Mitigating Social Bias in Language Modelling and Generation

Friederike Blatt, Xenia Heilmann
Technical University of Darmstadt
Darmstadt, Germany
friederike.blatt@stud.tu-darmstadt.de, xenia.heilmann@stud.tu-darmstadt.de

Abstract

Natural language models can reproduce social biases of the training data. We review two bias mitigation models described by Garimella et al. (2021), which successfully reduce gender and racial bias. We discuss their shortcomings and compare them to previous research. While Garimella et al. (2021) improve on previous models' performance, comparison is often difficult due to incongruent approaches and metrics. We finally argue that the unclear definition of social bias reflects an unresolved conceptual issue about the goal of bias mitigation.

1 Introduction

Reproduction and amplification of social biases by natural language processing (NLP) models can increase social inequality. NLP models adapt social biases when trained on biased data. In recent years, various approaches to mitigate social biases in NLP models have been presented.

We review the approaches presented in *He is very intelligent, she is very beautiful? On Mitigating Social Bias in Language Modelling and Generation* by Garimella et al. (2021). The goal of Garimella et al. (2021) is to reduce **representation bias** in NLP models. **Representation bias** is a form of bias where certain social groups (men, women) are associated with what Garimella et al. (2021) refer to as certain identities, meaning association of a soical group with a set of stereotypical attributes, such as man $\rightarrow$ *intelligent*, woman $\rightarrow$ *beautiful*. The considered NLP model is BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), a language representation model. In sentence completion tasks, BERT generates sentences like "He is very intelligent. She is very beautiful.". In these examples, words like "intelligent" are more likely to be associated with the male gender, whereas "beautiful" is picked more often for the female gender. To create a NLP model with reduced biases, Garimella et al. (2021) introduce additional loss functions for bias penalization.

Garimella et al. (2021) consider two target demographics, gender and race. For each target demographic, two groups are considered: male/female for gender and African American/Caucasian for race. The goal is to reduce bias between the two groups of each target demopgraphic.

For mitigation of social biases, Garimella et al. (2021) introduce modified models of BERT. This is possible as BERT is based on a multi-layer bidirectional transformer encoder (Devlin et al., 2018) which enables fine-tuning by adding one additional output layer. Initially, BERT is pretrained on unlabeled data. Figure 1 displays how the models build on one another. The two main approaches by Garimella et al. (2021) are:

1. DEBIASBERT: This model further pretrains BERT with additional loss functions for bias reduction. The **equalizing loss** aims to equalize associations between words indicating social groups and neutral words. In a follow-up pretraining step, the **declustering loss** aims to decluster word clusters based on social biases.

2. DEBIASGEN: This model focuses on mitigating bias during the task of summarization. To achieve this, DEBIASBERT is used as the encoder, and, on top of this, a decoder with an additional **bias penalizing loss** is fine-tuned.

Bias mitigation with DEBIASBERT and DEBIASGEN is performed for both target demographics, race and gender. Three datasets are used for fine-tuning different versions of BERT and DEBIASBERT. The output of these datasets show reduced SEAT scores (May et al., 2019) for DEBIASBERT compared to BERT and a bias mitigation of over

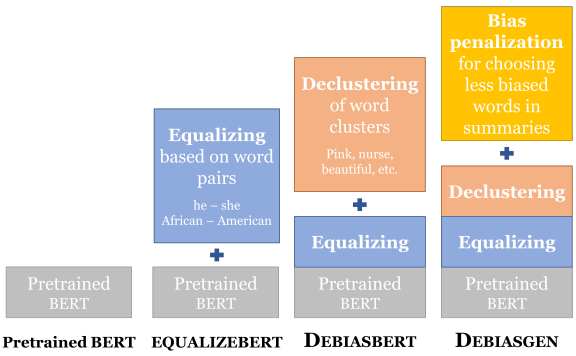

Figure 1: Overview of the different parts of the models presented in this paper.

30% according to human evaluation. For evaluation of DEBIASGEN, the CCO score (Garimella et al., 2021) is introduced. For this score, a reduced bias is found when comparing outputs of DEBIASGEN with BERT + decoder. Human evaluation also shows bias mitigation by DEBIASGEN.

2 DEBIASBERT

DEBIASBERT is a further pretrained version of BERT with mitigated social biases. Pretraining consists of two stages, **equalizing** and **declustering**. Equalizing attempts to associate neutral words to the same degree with male-defined and female-defined words (for gender) or with African American-defined and Caucasian-defined words (for race). Declustering addresses the problem that neutral words form clusters based on stereotypes and that these clusters are close to certain gender-defined or race-defined words. The goal is to increase the distance between the clustered words as well as their distance to the gender-specific or race-specific word.

Equalizing Two word pools are created, one for gender and one for race. They consist of words that define a specific gender or race. Opposing words are combined into word pairs like (he, she) or (African, American). For gender 65 word pairs are defined and 6 for race. Words not included in these groups are considered neutral. The word pairs are used to calculate the equalization loss

$$\text{EqLoss} = \lambda \frac{1}{k} \sum_{i=1}^{k} \left| \log \left(\frac{P([\text{groupA}_i])}{P([\text{groupB}_i])} \right) \right|. \quad (1)$$

Here, $P[\text{groupA}_i]$ and $P[\text{groupB}_i]$ represent the probabilities for the words of a word pair (for example he/she) to co-occur with a given neutral word. The number of these word pairs is given by k and $\lambda \geq 0$ is the equalization weight.

Declustering Words that have similar meanings or are used in the same contexts form clusters in embedding space. Hence, word clusters can also form based on social biases.

For DEBIASBERT, word representations from BERT are extracted with Brown Corpus (Francis and Kucera, 1967) as external signal. These representations are projected on axes indicating gender or race. Words with extreme positions on these axes are identified as "socially marked" (Garimella et al., 2021). These socially marked words are used in the loss function for declustering

$$\text{DeclustLoss} = \lambda \left| \log \left(\frac{\sum_{i=1}^{|A|} P([\text{social-groupA}_i])}{\sum_{i=1}^{|B|} P([\text{social-groupB}_i])} \right) \right|. \quad (2)$$

Here $|A|$ and $|B|$ refer to the number of socially marked words in each group. Garimella et al. (2021) do not explain the meaning of $P([\text{social-groupA}_i])$ and $P([\text{social-groupB}_i])$. We assume that they refer to the relative frequencies of socially marked words in the two word groups. With loss according to Eq. (2), the socially marked neighbors of each word should become more equally distributed among groups A and B.

2.1 Experimental Setting

Three datasets are used for pretraining and fine-tuning. The CNN/DM dataset (Hermann et al., 2015) consists of news articles from CNN and Daily Mail, WikiText-103 (Merity et al., 2016) is a collection of Wikipedia articles and Brown Corpus is a collection of texts about various topics. For each of these datasets one version of BERT is pretrained. This pretrained version (PT-BERT) is then further pretrained, first with equalization losses (EQUALIZEBERT) and then with equalization losses and declustering losses (DEBIASBERT).

Target Concepts	Attributes
European American names This is Katie.	*Pleasant* There is love.
African American names Jamel is here.	*Unpleasant* This is evil.

Table 1: Examples of subsets of target concepts and attributes adapted from May et al. (2019). Similar target concepts and attributes from Liang et al. (2020) are used for calculation of SEAT scores for DEBIASBERT.

All pretraining steps are done twice for each dataset, once with the goal to reduce gender bias and once to reduce racial bias. The weights λ from Eq. (1) and Eq. (2) are chosen such that SEAT scores are minimized.

2.2 Evaluation Methods

Bias mitigation is measured with the SEAT score. SEAT is based on WEAT (Anthony G. Greenwald and R., 2009) which measures the association between two sets of target concepts and two sets of attributes. SEAT compares sets of sentences. Each word is built into bleached sentence templates as shown in Table 1. Attributes used in the context of Garimella et al. (2021) are for instance *male* and *female*, target concepts are for example *family* and *career*. SEAT scores range from 0 to ∞. Lower SEAT scores indicate lower bias.

2.3 Experimental Results

Results in the form of SEAT scores are presented in Table 2. The first row displays SEAT scores for unmodified BERT trained on the Wikipedia and Book Corpus (Zhu et al., 2015). SEAT scores for different versions of BERT are displayed in the following rows. Values in brackets are the chosen λ values.

For gender, DEBIASBERT achieves low scores when trained on the CNN/DM dataset (0.1) or Brown Corpus (0.172). For EQUALIZEBERT the best score is measured when trained on WikiText-103 (0.173). Each of these best values (0.1, 0.172, 0.173) outperforms the score of 0.256 from SENTDEBIAS (Liang et al., 2020). The best score for bias reduction concerning race is achieved by EQUALIZEBERT trained on WikiText-103 (0.132) and Brown Corpus (0.222). The SEAT scores increase for race when EQUALIZEBERT and DEBIASBERT are trained on the CNN/DM dataset compared to BERT. SEAT scores were also calculated to test whether models trained to reduce gender bias also show less racial bias and vice versa. The SEAT score for gender bias is 0.26 for DEBIASBERT trained on the CNN/DM dataset to reduce racial bias. When DEBIASBERT is trained on WikiText-103 for gender bias mitigation, the SEAT score is 0.2 for racial bias. These results show that reduction of one bias type also tends to reduce the other.

Human Evaluation. Results from sentence completion are compared between BERT and DEBIASBERT. Sentences are labeled by 131 workers for gender and 140 workers for race. Labeling means workers have to decide whether a sentence is biased towards a specific group or unbiased. For gender, the proportion of biased sentences decreases from 38% for BERT to 4% for DEBIASBERT. For race, bias decreases from 48% for BERT to 6% for DEBIASBERT.

	MODEL	GENDER	RACE
	BERT	0.355	0.236
CNN/ Daily Mail	PT-BERT	0.352	0.490
	EQUALIZEBERT	0.135 (1)	0.368 (0.25)
	DEBIASBERT	**0.100** (1)	0.314 (1)
Wiki Text 103	PT-BERT	0.473	0.206
	EQUALIZEBERT	**0.173** (0.75)	**0.132** (0.5)
	DEBIASBERT	0.422 (1)	0.284 (1)
Brown Corpus	PT-BERT	0.373	0.396
	EQUALIZEBERT	0.255 (1.25)	**0.222** (0.75)
	DEBIASBERT	**0.172** (1)	0.274 (1)
	Liang et al. (2020)	0.256	–

Table 2: SEAT scores for comparison of different pretrained versions of BERT with EQUALIZEBERT and DEBIASBERT. Values in brackets are the λ values that led to the best performances for evaluation and declustering. The result from Liang et al. (2020) is included for comparison. Adapted from Garimella et al. (2021).

3 DEBIASGEN

The second approach which Garimella et al. (2021) discuss is DEBIASGEN. DEBIASGEN is a NLP model trained for summarizing input articles with the accessory goal of reducing social biases (race and gender). Concretely, Garimella et al. (2021) aim to avoid language which is seen as a generalization towards a specific group.

DEBIASGEN is set up with DEBIASBERT as the encoder and an additional fine-tuned transformer-based decoder. For the decoder, Garimella et al. (2021) use the framework from Liu and Lapata (2019). Further, a bias penalizing loss to mitigate input-specific biases and a negative log-likelihood loss are included. The benefit of this additional bias penalization term is that the decoder is encouraged to choose words and/or sentences that are less biased but still include the most important information of the input articles. The bias penalizing term

is defined as:

$$\text{BiasPenalizingLoss} = \sum_{i=1}^{|W|} \left(e^{b_i} \times P(W_i) \right). \quad (3)$$

Here, W is the set of all adjectives and adverbs in the vocabulary, b_i the bias score of word W_i and $P(W_i)$ the probability of W_i. The bias score b_i for a specific word W_i is calculated by

$$b_i = \frac{1}{k} \sum_{j=1}^{k} \left| \log \left(\frac{P(\text{groupA}_j, W_i)}{P(\text{groupB}_j, W_i)} \right) \right|. \quad (4)$$

In this equation k defines the number of gender/race-defined words, groupA and groupB contain definition words for female and male (for gender bias) or African American and Caucasian (for race bias). $P(\text{groupA}_j, W_i)$ is the probability of the j^{th} gender/race-defined word occurring together with W_i in the input articles.

Bias scores calculated for race with Eq. (4) are high compared to the bias scores calculated for gender. Therefore, Garimella et al. (2021) use $(1 + b_i)$ instead of e^{b_i} in Eq. (3).

3.1 Experimental Setting

Garimella et al. (2021) provide experimental results for three different settings: BERT as encoder with regular decoder (S1), DEBIASBERT as encoder with regular decoder (S2) and DEBIASGEN (DEBIASBERT as encoder + bias penalizing loss decoder; S3). As decoder, the 6-layered transformer decoder from Liu and Lapata (2019) without initial pretraining is used. During training the default parameters defined by Liu and Lapata (2019) are used. As datasets, CNN/DM (see Section 2.1) and XSum (BBC articles and accompanying single sentence summaries; (Narayan et al., 2018)) are used.

3.2 Evaluation Methods

To evaluate the quality of the output summaries, Garimella et al. (2021) use ROUGE (Lin, 2004), perplexity (from BERT), SLOR (Kann et al., 2018), and introduce the **Constrained Co-Occurrence (CCO)** score which is based on the Co-Occurrence bias score introduced by Qian et al. (2019).

ROUGE determines the quality of a summary by comparing it to (ideal) reference summaries created by humans. In Garimella et al. (2021), three individual ROUGE scores are of importance. ROUGE-1 represents the overlap of unigrams between the summary and its reference summary. ROUGE-2 calculates the overlap of bigrams between the given summary and reference summary. Lastly, ROUGE-L measures the longest common subsequence of words between the given summary and reference summary.

Perplexity shows how probable, meaningful and grammatically well-formed a sequence of words or a sentence is. A sentence is more fluent if its perplexity score is low.

SLOR is a normalized language model score that evaluates the fluency of a generation task output at sentence level. Higher scores indicate a more fluent text.

CCO estimates the bias in a given text by comparing co-occurrences of neutral words with definition words of given groups. It is defined as

$$\text{CCO(text)} = \frac{1}{N} \sum_{w \in N} \left| \log \left(\frac{\sum_{a \in A} c(w, a)}{\sum_{b \in B} c(w, b)} \right) \right|, \quad (5)$$

where N is the set of adjectives and adverbs in the text[1], A and B the gender/race-defined words, $c(w, a)$ and $c(w, b)$ the number of co-occurrences of word w with each gender/race defined word in its context. The CCO score can take values in $\{0, \infty\}$, higher values indicate more bias.

3.3 Experimental Results

Garimella et al. (2021) show that regarding the quality of content (ROUGE) and linguistic fluency (perplexity and SLOR), settings S1, S2 and S3 have approximately the same results. In contrast, the CCO scores are highest for S1. They decrease for S2 and are cut by half for S3. In Table 3, results for the CNN/DailyMail dataset are provided.

From their results, Garimella et al. (2021) conclude that DEBIASGEN can generate summaries with reduced bias while maintaining the quality and fluency of standard summaries. Further, Garimella et al. (2021) note that, if an input article is already highly biased, DEBIASGEN's generated summary still contains some of the input's bias.

Human Evaluation. Garimella et al. (2021) provide a human based survey on the resulting summaries for racial bias. Twenty-one summaries obtained by the models S1 and S3 are rated. For each summary, the 82 workers label the extent of bias towards either the African American group or the Caucasian group. Six out of the 21 summaries from the S1 setting are labeled as more biased against the

[1] We think that there are missing absolute value bars in the first fraction, since it is not possible to divide through a set.

CNN/DailyMail						
	R1	**R2**	**RL**	**CCO**	**PPL.**	**SLOR**
GENDER						
S1	40.74	18.66	37.90	1.902	1.938	19.921
S2	40.15	18.13	37.18	1.833	1.894	19.951
S3	40.03	18.07	37.18	**0.991***	1.908	19.897
RACE						
S1	40.74	18.66	37.90	0.068	1.938	19.921
S2	40.29	18.31	37.40	0.065	1.905	19.943
S3	40.32	18.27	37.51	**0.044***	1.913	19.894

Table 3: This table shows the ROUGE (R1, R2, RL), CCO, perplexity (PPL.) and SLOR scores for the summaries resulting from the three different model settings (see Section 3.1) on the dataset CNN/DM. Adapted from Garimella et al. (2021). $*p < 0$

African American group. Garimella et al. (2021) see this as an additional proof that summaries from DEBIASGEN mitigate bias.

4 Related Work

Mitigating social biases in natural language models receives growing interest from the scientific community. In recent years, many different methods have been proposed for mitigation of these biases. A focus has been gender bias.

An early approach by Bolukbasi et al. (2016) focuses on mitigating gender bias in word embeddings. By neutralizing and equalizing or softening gender subspaces in the embeddings, Bolukbasi et al. (2016) reduce gender bias. After application of this method gender stereotypes are reduced from an initial 19% to 6%. Although the percentage of initially biased sentences is lower than for Garimella et al. (2021), the percentage of biased sentences output is higher than for DEBIASBERT (6% versus 4%).

Qian et al. (2019) propose several methods which are similar to the methods applied by Garimella et al. (2021). Qian et al. (2019) propose to equalize gender bias by applying a novel loss function. The same loss function is applied to equalize word pairs in DEBIASBERT (see Eq. (1)). Also, Qian et al. (2019) introduce a co-occurrence bias score on which the definition of the Constrained Co-Occurrence applied by Garimella et al. (2021) (see Section 3.2) is based. The best co-occurence bias score from experiments by Qian et al. (2019) is 0.205 for the Daily Mail dataset (Hermann et al., 2015). Qian et al. (2019) only provide co-occurrence bias scores and do not work on the task of summarization. Therefore, it is not possible to compare the provided results with the results of Garimella et al. (2021).

Liang et al. (2020) develop a method called SENT-DEBIAS which adds an additional contextualization step. They define bias attributes which are similar to the word pairs used by Garimella et al. (2021). Liang et al. (2020) then insert these bias attributes into sentence templates that are as diverse as possible. On the basis of this, a debiasing step is performed. Reduced SEAT scores are achieved for most experimental settings by Liang et al. (2020). As mentioned in Section 2.3, SEAT scores achieved by DEBIASBERT outperform SENT-DEBIAS in all provided experiments.

Sheng et al. (2020) use adversarial triggers to influence social bias between demographic groups. This method proves to be efficient in mitigating bias as well as in inducing negative or positive bias towards one specific target group. Sheng et al. (2020) provide experimental results on gender, race and sexual orientation biases.

Other related work is described in a broad survey of social biases in language generation by Sheng et al. (2021).

5 Discussion

Garimella et al. (2021) present two NLP models for mitigation of social bias based on BERT. The first one, DEBIASBERT, shows bias reduction in human evaluation of 34% for gender and of 42% for race. The second one, DEBIASGEN, shows a reduced CCO score, which indicates reduced social bias.

The majority of research in the field of bias mitigation in NLP models has focused on gender bias. Garimella et al. (2021) conduct experiments on mitigation of racial bias as well. We see this as a step towards broadening the research of social bias in NLP models. Garimella et al. (2021) state that little research is done in the field of debiasing language generation tasks. By mitigating social bias in sentence completion and summarization tasks, Garimella et al. (2021) contribute to also debiasing language generation models.

With DEBIASGEN, a concrete language generation task applicable to daily life is introduced. This method could be a basis for transferring bias mit-

igation to a variety of language generation tasks. Further research is needed to identify concrete tasks to which DEBIASGEN can be extended. We believe that the research by Garimella et al. (2021) can help make social bias mitigation applicable to a larger set of NLP applications. For example, the concept of using word pairs as input to additional loss functions is extendable to other groups who suffer from social bias, such as homosexuals. Another direction of research could be to transfer DEBIASBERT and DEBIASGEN to other languages similar to English.

SEAT Score Evaluation of EQUALIZEBERT and DEBIASBERT is done using SEAT. While SEAT is well suited to detect the presence of bias, it cannot show its absence (May et al., 2019). In particular, a decreased score does not necessarily mean that bias mitigation was successful. Furthermore, SEAT has no fixed scale, which means that results of different studies are not comparable. In summary, further research on the behavior of SEAT or the development of a better metric is necessary. As long as the behavior of SEAT is not fully understood, it should be interpreted with great care.

For the dataset WikiText-103, SEAT scores increase for DEBIASBERT compared to EQUALIZEBERT and BERT. This happens for both, gender and race. Garimella et al. (2021) provide no explanation for this behavior. In this case, dataset-specific results from human evaluation might have been insightful. A direct comparison of trends indicated by results from both evaluation methods would help judge how well suited the SEAT score is for measuring bias mitigation.

CCO Score With DEBIASGEN, Garimella et al. (2021) tackle a new topic in the scope of bias mitigation: bias mitigated summaries. Yet, as there is no other research in this field, it is difficult to evaluate the results Garimella et al. (2021) provide. The greatest limitation is the CCO score. This score is introduced by Garimella et al. (2021) and used as the only score to evaluate the bias still contained in the summaries generated by DEBIASGEN. While this score shows that DEBIASGEN mitigates bias compared to BERT (see Table 3), the question is how well this score actually shows bias. Theoretically, the score could be well suited for evaluating bias in long generated texts and the paper by Qian et al. (2019) could be an indication for this. But as there are no other papers on bias mitigation for summarization tasks which measure their results with the CCO score, it is not clear that the CCO score is a good measure for bias in summaries. Further research is needed to qualify the CCO score as a valid measure for bias in summaries.

Representation of Demographics As mentioned in Section 2.3, the SEAT scores for race increase compared to BERT for the CNN/DM dataset. Garimella et al. (2021) provide a possible explanation: To indicate racial groups, SEAT uses templates around names that are more likely to appear in a certain racial group (like Hakim for African American). Gender on the other hand is indicated by group terms like (boy, girl) in SEAT. Names may not be sufficient to encapsulate the complexities of racial bias. An alternative explanation Garimella et al. (2021) provide is that the use of word pairs may be insufficient. While gender can be divided in two groups, race is a more complex construct and may not be captured by single words. Further research to investigate how race is indicated by language is needed.

Garimella et al. (2021) state another problem related to the above mentioned word pairs: All studies in the field discussed here depend on manually predefined word pairs. Either these word pairs are based on existing word pair research (e.g. (Zhao et al., 2017)), or word pairs for target demographics are identified by the researchers themselves. This is a great limitation when trying to transfer an existing approach to a new demographic group. We support the suggestion by Garimella et al. (2021) that more research is needed to propose approaches independent of word pairs. This could result in bias-mitigated NLP models applicable to all demographics and extend the current research, which mainly focuses on gender bias. We believe that automatic identification of words specific for the target demographic could help make models generalize better and thereby drive adoption in more use cases.

Human Evaluation Garimella et al. (2021) do not conduct extensive human evaluation of DEBIASGEN. Only 21 summaries are evaluated for race bias mitigation by humans. From these summaries, only 6 prove to have mitigated racial bias. While this small number aligns well with differences in the CCO score for mitigation of race bias (0.068 vs. 0.044), the strong decrease of the CCO score in the gender bias (1.902 vs. 0.991) is not backed by

human evaluation. We think that especially since Garimella et al. (2021) introduce a new score for evaluating bias mitigation in text generation tasks, more extensive human evaluation should have been conducted.

Regarding the results from human evaluation for DEBIASBERT and DEBIASGEN, although they appear clear-cut, we think some simple statistical tests could have been insightful, for example a χ^2-Test to validate statistical significance. Another question here is how representative the results from human evaluation are with respect to the sample of workers. Garimella et al. (2021) only report the number of workers and that all workers have a US background. Since no further demographics are provided, it is impossible to judge whether the human evaluation suffers itself from implicit social bias. Even people that are aware of their implicit social biases tend to reproduce them (Jackson et al., 2014). This is related to the question of who the intended recipient of social bias mitigation is. As we pointed out, the perception of bias differs between social groups. When judging gender bias, for instance, men may be more forgiving than women. The same holds for different racial groups. This question is not considered by Garimella et al. (2021) but is likely to have a strong effect on results. Without clarification on this question it is even more difficult to compare results from different approaches. Admittedly, this broad goal-setting issue plagues any approaches to bias mitigation.

Exceptions When trying to mitigate bias in NLP models, unwanted side effects may occur. As an example, Garimella et al. (2021) mention words that canonically belong to a certain gender or race. For instance, the word "dress" is usually associated with the female gender. Further work is needed to identify such words and exclude them from debiasing operations. An approach which tackles this problem is described by Bolukbasi et al. (2016). There, a softening function is applied which reduces differences in both the already debiased word space and the original word space, with the goal to reinforce certain distinctions (such as female $\rightarrow$ *dress*).

Racial Bias In Garimella et al. (2021) and similar research with the goal of reducing racial bias, single words are used as indicators for race. Such words are either directly describing the racial heritage (like "African American") or names that are typical for a certain racial background (like "Hakim"). Garimella et al. (2021) stipulate that single words or word pairs cannot capture the complex manner in which racial concepts are embedded in language. Yet, to our knowledge, insights from linguistics or social sciences are not incorporated in the development of bias mitigation models. We believe that an interdisciplinary study with linguists and computer scientists could shed light on the matter. If indeed word pairs turn out to not suffice as indicators for race, models as simple as the one by Garimella et al. (2021) may have to be extended.

Transfer-Learning In future work on the topic of mitigating social bias an interesting phenomenon to study is transfer-learning (Pan and Yang, 2010). Transfer-learning occurs when a model trained on one task shows above-baseline performance on a similar but different task. As shown in section 2.3, when DEBIASBERT is trained to reduce gender bias, racial bias is reduced as well and vice versa. The SEAT score for racial bias (0.2) for DEBIASBERT trained on WikiText-103 for reducing gender bias is even lower than the one for the model trained for reducing racial bias. We believe that this indicates a form of transfer-learning in DEBIASBERT. While training DEBIASBERT to mitigate racial bias alone does not work well, complementing it with transfer-learning from other types of biases could improve results. Using loss terms for gender and race simultaneously or in two separate pretraining steps could be two ways to achieve this.

6 Conclusion

In the paper *He is very intelligent, she is very beautiful? On Mitigating Social Bias in Language Modelling and Generation*, Garimella et al. (2021) present two new models for mitigation of social bias in language generation tasks. Results show a reduction of bias for both models while keeping the quality of generated summaries at high levels. We have argued that the lack of a universal score for bias makes comparisons between different approaches difficult. We further noted that even such a universal score would suffer from the ill-defined goal of bias reduction. We made suggestions to future research directions, like consulting linguists for a better understanding of how concepts like race are encapsulated in language, or using transfer-learning to improve performance.

References

Eric Luis Uhlmann Anthony G. Greenwald, T. Andrew Poehlman and Mahzarin R. 2009. Understanding and using the implicit association test: Iii. meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, 97(1):17–41.

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

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

Winthrop Nelson Francis and Henry Kucera. 1967. Computational analysis of present-day american english. *Providence, RI: Brown University Press. Kuperman, V., Estes, Z., Brysbaert, M., & Warriner, AB (2014). Emotion and language: Valence and arousal affect word recognition. Journal of Experimental Psychology: General*, 143:1065–1081.

Aparna Garimella, Akhash Amarnath, Kiran Kumar, Akash Pramod Yalla, N Anandhavelu, Niyati Chhaya, and Balaji Vasan Srinivasan. 2021. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4534–4545.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Sarah M Jackson, Amy L. Hillard, and Tamera R. Schneider. 2014. Using implicit bias training to improve attitudes toward women in stem. *Social Psychology of Education*, 17:419–438.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! *arXiv preprint arXiv:1809.08731*.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. *arXiv preprint arXiv:1905.12801*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2020. Towards controllable biases in language generation. *arXiv preprint arXiv:2005.00268*.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Yukun Zhu, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 19–27.