On The Effectiveness of Gender Debiasing Methods in Removing Gender Information From Model Representations

Anonymous ACL submission

Abstract

Large pre-trained models such as BERT have been shown to demonstrate biased behavior towards different demographic groups, such as gender, race, or religion. Despite the development and proposal of various debiasing methods, there is a paucity of prior research focusing on the efficacy of debiasing methods in removing the latent demographic information encoded in internal representations. We examine the effectiveness of some recent bias mitigation methods in removing stereotypical gender information from internal model representations using Minimum Description Length (MDL) probing. We discover that the effectiveness of current debiasing techniques might not necessarily be indicative of reduced latent gender bias in representations. Furthermore, we investigate the effect of debiasing methods on internal representations using layerwise probing, showing that they tend to concentrate gender information in a few layers. We additionally apply a number of state-of-the-art debiasing methods to the layers with the highest gender information concentration, finding that by focusing on these layers, there is only a minimal change in model behavior with respect to fairness and performance.

1 Introduction

Recent research indicates that pre-trained language models, such as BERT (Devlin et al., 2019), exhibit different societal stereotypes, including racism and sexism. Given the extensive implementation of these models and the numerous concerns it can cause, various methods have been proposed to mitigate bias in these models, either by manipulating datasets (Zhao et al., 2018a), refining the learning algorithm (Kaneko and Bollegala, 2021a), or by modifying the architecture of the network (Lauscher et al., 2021). Despite all these efforts, to our knowledge, no research has so far focused on the effectiveness of these methods in removing gender information from model representations. As a result, there is limited evidence demonstrating whether these debiasing strategies eliminate encoded gender-biased information.

In this paper, we carry out a set of experiments to determine if the existing debiasing techniques used to mitigate gender bias are also effective in reducing the captured bias information in model representations. We study three different debiasing techniques, from those that change the training dataset or the learning objective to those that directly alter model's architecture. We evaluate the amount of captured gendered information by BERT's representations using two probing datasets, BiosBias (De-Arteaga et al., 2019), and Funpedia (Dinan et al., 2020). We find that the significant performance improvements of debiasing techniques on bias datasets might not necessarily indicate that the gender information is discarded (or even reduced) from their representations. While some methods, such as counterfactual augmentation (Zhao et al., 2018a), tend to significantly reduce the encoded gender information in some cases, others either have negligible effect on BERT's internal representation or even amplify the gender information that they encode.

Furthermore, we apply MDL probing, an information-theoretic probing classifier proposed by Voita and Titov (2020) in a layerwise setting in order to determine the layers that encode the gendered information the most. We find that it is indeed the case that some layers encode more of the gendered information in comparison to other layers, with deeper layers consistently having higher gender information concentration in comparison to earlier layers. We apply MDL probing to the base, fine-tuned, and debiased models to determine the effects of debiasing on intermediate representations. We hypothesize that an effective debiasing method should have the largest effect on layers that encode the gendered information the most.

We finally apply counterfactual augmentation

(Zhao et al., 2018a) and adapter-based debiasing (Lauscher et al., 2021) only to the layers that encode the highest amount of gender information. We observe that by carefully selecting the layers that are to be debiased, we can reach a performance that is comparable to a full-model debiasing, in which every layer of a given model is debiased.

Our work is inspired by Mendelson and Belinkov (2021) who studied the impact of debiasing techniques used to reduce the model's reliance on spurious correlations between data and labels in natural language inference on model's representations. Our contribution is threefold:

- We utilize MDL probing to determine the encoded gendered information in pre-trained language models. We show that debiasing techniques do not necessarily reduce the encoded bias information in internal representations.
- We extend the probing to layer-wise analysis of pre-trained language models to determine the distribution of encoded information across layers. We find that some layers tend to encode this information more than the others. This observation can be used to develop efficient and effective debiasing techniques that focus on specific layers.
- To test our hypothesis, we apply two debiasing techniques only on layers with the highest gender information concentration, finding that it is indeed possible to develop models that are comparable to fully debiased models, while modifying only a small portion of model's weights.

2 Background

In this section, we discuss MDL probing, the technique we employ to measure gender information captured by model representations, as well as common measurement metrics used to quantify bias in neural networks.

2.1 MDL Probing

Traditionally, in order to extract the information encoded in a model's representations, a shallow classifier was trained using the model representations with the goal of predicting a linguistic feature (Belinkov, 2022). However, it has been shown that such models are unreliable, as they tend to classify representations of random data almost similarly to the representations of real data (Zhang and Bowman, 2018), highlighting the fact that these methods are inadequate to capture variations in representations, making their results hyperparameter-dependent.

To address this problem, Voita and Titov (2020) have proposed Minimum Description Length Probing, where in addition to the accuracy of the shallow classifier, this criteria measures how much effort does it need to extract that information from the model representations. Formally, they establish that a code exists to losslessly compress the labels using Shannon-Huffman code such that $L_p(y_{1,z}|x_{1,z}) = -\sum_{i=1}^{z} log_2 p(y_i|x_i)$. Note that this is the Cross-Entropy loss. Furthermore, they define the uniform code length as $L_{unif}(y_{i,z}|x_{i,z}) = z log_2(C)$ where C is the number of classes in our task.

Having calculated the uniform code length, they compare the Cross-Entropy loss against the uniform code length to find the final compression. Given a model $P_\theta(y|x)$ with learnable parameters θ, they choose blocks $1 = n_0 < n_1 < ... < n_s = N$ and encode data by these blocks. The model starts by transmitting the data using the uniform code length for the first chunk. The model is then trained to predict labels y from the data x, and also used to predict the labels. The next block is transmitted using this trained new model. This process continues until the entire dataset is covered. Final compression is calculated as follows:

$$L^{\text{online}}\left(y_{1:z} \mid x_{1:z}\right) = z_1 \log_2 C - \sum_{i=1}^{S-1} \log_2 p_{\theta_i}\left(y_{n_i+1:n_{i+1}} \mid x_{n_i+1:n_{i+1}}\right) \quad (1)$$

Note that this encourages the model to perform well with smaller blocks, as if the model performs well in compressing the data in the block n_i, the compression will be increased for the subsequent block n_{i+1}.

2.2 Bias Measurement Methods

Fairness metrics are measurement criteria which are used to observe a model's performance with respect to protected variables such as gender. Various methods have been proposed to measure gender bias in machine learning models. One of the common approaches to measuring gender bias is by looking at the statistical differences across multiple values of the protected variables. Statistical parity, for instance, states that a classifier should have an equal probability of assigning true output for samples with different values for protected variables. In this study, we utilized differences in recall, precision, and F1 scores for measuring bias.

3 Methodology

To investigate the effect of gender debiasing methods on internal model representations, we developed a general framework based on the online code length, a variation of MDL probing proposed by Voita and Titov (2020), to quantify the gender information contained in the model representations. We have conducted our experiments partially utilizing the code provided by Orgad et al. (2022)[1] and using two datasets and three debiasing techniques.

Datasets. Probing datasets are defined as $D = \{X, Y_p\}$, where X is the textual input and Y_p is the label of the knowledge characteristic we are investigating, which is gender information in our study. A number of datasets have been proposed with the goal of measuring fairness, either in specific tasks, or language modeling in general. Task specific datasets aim to measure societal bias using a downstream task. Datasets such as WinoBias (Zhao et al., 2018a), EEC (Kiritchenko and Mohammad, 2018), and BiosBias (De-Arteaga et al., 2019) fall into this category. On the other hand, datasets such as StereoSet (Nadeem et al., 2021), and CrowS-Pairs (Nangia et al., 2020) aim to measure societal biases using the language modeling capabilities of a pre-trained model. BiosBias (De-Arteaga et al., 2019) and Funpedia (Dinan et al., 2020) were used in our experiments, with the gender feature as the probing label. BiosBias is a set of 396,347 biographies with the occupation of the target person being the target label. Gender labels for each biography are also provided which are used for our probing task. Funpedia is a set of 23,000 biography sentences pulled from Wikipedia and rephrased to be conversational. The target label for Funpedia is the gender of the target person of the sentence. We test all of our models on 20% of the BiosBias dataset and the Funpedia evaluation set; therefore, we have adequate data to train the probing classifier as well as sufficient data to evaluate the model representations.

Model. Textual input is represented using a language model $f_\theta : X \to Z$, where X is the textual input, Z is the latent representation of the text, and θ contains the weights of the model. Experiments are conducted using model-generated representations Z. More specifically, we employed BERT base uncased model prior to and following the execution of multiple debiasing techniques. We additionally test our approach on BERT models pretrained on BiosBias and Funpedia with respect to their original objectives (occupation classification and gender classification, respectively) in order to determine the effect of pre-training in injecting gender information into model representations across various datasets.

Debiasing Methods. Debiasing methods are techniques for modifying model's weights θ using either continuous training on modified algorithms or training objectives, or by modifying the representation space using an auxiliary algorithm. To implement our framework of measuring gender information in the representations generated by debiasing methods, we investigate the following three debiasing techniques:

- Proposed by Zhao et al. (2018a), counterfactual data augmentation (CDA) is the process of automatically generating text instances that counter the stereotypical bias presented in representation. Using general terms and nouns to describe the involved groups, this technique is widely used to counteract various types of bias, particularly gender and ethnicity.

- Lauscher et al. (2021) proposed ADELE (adapter-based debiasing), in which they inject adapter modules into original pretrained language model architecture and train adapter modules using a counterfactually augmented dataset, while maintaining the original PLM parameters. They observe that their proposed method improves model's fairness without much alteration in the initial knowledge.

- Kaneko and Bollegala (2021b) proposed a post-processing debiasing method that can be applied to token-level or sentence-level representations. They assert that their proposed debiasing technique preserves semantic information captured in contextualised embeddings while removing gender-related bias through an orthogonal projection at the intermediate layers.

[1] https://github.com/technion-cs-nlp/gender_internal

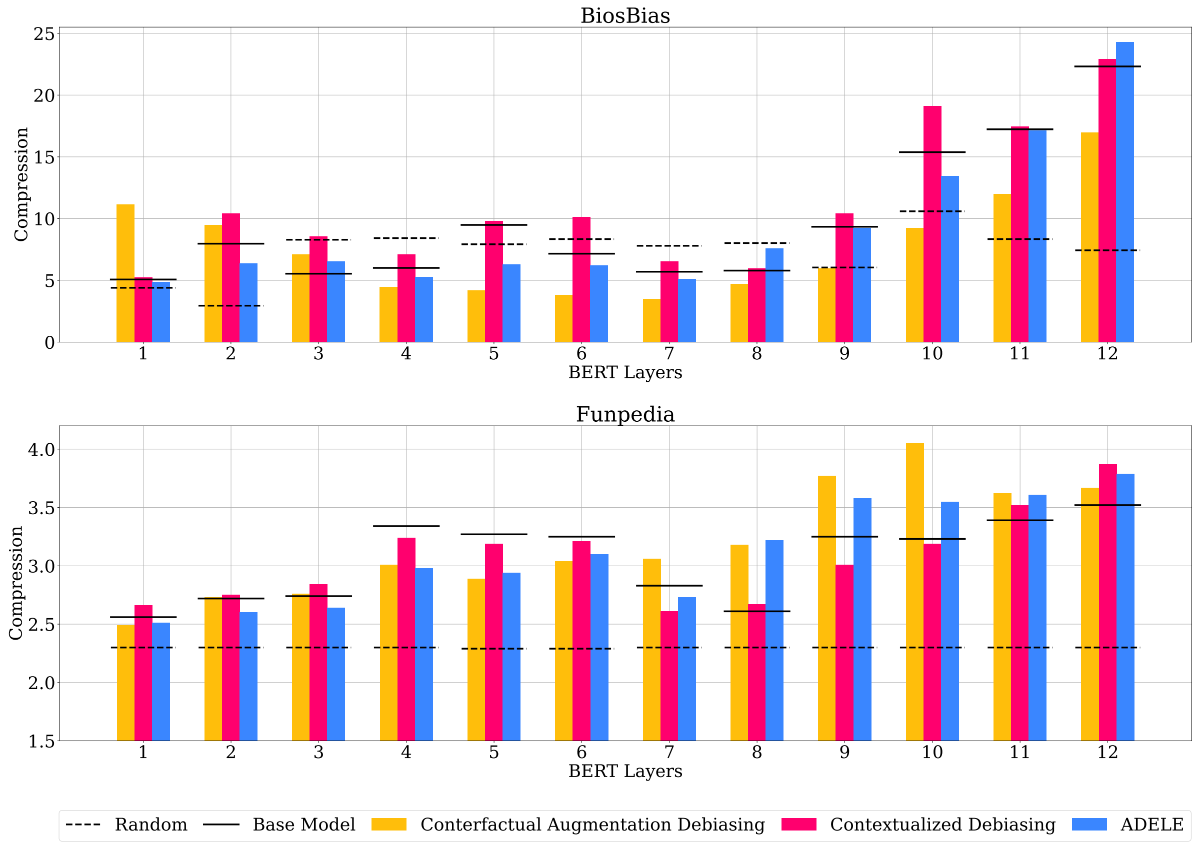

Figure 1: Layerwise compression of BERT models on Funpedia and BiosBias probing dataset. Higher values indicate that the layer contains more gender information. (See the Appendix for result tables)

Model	**Compression**
Random	7.43
Base	22.99
Fine Tuned	7.37
Contextualized Debiasing	22.91
CDA	16.98
ADELE	24.29

Table 1: Results indicating the captured gender information prior to, and after applying the debiasing techniques on the BERT base model using BiosBias dataset, as well the captured gender information when the model is fine-tuned on the occupation prediction task, or randomly initialized.

4 Representation-Level Analysis

In this section, we detail the first experiment we conduct to determine the efficacy of debiasing techniques in removing gender signals from model representations. We begin by describing our experimental setup, and then analyze and explain our findings.

Model	**Compression**
Random	2.30
Base	3.52
Fine Tuned	6.06
Contextualized Debiasing	3.87
CDA	3.67

Table 2: Results indicating the captured gender information prior to, and after applying the debiasing techniques on the BERT base model using Funpedia dataset, as well the captured gender information when the model is fine-tuned on the gender prediction task, or randomly initialized.

4.1 Experimental Setup

For our first experiment, we employ the probing datasets described in Section 3 and compute online code length, and subsequently, the compression for model representations. We carry out our experiments on a BERT base model before and after applying the three debiasing techniques described in the previous section. We followed the hypter-parameter setting of Lauscher et al. (2021) to implement counterfactual augmentation and adapter-based debiasing techniques. The Wikipedia dataset was augmented with the word pairs employed by Lauscher et al. (2021), trained both models using the standard MLM procedure for BERT training, and masked 15% of the tokens on the CDA dataset over the course of two epochs. For the experiments on contextualised representation debiasing Kaneko and Bollegala (2021b), we used the models provided in their GitHub repository.

In addition, we conduct our experiments with randomly initialized BERT base weights as a baseline for gender information extractability of representations of a random model. We expect that a randomly initialized model will capture less gender information in comparison to other models. Additionally, we conduct our tests using fine-tuned models on BiosBias and Funpedia datasets using occupation prediction and gender prediction tasks, respectively, to measure the gender information injected into the model as a result of fine-tuning. We hypothesize that the captured gender information by model representations largely depends on the task on which the model is fine-tuned. Tasks requiring gender information will lead to higher gender information captured by model representations, whereas tasks that require little gender information might decrease this information.

To determine what layers of the model capture the most gender information, we conduct probing experiments in a layerwise setting. We extract the representations of the model for each layer given a dataset, and apply Minimum Description Length probing to each representation individually and compute the associated compression.

4.2 Results

Effectiveness of Debiasing Methods in Removing Gender Information. Tables 1 and 2 show the results of layerwise probing experiments for the BiosBias and Funpedia gender prediction tasks, respectively. For the **BiosBias** dataset, we find that out of the three tested debiasing techniques, counterfactual augmentation of the dataset is the only technique that results in a reduced compression in Minimum Description Length Probing. This indicates that the other techniques fail to meaningfully reduce the gender information captured in model representations, and in the case of ADELE, increase it. This finding is particularly interesting as ADELE adapters are trained using the same procedure as counterfactual fine-tuning of the model. We believe that it might be the case that these debiasing techniques, make use of gender information to make fairer decisions with respect to a gender, rather than removing it completely. Our results in Section 5 further conforms with this hypothesis.

In the case of **Funpedia**, we find that fine-tuning a model on the gender-prediction task significantly increases the captured gender information. This contradicts our observation on the previous task, in which fine-tuning a model on the occupation prediction task significantly decreases the compression. This is in line with our previous assumption that the captured gender information largely depends on the task on which the model is trained on, meaning that when a model does not require the captured gender information, it simply discards it. Furthermore, we find no meaningful decrease in gender information when applying other debiasing techniques, showcasing the inefficacy of such techniques in removing gender information and conforming with our previous results.

Gender information is Captured in The Final Layers. Figure 1 showcases our results from the layerwise analysis experiment. We observe that later layers, layer 10 and onwards in particular, boast significantly higher compression in comparison to earlier layers. This means that these layers are extensively used during model inference regarding gender tasks. Inferring the gender of a person from a given text requires semantic knowledge over the input text to handle the required agreement between different parts of the sentence. Thus, our finding is in line with a previous work by Jawahar et al. (2019) in which they show that semantic information is mostly encoded by the later layers of the BERT model.

We find this information useful as it can be utilized while developing truly gender-neutral models by mainly focusing on layers that carry the most gender information during the debiasing phase and significantly decrease the number of trained parameters in such models.

5 Partial Debiasing

Results obtained in section 4 indicates that most of the gender information is concentrated in only a few layers of the BERT model. Namely, layers 9 through 12 contain the highest amount of encoded gender information. In this section, we apply two debiasing methods only on layers that contain the

Model	Female			Male			Δ Recall	Δ F1	Δ Precision
	Recall	Precision	F1	Recall	Precision	F1			
Base	62.22	68.22	65.08	77.53	70.39	73.79	15.31	8.71	2.17
Zari	71.11	53.01	60.74	45.75	73.85	56.5	-25.36	-4.24	20.84
CDA Full	58.47	67.89	62.83	78.87	68.71	73.44	20.39	10.62	0.82
CDA Last-4	57.04	71.86	63.60	80.92	68.56	74.23	23.88	10.63	-3.3
ADELE Full	60.72	69.15	64.66	77.03	72.93	69.24	16.31	8.27	0.09
ADELE Last-4	55.00	75.9	63.78	83.05	67.4	74.41	28.04	10.63	-8.5

Table 3: Performance Results for Base and Debiased BERT models in scrubbed gender prediction task. Δ indicates difference in a given metric and is calculated using *Metric(Male) − Metric(Female)*

Model	Female			Male			Δ Recall	Δ F1	Δ Precision
	Recall	Precision	F1	Recall	Precision	F1			
Base	78.90	78.44	78.67	81.13	80.20	80.66	2.22	1.99	1.76
Zari	82.42	79.8	81.09	84.18	81.6	82.87	1.75	1.78	1.8
CDA Full	78.66	78.44	78.55	80.75	80.39	80.57	2.09	2.02	1.95
CDA Last-4	79.00	78.54	78.77	81.17	80.59	80.88	2.17	2.12	2.05
ADELE Full	79.01	77.66	78.33	81.14	80.46	80.80	2.13	2.47	2.80
ADELE Last-4	78.53	78.17	78.35	80.68	80.74	80.71	2.15	2.36	2.57

Table 4: Performance Results for Base and Debiased BERT models in occupation prediction task. Δ indicates difference in a given metric and is calculated using *Metric(Male) − Metric(Female)*

most gender information and report our observations.

We find that debiasing the layers with the highest gender information does not adversely affect the model performance and fairness by a significant margin in comparison to debiasing the entire model, and requires the training of only a portion of model parameters. Furthermore, we find results that further support our previous hypothesis regarding the usage of gender information by debiased models to yield fairer results and not removing this information entirely.

5.1 Experimental Setup

To develop partially debiased models, we take two of the aforementioned debiasing techniques, counterfactual augmentation and ADELE adapter debiasing, and apply them to the final 4 layers of the BERT model. The training process remains the same as Section 4, but we additionally freeze the initial 8 layers of the model so that debiasing is applied only to layers 9 through 12. In the case of ADELE, adapter modules are added only to the final 4 layers of the model.

We test our models in two settings. First, we use the scrubbed version of BiosBias in which all words containing a gender indicator are replaced by a meaningless token ("_" in this case) and train a shallow classifier to predict the associated gender of each input. The shallow classifier utilizes a Sigmoid activation function. For the second test, we again use the BiosBias dataset and train a shallow classifier to predict the associated occupation of each person that the input text mentions. The dataset contains 28 classes, and Softmax activation function is used for the shallow classifier. In both cases, we use 20% of the data for testing, and the rest of the data for training. We run our tests 5 times for each model and report the average performance.

As our fairness metric, we calculate the difference in Recall, Precision and F1-score with respect to gender in both settings. e.g. the number of correctly predicted occupations for females out of all female instances of the dataset.

5.2 Results

In this section, we demonstrate and analyze our findings achieved by running BERT base and BERT debiased models on scrubbed gender prediction and occupation prediction tasks.

5.2.1 Scrubbed Gender Prediction

Table 3 showcases the results for scrubbed gender prediction task for each of our models. Somewhat surprisingly, we find that BERT base model performs the best with respect to difference in Recall out of all models, with debiased models performing noticeably worse. To validate our observations and our implementation of debiasing techniques, we utilize Zari (Webster et al., 2020), a BERT large variant pre-trained from scratch using Counterfactual Augmentation, and test it alongside our original models. We find that Zari, alongside other debiased models, perform worse than the base model with respect to Recall difference. More interestingly, we observe a trend in which models yield a higher Recall score in comparison to Precision score in male samples, while yielding a higher Precision score in female samples. Suggesting that models often assign a false-negative value to female samples, while assigning a false-positive value to male samples. This observation indicates that models have the tendency of predicting "Male" as the true label across all debiased models. The only exception to this observation is Zari, in which female samples have a higher recall score. We believe that this behavior by Zari is due to it being pre-trained from scratch using Counterfactual Augmentation, which has created different associations in comparison to the original BERT model.

We believe that this observation bolsters our previous hypothesis of debiasing techniques utilizing gender information to perform fairly in downstream tasks. With gender indicators removed from the input data in scrubbed gender prediction task, models fall back to utilizing correlations to make predictions. This observation indicates that the tested debiasing techniques do not remove underlying correlations between gender and profession in a representational level, but simply make use of the gender information that is encoded in the input data to make fairer predictions.

5.2.2 Occupation Prediction

Table 4 showcases the results for gender prediction task for each of our models. Unlike our previous observation, we find that the difference in Recall and Precision scores across genders to be much closer in this case. Furthermore, we find that the previously mentioned trend does not hold in the gender prediction task, in which models yield a higher Recall score to female samples, indicating that models refrain from using stereotypical behavior when exposed to gender information in the input data.

We find that all debiased models, including Zari and partially debiased models, increase the predictive parity (reducing the difference in Recall) in comparison to the BERT base model. Meaning that $P(\hat{Y} = 1|Y = 1, G = M) = P(\hat{Y} = 1|Y = 1, G = F)$ is further maintained in these models. On the other hand, we observe a decrease in the predictive equality (increasing the difference in Precision) in debiased models in comparison to the BERT base model. Meaning that $P(\hat{Y} = 1|Y = 0, G = M) = P(\hat{Y} = 1|Y = 0, G = F)$ is weakened in these models. We believe that this behavior might be due to the nature of the BiosBias dataset, in which most occupations have a stronger male correlation. Debiasing the model decreases the false-positive-rate of these classes for male samples, thus increasing the precision by a relatively significant margin. Female samples, however, have a weaker correlation with the occupations present in the dataset, thus their false-positive-rate is either unchanged or changed by a small margin.

Furthermore, we observe that models debiased using only the final four layers of the model exhibit no significant decrease in performance or fairness. Both partially debiased models perform comparable to the Base model, and yield a stronger predictive parity. In comparison to the fully debiased models, we observe a slight decrease in fairness metrics in partial models, which is expected due to their limited focus during the debiasing stage. Further investigation is required to completely understand the effects of partial debiasing on model fairness and behavior. However, our initial tests demonstrate promising results which can be applied to any other debiasing approach.

6 Related Work

6.1 Gender Bias

Early studies concerning gender bias in language models demonstrated that static embeddings not only encode but also amplify human-like biases in their representations (Islam et al., 2016; Bolukbasi et al., 2016). A number of studies have suggested methods for manipulating the embedding space or learning algorithm to mitigate bias in such models (Bolukbasi et al., 2016; Zhao et al., 2018b). But as demonstrated by Gonen and Goldberg (2019), these techniques only superficially remove biased information from the embedding space of the model.

The introduction of contextualised word embeddings such as BERT has raised the significance of this challenge, as manipulation in representation space is no longer as trivial as it was with static embeddings. It has been shown that contextualized language models also exhibit bias against demographic groups such as race, gender, and religion (Zhao et al., 2019; Silva et al., 2021). Similar to static embeddings, a number of techniques have been proposed to mitigate bias at various levels, including methods that modify the language model itself and methods that are applied when fine-tuning the language model for a specific downstream task. In Section 3, we discussed some of the most notable approaches for debiasing language models, which are used to reduce bias at the level of language modelling.

6.2 Bias Probing

Probing is a convenient technique for determining the nature and extent to which a model captures a particular knowledge characteristic. With the advancement of methods used to interpret model behaviour and the introduction of methods such as Minimum Description Length (Voita and Titov, 2020, MDL) (which was thoroughly explained in Section 4), many studies have built upon this technique to further investigate the knowledge captured by language models.

Mendelson and Belinkov (2021) used MDL to demonstrate that debiasing methods used to make models robust against spurious correlations between linguistic features and task labels in datasets cause the model to encode more biased information in its representations. More recently, Orgad et al. (2022) utilised MDL as a metric for assessing bias in model representations. They demonstrated that compression as an intrinsic bias metric, as compared to CEAT, the most prominent intrinsic bias measurement technique, has a much stronger correlation with extrinsic bias metrics used in conjunction with extrinsic bias mitigation techniques. Therefore, they argue that compression is a superior intrinsic bias metric than CEAT. In contrast, we investigate the retention of gender information through MDL compression after intrinsically debiasing a base model. In addition, MDL is applied layer-by-layer to determine the gender information captured by each layer.

7 Conclusions

In this work, we apply Minimum Description Length probing using two large datasets to identify the effectiveness of gender debiasing methods in removing the gender information encoded in BERT model representations. We find that, despite the success of such methods in forcing the model to reduce biased behavior in downstream tasks, they do not have a significant impact on the amount of encoded gender information in model representations.

Additionally, we conducted evaluations in a layerwise setting, showing that gender information is mostly concentrated in the later layers of the model, with the highest concentration being in layers 9 through 12. We hypothesized that the observation can be utilized to develop debiasing methods that only focus on layers with the highest gender information, decreasing the number of parameters to optimize and making more targeted changes to the original model. To test our hypothesis, we applied Counterfactual Augmentation debiasing and ADELE debiasing only to the final four layers of a BERT model. Using the occupation prediction task, we found that debiasing only the layers with the highest gender information yields no significant drawbacks with respect to model performance and fairness, making this approach worthy of investigation in future work. Additionally, and somewhat surprisingly, we found that when gender information was scrubbed from the input sentences, debiased models revert back to associating certain professions with a gender. This observation provides further support for our hypothesis that debiasing methods do not necessarily remove the encoded gender-information. On the contrary, debiased models seem to utilize this inherent information to reduce the biased behavior in downstream tasks.

8 Limitations

Due to the large amount of resources required to conduct the extensive tests mentioned in sections 4 and 5, we can only confirm the correctness of our results for the BERT models. As different models tend to encode linguistic knowledge in different layers (Fayyaz et al., 2021), it is currently difficult to generalize our observation to other models. Further testing on other models is required to find the layers that encode the gender information and observe their behavior when partially debiased.

Furthermore, our technique requires the presence of gender labels to measure the encoded gender information. This significantly reduces the datasets that our method can be applied on, which reduces its generalizability. Further methods, especially those not requiring explicit gender labels, will help in both confirming, or refuting our observations, and generalizing this approach to a more general setting.

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

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

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

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.

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

Mohsen Fayyaz, Ehsan Aghazadeh, Ali Modarressi, Hosein Mohebbi, and Mohammad Taher Pilehvar. 2021. Not all models localize linguistic knowledge in the same place: A layer-wise probing on BERToids' representations. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 375–388, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2021a. Debiasing pre-trained contextualised embeddings. In *Proc. of the 16th European Chapter of the Association for Computational Linguistics (EACL)*.

Masahiro Kaneko and Danushka Bollegala. 2021b. 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.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. 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.

Michael Mendelson and Yonatan Belinkov. 2021. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1557, Online and Punta Cana, Dominican Republic. 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.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. 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.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

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

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. 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.

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

A Result Tables

Model	Layerwise Compression												Compression Variance
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
Random	4.4	2.95	8.29	8.42	7.92	8.34	7.8	8.02	6.04	10.59	8.34	7.43	3.77
Base	4.85	6.94	5.52	5.69	8.83	7.22	5.67	5.86	9.35	**14.17**	**17.43**	**22.99**	29.90
Fine Tuned	5.29	7.64	10.12	5.76	13.57	8.44	7.34	7.34	**12.12**	**13.47**	**12.1**	7.37	7.96
Contextualized Debiasing	5.23	10.41	8.56	7.09	9.81	10.15	6.54	5.98	10.43	**19.11**	**17.47**	**22.91**	29.30
CDA	**11.14**	9.49	7.11	4.46	4.18	3.81	3.51	4.71	5.98	9.24	**11.99**	**16.98**	17.25
ADELE	4.87	6.39	6.52	5.3	6.29	6.2	5.11	7.59	9.26	**13.44**	**17.12**	**24.29**	32.83

Table 5: Layerwise compression of BERT models on BiosBias probing dataset. Each cell represents the compression achieved using either a base or debiased model from the representation extracted from the layer. Highlighted cells represent the top three layers with the highest compression.

Model	Layerwise Compression												Compression Variance
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	
Random	2.30	2.30	2.30	2.30	2.29	2.29	2.30	2.30	2.30	2.30	2.30	2.30	$1.38e{-}05$
Base	2.56	2.72	2.74	3.34	3.27	3.25	2.83	2.61	**3.25**	3.23	**3.39**	**3.52**	0.33
Fine Tuned	2.62	2.86	2.84	3.45	3.49	3.34	3.32	3.35	5.54	**6.24**	**6.18**	**6.06**	1.89
Contextualized Debiasing	2.66	2.75	2.84	3.24	3.19	3.21	2.61	2.67	3.01	**3.19**	**3.52**	**3.87**	0.13
CDA	2.49	2.73	2.76	3.01	2.89	3.04	3.06	3.18	**3.77**	**4.05**	3.62	**3.67**	0.21
ADELE	2.51	2.6	2.64	2.98	2.94	3.1	2.73	3.22	**3.58**	3.55	**3.61**	**3.79**	0.18

Table 6: Layerwise compression of BERT models on FunPedia probing dataset. Each cell represents the compression achieved using either a base or debiased model from the representation extracted from the layer. Highlighted cells represent the top three layers with the highest compression.