Improving Fairness by Debiasing Intersectional Biases in Contextual AI Models

***Abstract*—Contextual Artificial Intelligence (AI) models such as BERT and ALBERT have become foundational in natural language processing tasks due to their ability to capture semantic relationships in context. However, these models often encode social biases present in training data. While prior debiasing methods have primarily addressed single-axis biases (e.g., gender or race), intersectional bias, arising from the compounded interaction of multiple identity attributes, remains largely unaddressed in contextual embeddings. In this work, we extend a projection-based debiasing technique to mitigate intersectional bias across gender, race, and religion. We construct individual bias subspaces using word-pair difference vectors and apply a nonlinear optimization to learn an intersectional subspace. We project contextual embeddings orthogonally to this subspace and regularize the model to preserve linguistic integrity. We evaluate our method using BERT and ALBERT on the CrowS-Pairs and StereoSet benchmarks, observing consistent reductions in religious bias and minimal degradation to downstream task performance. Our results demonstrate the feasibility of intersectional debiasing in contextual models and highlight the need for fairness-aware representation learning techniques that scale with the complexity of modern NLP systems.**

***Index Terms*—AI model, intersectional bias, contextual embeddings, projection, responsible AI, fairness, Natural Language Processing**

I. Introduction

Context Artificial Intelligence (AI) models are widely used for a range of tasks, including question answering, translation, documentation, and text summarization [1], [2]. These models are increasingly embedded in tools used by individuals and organizations, raising essential concerns around their outputs' fairness, inclusivity, and potential social impact.

At the core of these models are word embeddings, which are vector representations that capture semantic relationships between words. Traditional word embeddings, such as Word2Vec [3] and GloVe [4], which have been widely adopted due to their ability to store linguistic information in a low-dimensional space but do not look at the surrounding information. Contextual word embeddings such as BERT [5] and ELMo [6] work in a different way by looking at the surrounding context of every word and calculating the word vector. While contextual embeddings have improved performance across tasks, they also inherit and potentially amplify the social biases in training data [7].

Imagine a resume screening tool powered by an AI model. While it may seem neutral, the model may rank "Fatima" — a Muslim woman — consistently lower than "James" — a white man — even when their qualifications are equivalent. This subtle but harmful discrepancy is not due to explicit rules but arises from deeply embedded biases in the model's representations. These types of intersectional harms can go unnoticed in real-world systems if fairness interventions target only one identity dimension at a time.

Typically, when attempting to debias word embeddings, researchers use social categories such as gender or race [8]. Often these social categories are debiased individually, irrespective of each other. However, these categories frequently intersect in non-trivial ways, and debiasing them individually may lead to information loss. To address this issue, researchers have proposed techniques for static word embeddings. One of these approaches leverages nonlinear geometry and uses the planes where the biases are to identify and mitigate the biases [9]. This approach is more nuanced in dealing with intersectional biases than linear techniques.

Intersectional biases are based on the compounding effect of the biases that exist within individual groups, in which the sum of the biases does not reflect the correlation or intersection of the biases [10]. For example, "hair weaves" is a term that is associated with African American females in a way that is not parallel to the association of African Americans and females as separate groups, added together [11].

Recent studies have also used Natural Language Processing (NLP) in real-world applications across critical domains. For instance, Ozyegen et al. work on medical text highlighting demonstrated how transformer-based models can reduce physician overload by prioritizing clinically relevant terms [12]. Another study by Topal et al. highlighted how fine-tuning strategies like QLoRA can improve sentiment analysis models for emotionally aware applications [13]. These examples underscore that nuances in text processing are not a purely academic concern, as they directly affect the trustworthiness, usefulness and inclusivity of AI tools in practice.

While much progress has been made in debiasing contextual embeddings and using NLP in other domains, there has been little research into debiasing intersectional biases. This study aims to extend the existing work of debiasing intersectional biases from non-contextual embeddings to contextual embeddings, aiming at improving fairness in artifacts that rely on them. The following research questions guide our research:

RQ1. How can nonlinear debiasing techniques designed for non-contextual embeddings be adapted to debiasing contextual embeddings? This question is important because the changes that have occurred in the usage of LLMs, primarily in those models that use non-contextual embeddings, have been widely phased out of use, being replaced with models that use contextual embeddings. However, the debiasing methods that work on non-contextual embeddings are not directly

applicable to those that work on contextual embeddings. To keep contextual embeddings fair, one must modify existing methods to adequately debias those embeddings.

RQ2. How effective is this debiasing model as compared to models that are linear? This question is important because the outputs of these models must be comparable to those of their unbiased counterparts. If these models' performance degrades significantly, the gains in the lack of bias will be meaningless.

To answer the research questions and address the gap, we adapt the subspace projection method by Kaneko & Bollegala [14] and integrate it with the Intersectional Hypothesis proposed by Cheng et al. [9]. This approach enables us to define and remove a biased direction that spans multiple social dimensions (e.g., gender, race, and religion) simultaneously in contextualized word representations. We then fine-tune BERT and ALBERT models using this intersectional subspace and evaluate both bias mitigation and downstream task performance.

Our results show that while intersectional debiasing via subspace projection is feasible, its effectiveness varies across social categories. It achieves notable improvements in reducing religious bias but more limited gains for gender and race, highlighting both the promise and current limitations of this approach.

II. Related Work

We review relevant work in two main areas: (1) intersectional bias in language representations and (2) debiasing techniques for contextualized embeddings.

A. Intersectional Biases

The biases that exist in LLMs often extend beyond the individual categories, such as gender, race, religion, etc. At the intersections of these categories, we get intersectional bias, which was first conceptualized by Crenshaw, referring to the fact that individually analyzing the categories does not reflect the compounding effect of multiple categories [15].

Other studies investigate how effective specific approaches would be specifically for intersectional biases [16]. These approaches are iterative null space projection and a bias-constrained model, showing that adding intersectional identities gives better results for iterative null space projection and for bias-constrained models, the gain is only mild. Magee et al., shows how prevalent intersectional biases are and the sentiments that come along with that; for example gendered nouns had more of a negative sentiment as opposed to gender neutral terms, further they found that these intersections did not add the negative sentiment of terms but multiplied them together resulting in much worse sentiments [17].

Cheng et al. proposed joint and intersectional debiasing (JoSEC); this method uses nonlinear geometry to account for multiple bias subspaces simultaneously [9]. This is done by taking the subspace representation of the biases and learning the corresponding intersectional subspace from them. This approach also does not need sets explicitly created for intersectional categories, which limited the applicability of the previous approaches as they would require a new set for each intersection.

B. Contextualized Embedding Debiasing

Unlike non-contextual embeddings, which have a fixed representation for each word, contextual embeddings create representations based on the surrounding words. This makes debiasing a much more challenging task as words may seem unbiased individually, but may have bias only within specific contexts.

Zhao et al. used data augmentation to swap the gendered words in the dataset and neutralization, creating gender-swapped versions of the testing set and then averaged the results between the original set and the gender swapped set [18]. These methods effectively eliminated the gender bias within the ELMo system.

Kaneko & Bollegala removed gender related biases through orthogonal projections in hidden layers on the token and sentence levels [14]. Their results showed that they could debias gender biases while keeping useful semantic information in the embeddings. These results also showed that projection-based debiasing methods were possible on contextualized layers, whereas previous research only showed that projection worked on the noncontextualized layers [19].

Despite these advances, most prior work addresses single-axis bias, typically gender, and few approaches attempt to generalize across multiple identity categories in contextual embeddings. To our knowledge, no prior work has adapted intersectional subspace projection to modern contextual models and empirically evaluated both bias mitigation and downstream task effects.

III. Methodology

To address the research questions outlined in the Introduction, we adapt a nonlinear subspace projection method to identify and mitigate intersectional bias in contextual word embeddings. Our method builds on prior work by Kaneko & Bollegala [14], extending it using the Intersectional Hypothesis proposed by Cheng et al. [9]. As presented in Figure 1, the process consists of three main stages: (1) constructing individual bias subspaces for social categories (e.g., gender, race, religion), (2) learning an intersectional bias subspace from these individual components, and finally, (3) debiasing contextual embeddings by projecting them orthogonally to the intersectional subspace while preserving semantic information.

A. Bias Subspace Construction

We begin by collecting contextual embeddings for word pairs that represent specific social categories (e.g., "he"/"she" for gender, "Black"/"White" for race). Each word is embedded in a series of neutral sentence templates. These sentences will minimize confounding variables using templates like:

"[WORD] is a good worker." or "The [WORD] doctor wore a coat."

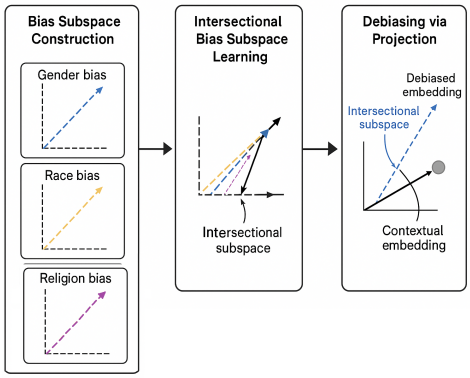

Fig. 1: Overview of the intersectional debiasing methodology.

In other words, each of these words will then be placed in several neutral contexts to keep them from being influenced by other social categories. Using a pre-trained contextual model, we extract the embeddings at the token level for each word in the respective sentences. Then average the difference of the vectors of each word pair across multiple sentences:

$$\Delta(w_1,w_2) = \text{avg\_emb}(w_1) - \text{ave\_emb}(w_2)$$

The bias subspace B will then be defined by the first K components of a Principal Component Analysis evaluation [20]. The result of this is a set of bias subspaces $\{\mathcal{B}_1, \mathcal{B}_2,...,\mathcal{B}_N\}$ where $\mathcal{B}_i$ represents a bias subspace of social category i and N is the number of considered social categories.

B. Finding The Intersectional Subspace

Once the bias subspaces are collected, we follow Cheng et al's Intersectional Hypothesis and method to find the intersectional subspace [9]. The goal is to find a direction vector $\mathcal{B}_{sec} = \hat{u}$ such that it is as close as possible to all bias subspaces. Let u be a unit-length vector. The problem can then be turned into an optimization task to minimize the squared distance between u and the bias subspaces $\mathcal{B}_i$ as shown in (1).

$$\hat{u} = \underset{||u||=1}{argmin} \sum_{i=1}^{N} d(u,\mathcal{B}_i)^2 \tag{1}$$

d(u,$\mathcal{B}_i$) is the shortest ℓ_2-distance between u and $\mathcal{B}_i$ and can be found using (2) where $\{v_{i1},...,v_{iK}\}$ are the K principal components that represent the respective bias subspace.

$$d(u,\mathcal{B}_i) = \sqrt{(||u||)^2 - \sum_{k=1}^{K}(u^T v_{ik})^2} \tag{2}$$

Using (1) and (2), the equations can be rearranged to get:

$$\hat{u} = \underset{||u||=1}{argmax} \sum_{i=1}^{N}\sum_{k=1}^{K}(u^T v_{ik})^2 \tag{3}$$

Using (3), $\hat{u}$ can then be calculated using the first principal components of $\{v_{ik}\}_{i=1,...,N;k=1,...,K}$.

C. Debiasing via Projection and Regularization

To remove the influence of the intersectional subspace from the embeddings, we apply orthogonal projection. Building on the approach by Kaneko and Bollegala [14], let E be a contextualized word embedding model and θ_e be the pre-trained model parameters. Then let $E_i(w,x;\theta_e)$ represent the embedding of token w in the i-th layer of E for a given sentence x. We can use the intersectional subspace as a replacement for the bias direction term for the non-gendered words. This would result in the debiased word embeddings being orthogonal to the intersectional subspace, which would result in the following loss L_i:

$$L_i = \sum_{t\in\mathcal{V}_t}\sum_{x\in\Omega(t)}(E_i(t,x;\theta_e)\hat{u})^2 \tag{4}$$

Where $\Omega(t)$ is a set of sentences containing non-social category-specific words.

To preserve as much useful information as possible while debiasing, we use the regulariser by Kaneko and Bollegala (2021). This regulariser measures the ℓ_2-distance between the contextualized word embedding and the debiased model as shown:

$$L_{reg} = \sum_{x\in\mathcal{A}}\sum_{w\in x}\sum_{i=1}^{N}||E_i(t,x;\theta_e) - E_i(t,x;\theta_{pre})||^2 \tag{5}$$

Where i is the ith layer of the model and θ_{pre} are the parameters based on the pretrained model before any debiasing.

This formulation ensures that embeddings are orthogonal to the bias subspace while retaining as much useful contextual information as possible. In other words, the weighted sum of the two loss functions so as to maximize information kept and debiasing achieved as seen in Equation (6).

$$L = \alpha L_i + \beta L_{reg} \tag{6}$$

Let $\alpha,\beta \in [0,1]$ such that $\alpha+\beta = 1$

IV. Experiments and Results

We evaluate the proposed debiasing method using pre-trained contextual models, BERT and ALBERT, across two dimensions: (1) reduction of intersectional bias, and (2) preservation of downstream task performance.

A. Datasets

To generate the bias subspaces, we altered the CrowS-Pairs dataset to mask the most important term according to the bias type, and that would then be substituted based on the respective social category [21]. All of the sentences that did not have bias relating to gender, race or religion were removed from the dataset. The terms used for substitution are taken from Manzini et al. and are for gender, race and religion. When there are more than two items in the pair, then two were chosen at random for $\Delta(w_1,w_2)$. For fine-tuning, the news-commentary-v18 dataset was used as development

data. The models used are BERT [5], and ALBERT[22]. The benchmarks used are evaluated on the StereoSet [23] dataset as well as the CrowS-Pairs dataset (CrowS-Pairs was split to ensure that the training set (n=100) and testing set are exclusive).

TABLE I: All Unmasked Likelihood Without Attention

Crows	Gender	Race	Religion	Bias Score
BERT	53.82	55.04	53.33	53.98
BERT (Intersect)	53.44	51.36	54.90	50.46
ALBERT	48.47	50.01	56.19	53.18
ALBERT (Intersect)	48.04	51.24	56.73	52.71
Stereo	Gender	Race	Religion	Bias Score
BERT	49.8	51.56	46.84	50.28
BERT (Intersect)	54.12	46.76	46.83	47.48
ALBERT	64.31	55.09	55.43	58.07
ALBERT (Intersect)	60.02	49.37	53.78	53.82

TABLE II: All Unmasked Likelihood With Attention

Crows	Gender	Race	Religion	Bias Score
BERT	52.13	52.13	57.14	52.85
BERT (Intersect)	51.91	49.61	57.95	49.20
ALBERT	46.95	49.61	56.19	52.98
ALBERT (Intersect)	48.22	47.9	53.64	52.16
Stereo	Gender	Race	Religion	Bias Score
BERT	48.63	54.57	51.9	51.38
BERT (Intersect)	51.76	47.07	48.3	49.44
ALBERT	62.35	54.99	56.96	58.31
ALBERT (Intersect)	58.39	51.73	52.05	53.94

TABLE III: CrowsPairs Score

Crows	Gender	Race	Religion	Bias Score
BERT	57.63	54.26	66.67	58.69
BERT (Intersect)	59.16	57.95	61.38	59.08
ALBERT	56.11	59.11	76.19	60.41
ALBERT (Intersect)	57.02	59.69	66.39	58.85
Stereo	Gender	Race	Religion	Bias Score
BERT	68.63	55.3	63.29	59.83
BERT (Intersect)	61.18	58.42	62.03	59.88
ALBERT	65.88	61.12	54.43	62.06
ALBERT (Intersect)	60.21	60.74	52.93	59.36

TABLE IV: StereoSet Score

Crows	Gender	Race	Religion	Bias Score
BERT	66.79	49.22	51.43	53.38
BERT (Intersect)	65.65	52.13	52.38	55.44
ALBERT	57.63	51.55	72.38	57.56
ALBERT (Intersect)	53.91	51.92	66.43	56.22
Stereo	Gender	Race	Religion	Bias Score
BERT	63.14	53.55	58.23	57.26
BERT (Intersect)	61.67	55.2	60.76	58.64
ALBERT	63.53	56.34	58.23	58.88
ALBERT (Intersect)	59.46	58.1	54.73	55.51

B. Bias Reduction

To measure the effectiveness of the debiasing, we use four bias metrics:

- All Unmasked Likelihood without Attention (AUL)
- All Unmasked Likelihood with Attention (AULA)
- CrowS-Pairs Score (CPS)
- StereoSet Score (SSS)

In all four, scores closer to 50 indicate lower bias. Our results (Tables I–IV), where Table I shows AUL results, Table II AULA, Table III CPS and Table IV SSS show the greatest reduction in religious bias, particularly in ALBERT. Changes in gender and race bias were smaller and inconsistent. For example, ALBERT reduced its CPS religion score by 9.8 points post-debiasing, but improvements for gender and race were marginal. Some scores even fell below 50, suggesting potential overcorrection in BERT's projections.

Across all tests, religious bias was the most consistently reduced following the debiasing. The models either decreased religious bias or maintained it at around the same level, with notable improvements, particularly for ALBERT. Gender and race bias, in contrast, remained relatively stable. There were minor changes in some settings, but no consistent or substantial improvement across either of the models or datasets. The debiased models did bring the bias scores closer to 50 in all but the SSS using BERT. These improvements were typically minor, but these bias scores include scores of social groups that were not debiased and may have been adversely affected by the debiasing of the groups used. The performance of BERT compared to ALBERT shows some discrepancies in the effectiveness of this approach. At times, BERT had the bias score reduced below 50, showing anti-stereotypical behaviour and that the intersectional subspace may be over- or under-representing specific individual biases. It is also possible that the intersection of the biases leads to some compounding effect on the individual biases when debiased. ALBERT's base model has much larger biases present. As such, the nominal values found from the benchmarks may exaggerate this approach's effectiveness, as the overall bias scores are typically worse than those of BERT.

C. Downstream Performance

To measure downstream performance, we used the token prediction accuracy of the bias reduction methods, namely: AUL/AULA, CPS, and SSS [24]. The impact on token accuracy was minimal across both models and both datasets, with minimal decreases in accuracy. BERT and ALBERT only showed minimal changes to their token prediction accuracy, with their change being less than 2% in AUL/AULA benchmarks. In the CPS benchmark, they are all minimally changed at a maximum difference of 2.02%. This suggests that the intersectional subspace carries the same information as the bias direction terms for the set of words within the social categories considered. Results are shown in Table V.

D. Discussion

The findings demonstrate that it is possible to debias contextual embeddings using an intersectional subspace, offering a method for mitigating intersectional social biases. The results suggest that intersectional debiasing through subspace projection is only partially effective. While the approach showed

TABLE V: Token Prediction Accuracy

Crows	AULA/AUL	CPS	SSS
BERT	82.76	62.98	7.66
BERT (Intersect)	82.75	61.31	7.24
ALBERT	88.02	56.18	6.34
ALBERT (Intersect)	87.9	56.32	6.08
Stereo	AULA/AUL	CPS	SSS
BERT	75.07	55.22	2.20
BERT (Intersect)	76.73	57.26	2.41
ALBERT	81.19	51.14	2.21
ALBERT (Intersect)	80.33	51.55	2.24

encouraging reductions in religious bias, its performance on gender and race was less impactful. This could be due to the smaller size of the dataset used to create the intersectional subspace, and may suggest that a larger one would have more substantial results. The results also showed that there are times when the model overcorrected the bias on some of the tests, which could also suggest that the projection of a subspace on contextual embeddings does not uniformly reduce bias. Additionally, while our method demonstrates promise, the observed bias reductions, particularly for gender and race, were modest and not consistently significant across benchmarks. This raises important questions about the stability and generalizability of intersectional subspace projection as a debiasing technique. Biased distributions in contextualized representations may be more entangled and nonlinear than current projection techniques can fully address. Further research is needed to determine whether these biases are best mitigated through projection alone or combined with other methods such as attention modulation or counterfactual data augmentation. The minimal changes in the downstream performance show that the effect of using the intersectional subspace as a replacement for hidden attributes is primarily on the biases and does not largely affect the utility of the word embeddings. This work provides empirical evidence of the Intersectional Hypothesis proposed by Cheng et al. [9], as the intersectional subspace could capture biases while not individually targeting those biases.

While this approach does advance work on intersectional debiasing, many limitations should be mentioned. Firstly, there are no robust benchmarks for intersectional biases, so it is difficult to truly see how effective any model is at debiasing these intersectional biases. We are only able to gauge the effectiveness of this model against the individual social categories that are tracked in the benchmarks used. Secondly, this approach was only done on BERT and ALBERT, so the effectiveness of projecting the intersectional subspace on decoder-only models (such as GPT) must be considered. The biases in these models may be present in more specific ways, only in particular debiased layers, or this approach may not function correctly on those models. Further, due to the much larger size of generative models, it is possible that fine-tuning in the way described initially by Kaneko & Bollegala is not possible [14]. Thirdly, once the intersectional subspace was found, it was treated as static during the fine-tuning process. Multiple intersectional subspaces that work with intersections of their respective groups should be looked at further. Another limitation is the lack of ablation analysis in our current evaluation. We did not independently assess the effects of debiasing only one or two identity dimensions (e.g., religion-only or gender-only), which would help isolate the impact of each component on bias mitigation and model utility. An ablation study could reveal whether intersectional subspaces capture emergent bias directions or aggregate known ones. Additionally, we recognize that our evaluation was restricted to BERT and ALBERT models, limiting our results' generalizability. Future work should assess whether similar intersectional debiasing can be achieved in decoder-only architectures such as LLaMA, GPT, or Claude, especially under parameter-efficient fine-tuning paradigms like LoRA or QLoRA.

Beyond social fairness, the same principles can be extended to technical biases that exist within systems when creating software artifacts, for example. By using datasets that are fair towards these technical biases, such as a preference for specific platforms or environments, we can ensure that systems do not create artifacts that favour one platform or another. Further, by using a method such as this, we can also create biased subspaces that may ignore certain platforms or environments by using a training set that doesn't contain certain topics, and through finding the intersectional subspace, we still have a robust system but without preference towards platforms that are not required.

V. Conclusion and Future Work

We presented an approach for debiasing contextual word embeddings by projecting them away from a learned intersectional subspace, combining prior techniques in projection-based debiasing and nonlinear geometric modelling. Our method was applied to BERT and ALBERT using bias subspaces derived from gender, race, and religion word pairs. Experimental results demonstrated modest but consistent reductions in religious bias, with less substantial and more variable improvements for gender and race. Notably, these fairness gains were achieved with minimal degradation to downstream task performance.

Our results highlight the promise and limitations of intersectional debiasing for contextual models. While effective in certain cases, our method occasionally produced overcorrections and struggled to generalize across all identity categories. This suggests that intersectional bias is not only multi-dimensional but dynamically distributed across layers and tasks, posing challenges for static projection methods.

The proposed intersectional debiasing technique offers a lightweight and modular approach that can be integrated into production NLP pipelines, particularly during the fine-tuning or post-processing stages of contextual models like BERT, ALBERT, or LLaMA. This makes it feasible for organizations deploying large-scale language models to improve fairness without sacrificing performance. For developers leveraging commercial LLM APIs such as GPT or Claude, our methodology provides a foundation for external bias auditing and targeted prompt refinement based on learned intersectional

subspaces. Moreover, by isolating social biases in embeddings, this work enables the development of fairness-aware applications in domains such as HR automation, educational tools, healthcare decision support, and legal tech, where inclusive and equitable AI outputs are critical. Our findings thus contribute both technically and ethically to the broader goal of responsible AI adoption in real-world systems.

For future work, we plan to (1) explore dynamic, layer-wise debiasing mechanisms; (2) extend our evaluation to decoder-only models such as GPT and Claude; and (3) investigate scalable methods for constructing richer intersectional subspaces from larger, diverse corpora. Ultimately, we aim to contribute toward more transparent and fairness-aware AI systems by refining techniques that minimize social bias while preserving linguistic utility.

REFERENCES

[1] R. Dabre, B. Buschbeck, M. Exel, and H. Tanaka, "A study on the effectiveness of large language models for translation with markup," in *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, M. Utiyama and R. Wang, Eds., Macau SAR, China: Asia-Pacific Association for Machine Translation, Sep. 2023, pp. 148–159. [Online]. Available: https://aclanthology.org/2023.mtsummit-research.13/.

[2] S. Huo, N. Arabzadeh, and C. Clarke, "Retrieving supporting evidence for generative question answering," in *Proceedings of the annual international acm sigir conference on research and development in information retrieval in the Asia Pacific region*, 2023, pp. 11–20.

[3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, 2013. arXiv: 1301.3781 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1301.3781.

[4] J. Pennington, R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, A. Moschitti, B. Pang, and W. Daelemans, Eds., Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 1532–1543. DOI: 10.3115/v1/D14-1162. [Online]. Available: https://aclanthology.org/D14-1162/.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[6] M. E. Peters, M. Neumann, M. Iyyer, *et al.*, *Deep contextualized word representations*, 2018. arXiv: 1802.05365 [cs.CL]. [Online]. Available: https://arxiv.org/abs/1802.05365.

[7] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, "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)*, J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 629–634. DOI: 10.18653/v1/N19-1064. [Online]. Available: https://aclanthology.org/N19-1064/.

[8] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[9] L. Cheng, N. Kim, and H. Liu, "Debiasing word embeddings with nonlinear geometry," *arXiv preprint arXiv:2208.13899*, 2022.

[10] A.-M. Hancock, "When multiplication doesn't equal quick addition: Examining intersectionality as a research paradigm," *Perspectives on politics*, vol. 5, no. 1, pp. 63–79, 2007.

[11] N. Ghavami and L. A. Peplau, "An intersectional analysis of gender and ethnic stereotypes: Testing three hypotheses," *Psychology of Women Quarterly*, vol. 37, no. 1, pp. 113–127, 2013.

[12] L. Ozyegen, M. Cevik, and A. Basar, "Transformer-based text highlighting for medical terms," in *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*, 2024, pp. 1–10. DOI: 10.1109/CASCON62161.2024.10837964.

[13] M. B. Topal, A. Bozanta, and A. Basar, "Sentiment analysis with llms: Evaluating qlora fine-tuning, instruction strategies, and prompt sensitivity," in *2024 34th International Conference on Collaborative Advances in Software and COmputiNg (CASCON)*, 2024, pp. 1–10. DOI: 10.1109/CASCON62161.2024.10838185.

[14] M. Kaneko and D. Bollegala, "Debiasing pre-trained contextualised embeddings," *arXiv preprint arXiv:2101.09523*, 2021.

[15] K. Crenshaw, "Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics," 1989.

[16] S. Subramanian, X. Han, T. Baldwin, T. Cohn, and L. Frermann, "Evaluating debiasing techniques for intersectional biases," *arXiv preprint arXiv:2109.10441*, 2021.

[17] L. Magee, L. Ghahremanlou, K. Soldatic, and S. Robertson, "Intersectional bias in causal language models," *arXiv preprint arXiv:2107.07691*, 2021.

[18] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, "Gender bias in contextualized word embeddings," *arXiv preprint arXiv:1904.03310*, 2019.

[19] S. Dev and J. Phillips, "Attenuating bias in word vectors," in *The 22nd international conference on artificial intelligence and statistics*, PMLR, 2019, pp. 879–887.

[20] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[21] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, *Crows-pairs: A challenge dataset for measuring social biases in masked language models*, 2020. arXiv: 2010.00133 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2010.00133.

[22] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[23] M. Nadeem, A. Bethke, and S. Reddy, *Stereoset: Measuring stereotypical bias in pretrained language models*, 2020. arXiv: 2004.09456 [cs.CL].

[24] M. Kaneko and D. Bollegala, "Unmasking the mask – evaluating social biases in masked language models," in *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, Vancouver, BC, Canada, Feb. 2022.