3.2 CEAT

The CEAT (Guo and Caliskan, 2021) is designed to test for associations between the contextualised embeddings of targets and polar attributes (such as binary gender). The authors sampled sentences from Reddit where a stimuli (target or attribute term) occurred, and generated contextualised embeddings for the sentences. These contextualised embeddings were then used to calculate the effect sizes, based on a cosine similarity measure between the embeddings of the target and attribute tokens. They then measure the distribution of effect sizes for the terms in different contexts (to ensure that the choice of context does not unduly influence the final effect size metric). The authors then apply a random-effects model to calculate a combined effect size (CES) and significance, given the distribution of effect sizes. We adopt the same sample data and testing procedure.

We use the lists of identity terms for racial and intersectional bias given in Guo and Caliskan (2021), namely related to AA versus EA identities and MA women versus EA men, along with the SCM attribute terms, to establish the presence of stereotypes in the contextualised word embeddings using the CEAT.

In addition to using the SCM terms, we will also use the pleasant/unpleasant terms from Guo and Caliskan (2021)'s paper - this provides a comparison point for use of the SCM versus another set of non-demographic-specific terms.

We also measure how strongly the demographic specific stereotype terms for MA women and EA men are associated with the demographic groups, to see if demographic specific stereotype associations are reduced following demographic agnostic debiasing. Note that we removed the word "intelligent" from the EA men attributes list as this also occurs in the competence attributes list and we wanted to be totally confident that any observed reduction in bias was due to restructuring of the entire embedding space and not due to bias being removed from an overlapping word. The CEAT does not have equivalent demographic specific terms for the AA/EA groups, though for completeness we compare how strongly the MA female/EA male specific terms are associated with the AA/EA groups.

Again, we adopt the approach of always assuming the more marginalised group will be represented in the model as more cold and incompetent compared to the majority group. This is an oversimplification. For example, Cuddy et al. (2008) indicate that in a Western context neither men nor women are strongly associated with coldness. However, we adopt this simplifying assumption to maintain testing consistency and thus require less human intervention, as per our goals.

We apply the CEAT before and after debiasing, to measure the success of the fine-tuning approach using the SCM terms.

3.3 Language Modelling Benchmark

Meade et al. (2022) note that apparent reductions in bias can reflect a worsening of language modelling performance. To ensure our debiasing procedure does not come at the expense of model performance, we evaluate our model on the GLUE benchmark (Wang et al., 2018).

The GLUE benchmark consists of 10 primary tasks and one diagnostic test, which evaluate the performance of a model in different contexts. We chose to evaluate our models using only five of these tasks – MRPC, SST-2, STSB, RTE and WNLI – following Kaneko and Bollegala (2021). These five tasks have small datasets, meaning we can minimise the effect of task-specific fine-tuning when running predictions (Kaneko and Bollegala, 2021).

We run the tests using the public GLUE code from huggingface[3]. We will perform these tests before and after debiasing, and compare the results. We report results based on the provided evaluation data.

[2] https://files.pushshift.io/reddit/comments/

[3] https://github.com/huggingface/transformers/tree/main/examples/pytorch/text-classification/

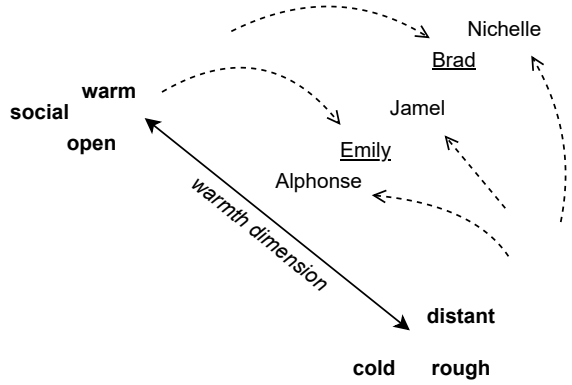

Figure 1: Diagram of intended orthogonal projection of target terms away from the warmth dimension, determined by attribute terms in **bold**. EA names underlined

.

4 Methodology

We use the 'bert-base-cased' model from the Hugging Face library[4]), henceforth BERT, although this same procedure should be applicable to any LLM with minimal modification.

We fine-tune the model following an adapted version of the procedure in Kaneko and Bollegala (2021). Namely, through a training objective that looks to minimise unwanted projection into the attribute category dimensions for the target words through an orthogonal projection, whilst also staying close to the contextualised embeddings of the pre-trained model to preserve semantics. We visualise this orthogonal projection in Figure 1. Adjusting the embeddings of the target terms to lie orthogonal to the warmth dimension (equidistant from the attribute terms) should ensure less negatively biased representations for minority groups (in the visualisation, AA names).

Crucially, we modified the original algorithm in Kaneko and Bollegala (2021) as we wish to remove unwanted projections into two dimensions, not just one: warmth/coldness and competence/incompetence. The first component of the loss function for layer i of our model is:

$$L_i = \sum_{d \in D} \sum_{t \in V_t} \sum_{x \in \Omega(t)} \sum_{a \in V_a} (v_i(a)^\top E_i(t; x; \theta_e))^2$$

where $E_i(t; x; \theta_e)$ represents the embedding of target word t in sentence x for model E_i, $v_i(a)$ is the average embedding for the attribute term across training sentences, and we calculate the inner product across all attributes $a \in V_a$, for all sentences containing the target $x \in \Omega(t)$, for all target words V_t, for all target dimensions, D_d.

The second component of the loss function is:

$$L_{reg} = \sum_{x \in A} \sum_{w \in x} \sum_{i=1}^{N} ||E_i(w; x; \theta_e) - E_i(w; x; \theta_{pre})||^2$$

where $E_i(w; x; \theta_{pre})$ is the contextualised embedding of a word, w, in a sentence, for the model before fine-tuning, and we calculate the squared ℓ_2 between this and the embedding after fine-tuning, for all layers, for all sentences and targets.

The final loss function is a weighted sum:

$$L = \alpha L_i + \beta L_{reg}$$

where α and β sum to 1.

Kaneko and Bollegala (2021) find debiasing all layers to be the most effective, so we do likewise.

5 Results

5.1 Baseline Performance

5.1.1 CEAT

Results for the CEAT for BERT are given in Table 1. We found there was a medium combined effect size (CES, between 0.5 and 0.8, as per the original paper's classification (Guo and Caliskan, 2021)) in the strength of association between EA names & warmth and AA names & coldness. We also found a medium strength association between EA names & competence and AA names & incompetence. As with the original paper, we found a small association between EA names & pleasantness and AA names & unpleasantness, suggesting this approach may be less able to detect the true scale of bias.

We also found a medium effect size association between AA names and the negative, MA women specific intersectional bias terms, and between the EA names and the EA male specific intersectional bias terms. This may be because the EA male stereotypes are relevant to all EA people.

For the intersectional power dynamic, we found a small association between EA male names & warmth and MA female names & coldness. We found a medium association between EA male names & competence and MA female names & incompetence. We found a very small association between EA male names & pleasantness and MA female names & unpleasantness, suggesting these

[4] https://huggingface.co/bert-base-cased

generic terms are less effective for detecting the true levels of bias in the model.

Finally, we found a medium effect size association between MA female names and the MA female specific bias terms, and between the EA male names and EA male specific bias terms - surprisingly, this association was weaker than than for the black/White demographic group, despite the fact that these stereotypes were chosen to be highly pertinent to the intersectional group.

5.1.2 GLUE

Table 2 shows the GLUE benchmark scores for BERT and DEBIAS, on the five chosen tasks.

The baseline BERT model performs very well on SST-2, MRPC and STS-B, with metric scores of around 90%. The lower scores come from the RTE and WNLI tasks. RTE assesses the model's ability to determine whether sentence A entails sentence B. WNLI assesses the model's ability to determine whether an inserted noun is correct. These specific grammatical situations seem to be the weaknesses of the model. The low score for WNLI is surprising and may indicate suboptimal hyperparameter choices during training. The training loss is comparable to that of a similar model on huggingface[5].

5.2 Debiasing Procedure

We adopt the values for α and β given in the original paper, namely 0.2 and 0.8 respectively, having trialed α 0.1 above and below and found 0.2 to be the best performing. Bar batch size and learning rate, all other hyperparameters were set to their default values for BERT. We trialed a number of starting learning rates and found the best to be 5e-5 (this is the same learning rate used in the original paper). Batch size was set to 32, as in the original paper. We train for 3 epochs (this is given in the code for the context debias paper but not specified).

We fine-tuned the model using the methodology detailed in Section 4.

5.3 Post-debiasing Performance

5.3.1 CEAT

The results of our post-debiasing CEAT tests indicate this debiasing procedure to be largely successful. We were able to reduce bias in DEBIAS and in all instances render the strength of stereotyped association to be very small.

[5] https://huggingface.co/gchhablani/bert-base-cased-finetuned-wnli

For DEBIAS, there is no longer an association between EA names & warmth and AA names & coldness, nor between EA names & competence and AA names & incompetence. Although our debiasing procedure involved only the SCM terms, it also had an impact on the other associations. The strength of association between EA names & pleasantness and AA names & unpleasantness has reduced to be very small. Intersectional bias was also reduced as to be very small. Though these very small effects are statistically significant, their practical impact will be negligible.

Similarly, we found that for DEBIAS, there is no longer an association between EA male names & warmth and MA female names & coldness, nor between EA male names & competence and MA female names & incompetence. The association with pleasantness was also reduced, although this effect size was very small to begin with. Intersectional bias was also reduced as to be very small.

5.3.2 GLUE

Table 2 shows the differences between GLUE benchmark scores for our model before and after debiasing. For most tests, the GLUE benchmark scores have very minor differences.

Our debiased model outperforms the baseline model on both the RTE and WNLI tasks, with the largest difference coming from WNLI. We suspect that the improvement regarding RTE is because the RTE dataset is constructed based on news and Wikipedia text (Wang et al., 2018), which are domains likely to contain significant bias. For WNLI, the task of resolving ambiguities requires real world knowledge, which is also highly influenced by bias. Removing bias from these datasets allows the model to focus on classifying entailment (RTE) or resolving ambiguities (WNLI) in a more reliable manner, without being "distracted" by stereotyped associations between particular groups and actions that are irrelevant to the task.

In general, these results show that debiasing the model did not hurt its performance, as would have been implied by Meade et al. (2022). On our five chosen GLUE tasks, any performance decreases were very minor, while the performance increases on RTE and WNLI were rather significant. Though not directly comparable to Kaneko and Bollegala (2021), as their paper considers 'bert-base-uncased', our results are inline with their findings showing debiasing along two "axes" does not unduly harm language modeling performance com-

	BERT		DEBIAS	
Test	CES	Sig.	CES	Sig.
EA,AA,Warm	0.77	*	**-0.12**	-
EA,AA,Comp.	0.67	*	**-0.18**	-
EA,AA,Pleas.	0.47	*	**0.16**	*
EA,AA,Inter.†	0.71	*	**0.15**	*
EAM,MAF,Warm	0.43	*	**-0.03**	-
EAM,MAF,Comp.	0.51	*	**-0.04**	-
EAM,MAF,Pleas.	0.17	*	**0.13**	*
EAM,MAF,Inter.	0.50	*	**0.08**	*

Table 1: Strength of combined effect size (CES) between attributes and targets for BERT before (BASELINE) and after (DEBIASED) debiasing. Sig. = significance. * = significant to $p < 0.05$. AA = African American names. EA = European American names. MAF = Mexican American female names. EAM = European American male names. Warm = warm/cold terms. Comp. = competent/incompetent terms. Pleas. = pleasant/unpleasant terms. Inter = Intersectional stereotypes.† **Bold** indicates that the debiasing procedure has reduced the absolute effect size to very small. †The intersectional stereotypes were intended as relevant to the EAM and MAF pair.

Benchmark	Baseline Score	Debiased Score
SST-2	**92.7**	92.5
MRPC	**89.5/85.0**	87.9/82.8
STS-B	**88.9/88.6**	88.7/88.5
RTE	66.1	**67.5**
WNLI	32.4	**42.3**

Table 2: GLUE Benchmark scores for both our baseline BERT, and our final DEBIAS models. Values correspond to the metrics described in Section 3.3. **Bold** indicates the best performance.

pared to debiasing along one axis.

6 Discussion

We found that our approach to bias measurement, informed by the SCM, proved to be an effective method for detecting bias in an LLM. We found that compared to using another list of generic, non-demographic specific attribute terms related to pleasantness, our approach seemed to give a more accurate measure of the level of bias in the model - our terms allow us to capture a stronger association between a minority group and negative stereotypes. It is possible that our approach exaggerates the level of bias in the model and in fact is less accurate. However, the effect sizes from our approach are closer to the effect size for association with demographic specific terms for the intersectional pair, suggesting it paints an accurate picture of negative bias in the model. Further, given how often BERT has been found to produce offensive content, it seems more likely that use of pleasantness terms is underestimating the level of bias in the model, rather than our approach overestimating it. The pleasantness terms were only slightly associated with EA male names compared to MA female names, yet BERT has been shown to consistently produce more favourable content about such individuals (Sheng et al., 2019).

Our finding that the intersectional bias terms were actually more strongly associated with the Black/white demographic groups highlights how the selection of demographic specific stereotypes for use in measuring bias and debiasing models can be challenging. That these stereotypes are actually more strongly associated with AA/EA names could suggest that the stereotyping captured by the model does not reflect the attitudes of the group of undergraduates responsible for generating these stereotypes (Ghavami and Peplau, 2013). It could also be that the model has not been exposed to sufficient (stereotyped) data to capture the category of MA females and the associated stereotypes.

The results might suggest that these demographic specific terms are actually rather "demographic agnostic", hence they are able to capture bias against AA people. However, intuitively, "sexy" and "feisty" (two MA female specific stereotypes) are not associated with people experiencing homelessness (and studies on public attitudes towards homelessness to our knowledge confirm this intuition), but the Stereotype Content Model is able to predict the contempt they experience due to being perceived as cold and incompetent (Cuddy et al., 2008), which is likely reflected in language

use and thus in an LLM.

After debiasing using the SCM informed approach, we were able to reduce bias in all instances. Not only did we reduce the association between competence, warmth and ethnicity, but we also reduced the association with pleasantness. Intuitively, this is likely a reflection of the semantic association between warmth and pleasantness - reducing projection in the warmth dimension may have impacted projection in the pleasantness dimension.

Crucially, we were able to reduce the association between the intersectional groups and their specific stereotypes, using a demographic agnostic approach that did not require prior knowledge of group specific stereotypes. Although we only ran the debiasing procedure for warmth and competence dimensions, there was a positive "knock on" effect, supporting our belief that debiasing at the more abstract level will reduce more specific bias associations as well, as these can be thought of as subcategories of these more generic stereotype concepts. We were able to successfully debias the model without impeding performance on benchmark NLI tasks, suggesting language modelling abilities have not been negatively impacted, and in two instances performance was actually improved, possibly due to the reduction in bias.

7 Conclusions

7.1 Future Work and Limitations

In future work we hope to make use of language models to generate the target identity terms, akin to Schick and Schütze (2021)'s use of LLMs to generate training data, using prompts such as "I am proud to identify as". This will further reduce the amount of human resource and *a priori* knowledge needed, making the approach more efficient and widely applicable. We may also try to introduce additional dimensions related to "universal" patterns of discrimination such as the use of dehumanising language (Cameron et al., 2016) and animal comparisons (Haslam et al., 2011).

Though we are hopeful that our proposed debiasing pipeline will show promising results, we acknowledge there are several inherent limitations we would look to address in future work.

First, the SCM has received significant support as a model for our perceptions of different groups, and its simplicity makes it ideal for use in our "demographic agonostic" approach. However, it has been shown that the model may fail to adequately capture stereotypes surrounding immigrant groups (Savaş et al., 2021). This might be addressed in future work by adopting additional attribute dimensions (i.e. diligence) to encompass a wider range of potential stereotypes. This will allow us to better measure and mitigate bias against groups which is not best captured by the warmth and competence stereotypes.

A second limitation is our use of Reddit data for both debiasing and testing for bias - it is not clear how robust the reduction in bias would be if tested using out-of-domain data.

A further limitation is that during the process of identifying suitable names from Tzioumis (2018) for our debiasing procedure, we found that some of the names used in CEAT tests to measure bias against Black Americans were not predominantly used by Black individuals (for example "Leroy"), an indication that relying on names to establish bias against a demographic group may be fallible.

Our use of the GLUE metric to evaluate language modelling performance is potentially problematic as this static benchmark is outdated and saturated for some tasks. Though using the same metric as Kaneko and Bollegala (2021) gave us confidence that debiasing along two axes did not unduly harm performance, we could better evaluate our model using modern dynamic benchmarks.

Finally, intrinsic measure of bias do not always correlate well with application bias (Goldfarb-Tarrant et al., 2021; Cao et al., 2022a), suggesting the CEAT may not accurately capture the extent of bias the model might be responsible for in downstream applications. In future work, we could evaluate the success of our debiasing approach using gender targets and an extrinsic measures such as Zhao et al. (2018), a gender bias in coreference resolution benchmark that could assess our model after finetuning for this task. We could also try to adapt the principles of this process to work in downstream tasks, for example amending the finetuning data to contain balanced stereotyped instances.

7.2 Conclusion

Our debiasing procedure has reduced stereotyped associations between minority groups and negative characteristics without the need for idiosyncratic target terms for each group, making it demographic agnostic and human resource efficient, in line with our goals. The debiasing procedure is able to effectively "neutralise" the presence of target dimen-

sions in the attribute embeddings, as well as decrease the association between more demographic specific stereotype attributes and the target demographics. The debiasing procedure did not come at the cost of performance, and even improved performance on RTE and WNLI.

Further, the finetuning procedure ran in a matter of hours on a single GPU, making it computationally efficient as well. This aligns with our goals, to establish a robust bias mitigation procedure that is efficient and widely applicable.

Our work can be thought of as a prototype for a promising debiasing procedure grounded in the SCM. In future, we hope to encompass automatic target term generation. We also plan to expand this work to more minority identities, and more importantly test the resulting model using a range of extrinsic bias measures and language modeling benchmarks, to evaluate the potential for a positive real world impact. The hope is that those using LLMs may apply our simple and efficient debiasing procedure before fine-tuning for their own purposes, helping to reduce the impact of stereotypes across the field.

8 Acknowledgements

We would like to thank our anonymous reviewers for their feedback. This work is in part supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics.

References

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 *ACL-IJCNLP 2021*.

C. Daryl Cameron, Lasana T. Harris, and B. Keith Payne. 2016. The emotional cost of humanity: Anticipated exhaustion motivates dehumanization of stigmatized targets. *Social Psychological and Personality Science*, 7(2):105–112.

Yang 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)*, page 561–570, Dublin, Ireland. Association for Computational Linguistics.

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

Amy J. C. Cuddy, Susan T. Fiske, Virginia S. Y. Kwan, Peter Glick, Stéphanie Demoulin, Jacques-Philippe Leyens, Michael Harris Bond, Jean-Claude Croizet, Naomi Ellemers, Ed Sleebos, Tin Tin Htun, Hyun-Jeong Kim, Greg Maio, Judi Perry, Kristina Petkova, Valery Todorov, Rosa Rodríguez-Bailón, Elena Morales, Miguel Moya, Marisol Palacios, Vanessa Smith, Rolando Perez, Jorge Vala, and Rene Ziegler. 2009. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, 48(1):1–33.

Amy J.C. Cuddy, Susan T. Fiske, and Peter Glick. 2008. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. volume 40 of *Advances in Experimental Social Psychology*, pages 61–149. Academic Press.

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.

Lewis Diuguid and Adrienne Rivers. 2000. The media and the black response. *The ANNALS of the American Academy of Political and Social Science*, 569(1):120–134.

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)*, page 600–616. Association for Computational Linguistics.

Negin Ghavami and Letitia Anne Peplau. 2013. An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses. *Psychology of Women Quarterly*, 37(1):113–127.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. 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), page 1926–1940. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, page 122–133. ACM.

Nick Haslam, Steve Loughnan, and Pamela Sun. 2011. Beastly: What makes animal metaphors offensive? *Journal of Language and Social Psychology*, 30(3):311–325.

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*, page 1256–1266. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

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.

Gandalf Nicolas, Xuechunzi Bai, and Susan T. Fiske. 2021. Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology*, 51(1):178–196.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Özge Savaş, Ronni M. Greenwood, Benjamin T. Blankenship, Abigail J. Stewart, and Kay Deaux. 2021. All immigrants are not alike: Intersectionality matters in views of immigrant groups. *Journal of Social and Political Psychology*, 9(1):86–104.

Timo Schick and Hinrich Schütze. 2021. Generating datasets with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. 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.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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)*, page 3405–3410, Hong Kong, China. Association for Computational Linguistics.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael Wick. 2022. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Konstantinos Tzioumis. 2018. Demographic aspects of first names. *Scientific Data*, 5(11):180025.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing NLU models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610, 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 H. Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. Technical report.

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.