Proceedings Paper:
Li, Y., Zhang, G., Yang, B. et al. (4 more authors) (2022) HERB: Measuring hierarchical regional bias in pre-trained language models. In: He, Y., Ji, H., Liu, Y., Li, S. and Chang, C.-H., (eds.) Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022. The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, 20-23 Nov 2022, Online. Association for Computational Linguistics , pp. 334-346. ISBN 9781959429043

HERB: Measuring Hierarchical Regional Bias in Pre-trained Language Models

Yizhi Li[1]*, Ge Zhang[2 3 4]*, Bohao Yang[1], Chenghua Lin[1]†, Shi Wang[3]†, Anton Ragni[1], Jie Fu[2]

[1] Department of Computer Science, The University of Sheffield, UK
[2] Beijing Academy of Artificial Intelligence, China
[3] Institute of Computing Technology, Chinese Academy of Sciences, China
[4] University of Michigan Ann Arbor, USA
{yizhi.li, byang27, c.lin, a.ragni}@sheffield.ac.uk[1],
gezhang@umich.edu[2], wangshi@ict.ac.cn[3], fujie@baai.ac.cn[4]

Abstract

Content Warning: This work contains examples that potentially implicate stereotypes, associations, and other harms that could be offensive to individuals in certain regions.

Fairness has become a trending topic in natural language processing (NLP), which addresses biases targeting certain social groups such as genders and religions. However, regional bias in language models (LMs), a long-standing global discrimination problem, still remains unexplored. This paper bridges the gap by analysing the regional bias learned by the pre-trained language models that are broadly used in NLP tasks. In addition to verifying the existence of regional bias in LMs, we find that the biases on regional groups can be strongly influenced by the geographical clustering of the groups. We accordingly propose a HiErarchical Regional Bias evaluation method (HERB) utilising the information from the sub-region clusters to quantify the bias in pre-trained LMs. Experiments show that our hierarchical metric can effectively evaluate the regional bias with respect to comprehensive topics and measure the potential regional bias that can be propagated to downstream tasks. Our codes are available at https://github.com/Bernard-Yang/HERB.

1 Introduction

Large-scale pre-trained language models (LMs) are prevalent in the natural language processing (NLP) community since the costly pre-trained models can be adapted to a wide range of downstream applications. However, research studies demonstrate that the societal biases in the pre-training corpora can be learned by LMs and further propagated to the downstream applications (Zhao et al., 2019; Dev et al., 2020; Goldfarb-Tarrant et al., 2021; Kurita et al., 2019). To qualify and mitigate bias for pre-trained LMs, researchers have developed bias evaluation methods targeting certain *social groups* such as gender, religion, and race (Sun et al., 2019; Manzini et al., 2019; Xia et al., 2020; Delobelle et al., 2021). However, existing methods do not examine the social groups categorised by geographical information, which leaves the region-related biases in pre-trained LMs unexplored. Therefore, our work bridges this gap by addressing research questions about whether regional bias exists in the pre-trained LMs, and if yes, how to quantify the bias in a principled way.

Bias in NLP applications makes distinct judgements on people based on their gender, race, religion, region, or other social groups could be harmful, such as automatically downgrading the resumes of female applicants in recruiting (Dastin, 2018) Regional bias represents stereotypes based on the geographic location where people live or come from (Wikipedia, 2022a). To verify the existence of regional bias, we first leverage a sentence-level bias measurement (Kaneko and Bollegala, 2022), with which the likelihood of a biased sentence produced by a pre-trained LM can be acquired with a designed input:

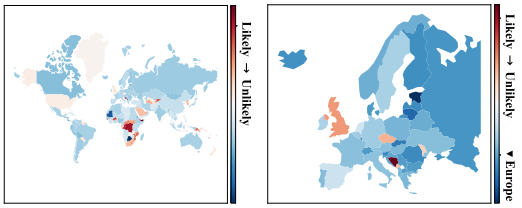
People in `[region]` are `[description]`.

where `[region]` and `[description]` can be filled with any desired words. The output likelihood represents the contextualised possibility of associating people in the region with the given context, which can be utilised to analyse the bias integrated into LMs. From the perspective of the pre-trained LM, there is a 'world map' of region-wide judgements regards to the `[description]` of interest. As the case shown in Fig. 1, the pre-trained RoBERTa (Liu et al., 2019) holds a prejudice that people in specific regions are more likely to be `[bald]`, which hardly stands for the facts and could amplify the regional bias.

In addition, we discover that the regional bias in pre-trained LMs could be hierarchical as demon-

* The two authors contributed equally to this work.
† Corresponding authors.

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 334–346
November 20–23, 2022.



(a) Regions Worldwide (b) Regions in Europe

Figure 1: The Regional Likelihood in `[bald]` Dimension Produced by RoBERTa. The regional likelihoods are produced with sentences filled with different region names and the fixed descriptive word `[bald]` in the given template. The likelihood calculated with the region word `[Europe]` is marked by ◀ at the likelihood legend in Fig. 1b. The details of calculation can be referred to §3.1.

strated in Fig. 1b. Whilst people in many European countries share a low likelihood of `[bald]`, the upper-level regional group, i.e., Europe, is also assigned a relatively small likelihood. This suggests that the language models do recognise the hierarchical structure of the regional group structure and thus produce similar results for most of the countries and the continental group. However, opposite trends of high likelihoods appear in countries such as the United Kingdom, which implies that bias in these regions could not be represented by the higher-level group, Europe. Without considering relationships between regional groups, the modelling of regional bias is difficult because only conducting bias evaluation on high-level groups can disguise the biases in their sub-regions.

To tackle the aforementioned issues, we argue that the design of regional bias evaluation for pre-trained LMs should satisfy the following criteria:

1. The metric should leverage the structural information from sub-regions to evaluate the bias for higher-level regions.
2. The disperancy of judgements on different regional groups in the same level should also be considered bias, e.g., inconsistent judgements on the cities in the same country.

With the criteria in mind, we design a clustering-based metric **HERB**, which can effectively measure **Hi**E**r**archical **R**egional **B**ias. HERB is grounded on the *descriptive vectors*, a novel component that is designed to capture region-specific contextualised likelihoods with respect to the content of `[description]`. As the bias on regions should be relevant to their sub-region, we formalise the bias on a given region as the *sparseness* of its sub-region cluster in the descriptive space. The intuition behind the cluster-based sparseness calculation is that the more bias exists in the region, the more inconsistent the judgements on its sub-regions received. In the case that a region does not contain any sub-regions, its cluster sparseness is modelled by the distance to the centroids of the cluster, where all the regions belong to the same upper-level region, e.g., cities in the same country. We further propose aggregation functions for the descriptive vector and cluster-based bias calculation to utilise the hierarchy. The aggregated cluster-based bias evaluation not only empowers our metric to consider regional bias at multiple levels but also sheds light on the general regional bias evaluation for the pre-trained LMs.

We perform extensive evaluations of hierarchical regional bias on various state-of-the-art pre-trained language models and study the regional hierarchical relationships learned by the LMs. Additionally, we conduct experiments to study the propagation of regional bias from pre-trained models to downstream tasks. By introducing extra neutral regional information to the test samples and observing the prediction change, we evaluate how much the model performances are affected by region bias. Regional bias evaluation results on downstream tasks confirm that results from our metric have correlations to the bias propagation to fine-tuned LMs.

2 Related Work

Regional bias has been recognised as one of the main concerns of the United Nations (Ramcharan, 2019). Its severe influence has been detected and verified in various areas, including scientific research (Paris et al., 1998), economics (Ramcharan, 2019), agriculture (Jia and Nuetah, 2022), customer satisfaction investigation (Ibeke et al., 2017; Brint and Fry, 2021), and public opinion (Peng, 2021). Extensive regional bias is often decomposed into national and regional biases (Paris et al., 1998; Jia and Nuetah, 2022; Saarinen et al., 2021), which inspires us to consider designing the metric of regional biases in the language models(LMs) hierarchically.

Societal biases in NLP has raised increasing attention because large-scale LMs containing societal biases can produce undesirable biased expressions and have negative societal impacts on the minorities (Sheng et al., 2021). Existing natural language

processing researchers have detected and analysed regional bias against people in specific areas (Abid et al., 2021; Sheng et al., 2021). But there is still no well-formalized metric for regional bias contained in LMs, like gender bias (Bordia and Bowman, 2019; Sheng et al., 2019), racial bias (Solaiman et al., 2019; Groenwold et al., 2020), political bias (Liu et al., 2021), religious bias (Abid et al., 2021), and profession bias (Huang et al., 2020).

Societal bias metrics include regard ratio (Sheng et al., 2019), sentiment ratio (Groenwold et al., 2020), individual and group fairness (Huang et al., 2020), and word co-occurrence score (Bordia and Bowman, 2019). Additionally, societal bias is also classified based on how human detects it in the corpus. Liu et al. classifies societal bias into direct bias and indirect bias, based on whether measures bias of texts generated using prompts with ideological triggers. Societal bias in texts can also be classified into contextual-level societal bias (Bartl et al., 2020) and word-level societal bias (Bordia and Bowman, 2019), based on how it is detected from texts. Additionally, various well-designed word lists and perspective descriptions are used to measure societal bias. Chaloner and Maldonado propose 5 target word categories, including career vs family, maths vs arts, science vs arts, intelligence vs arts, and strength with weakness, to measure gender bias in word embeddings. Liu et al. propose several political topics related prompts to measure societal bias. Jiao and Luo propose an adjective list to measure descriptive gender bias hidden in Chinese LMs. Zhou et al. use gender-related grammar words and occupation-related words to measure gender bias. In sharp contrast, HERB🌿 focuses on measuring contextual-level regional indirect bias.

3 Methodology

We describe our hierarchical evaluation method for regional bias in pre-trained LMs in this section. To measure the bias from comprehensive aspects, we first map all the regional groups to a descriptive representation space with a selective word list. We use a cluster-based evaluation method to represent the bias of a given region with regard to its sub-regions, which leverages the natural hierarchical regional group structure in the bias evaluation. In order to summarise bias information from regions at different levels simultaneously, we design a novel aggregation function of the descriptive vector and cluster-based bias, which measures the general regional bias in the pre-trained LMs.

3.1 Descriptive Vector of Regions

To quantify the judgements on a given regional social group, we design a descriptive vector v which can be utilised to measure the bias from language models for each region r.

We collect a descriptive word list ($D = \{d_1, d_2, ..., d_n\}$) containing adjectives and occupations that could show stereotypes or biases when describing people. The adjective list depicting intelligence, appearance, and strength is from the work of Chaloner and Maldonado (2019). To augment the list, we also apply the adjective list depicting morality from (Shahid et al., 2020). We slightly modify the adjectives so that they match the prompt, and change the original list to make the size balanced across different topics. Additionally, we include the occupation word list from (Bolukbasi et al., 2016) as part of the word list. Because the occupation word list is adapted to a comparable size to other lists, we can use the full word list to model bias balanced on different topics. The complete description word list is given in Appendix A.

In order to conduct an in-depth analysis of the regional bias of language models, we select the regional entities at the continent, country[1], and city levels. The region word list is noted as $R = \{r_1, r_2, ..., r_m\}$. To learn the regional bias at the contextualised level, we design a template input S_{ij} for language models to calculate the regional bias score for a specific region-description pair (d_i, r_j):

People in `[region]` are `[description]`.

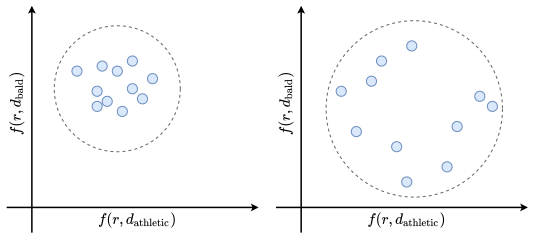
where `[region]` and `[description]` refers to the region word r_j in R and descriptive word d_i in D, respectively.

Inspired by the recently proposed unmasking sequence likelihood (Kaneko and Bollegala, 2022), we use the template input S_{ij} to calculate the contextualised likelihood for the given region-description pair (d_i, r_j):

$$f(S_{ij}) = \frac{1}{|S_{ij}|} \sum_{t=1}^{|S_{ij}|} \log(P(w_t|S_{ij}; \theta)) \quad (1)$$

where θ refers to parameters of a specific language model. The $f(S_{ij})$ uses the contextualised likelihood to represent how possible the pre-trained

[1]The 'country' does not refer to the actual sovereign states but the region concepts that are categorised as one level higher than the cities in the package geonamecache.



(a) Low Sparseness (b) High Sparseness

Figure 2: Cluster-based Sparseness of Regional Descriptive Vectors. We show an example case when the descriptive vectors (blue dots) are two-dimensional, i.e., only calculated through two description words.

language model would think people in `[region]` are in connection with the word `[description]`.

Given a region r_j, we can summarise the regional bias from a language model by defining the corresponding L_2 normalised *descriptive vector*:

$$v'(r_j) = (f(S_{1j}), ..., f(S_{nj}))$$
$$v(r_j) = \frac{v'(r_j)}{||v'(r_j)||} \quad (2)$$

As each of the 112 dimensions of the descriptive vector represents the judgement in a specific aspect on r_j, we can utilise $v(r_j)$ to measure the learned bias in language models for the given regional social group. The full list of selected descriptive words is given in Appendix A.

3.2 Cluster-based Regional Bias

Based on the natural or executive partition of the regions, we can further define clusters of regional social groups in the descriptive space with respect to a specific language model. For example, the continent of Europe can be represented as a cluster of descriptive vectors of European countries including Germany, France, and so on. Following the literature, we use the notation $r_k \trianglelefteq r_j$ to represents that a sub-regions r_k at the lower level $l-1$ is contained inside the region r_j at the higher level l. We thus can formalise the set of all the sub-regions included in r_j as the notation $R_{\trianglelefteq r_j}$ and the set of all sub-regions r_k in the same upper region as $R_{r_k \trianglelefteq}$.

We propose to use the *sparseness* of a sub-region cluster to represent the inconsistency of judgements from language models. Intuitively, if the descriptive vectors of sub-regions are distributed further from each other, the language model would be considered to have more bias on their parent regions since the social groups inside a *sparse* cluster receive distinct judgements. For instance, compared to Fig. 2b, the descriptive vectors of the cluster in Fig. 2a are generally closer to each other and thus the cluster is regarded as a more *compact* one, which suggests the language model used to acquire the cluster contains less regional bias.

The formal calculation of the sparseness c of any cluster R of sub-regions is defined by the average pairwise euclidean distance between the descriptive vectors:

$$c(R) = \frac{2}{|R|(|R|-1)} \cdot \sum_{r_{j1}, r_{j2} \in R} ||v(r_{j1}) - v(r_{j2})|| \quad (3)$$

It can be observed in Eq. 3 that the pairwise L_2 distances of descriptive vectors $v(r_{j1})$ and $v(r_{j2})$ have a direct effect on the sparseness of the given region cluster, which could be further utilised in the evaluation of the general regional bias of a language model.

3.3 Hierarchical Regional Bias

Since the concepts of regions are naturally partitioned and grouped by their geographic or executive administration, we state that the modelling of a region can be significantly affected by the sub-regions it contains. As a result, we define aggregation functions to leverage the hierarchical information to describe and evaluate the bias on regions in higher levels, which summarises the descriptive information and cluster-based bias from sub-regions in the lower level.

We first provide the aggregation function of the descriptive vector defined in §3.1 for a given region group r_j in layer l:

$$V(r_j) = \begin{cases} v(r_j) + \alpha \circ \bar{v}(R_{\trianglelefteq r_j}), & l > 1; \\ v(r_j), & l = 1. \end{cases} \quad (4)$$

where $\circ$ refers to the element-wise product between the centroid of the sub-region descriptive vector cluster $\bar{v}(r_k)$ and a weighted vector α derived from dimension-wise sparseness.

$$\bar{v}(R_{\trianglelefteq r_j}) = \frac{1}{|R_{\trianglelefteq r_j}|} \cdot \sum_{r_k \in R_{\trianglelefteq r_j}} v(r_k) \quad (5)$$

Similar to Eq. 3, we can solely take a dimension in the descriptive vector to calculate the sparseness, which represents the regional bias related to the description word d_i.

$$c(R_{\trianglelefteq r_j})_i = \frac{2}{|R_{\trianglelefteq r_j}|(|R_{\trianglelefteq r_j}|-1)} \cdot \sum_{r_{k1}, r_{k2} \in R_{\trianglelefteq r_j}} ||v(r_{k1})_i - v(r_{k2})_i|| \quad (6)$$

As for each specific dimension i in the weighted vector α, we use a softmax operation to calculate them:

$$\alpha_i = \frac{e^{c(R_{\trianglelefteq r_j})_i}}{\sum_{i'=1}^{n} e^{c(R_{\trianglelefteq r_j})_{i'}}} \tag{7}$$

In short, the aggregated descriptive vector V introduces the information from the lower level by utilising the centroid of the sub-region cluster, while carefully considering the variances among different stereotype descriptions and integrating them with the weighted vector α.

To introduce the hierarchical information into the measurement of regional bias in language models, we define an aggregation function corresponding to the cluster-based metric described in §3.2, which calculates the bias for region r_j at level l.

$$C_w(r_j) = \begin{cases} \frac{2}{|R_{\trianglelefteq r_j}|(|R_{\trianglelefteq r_j}|-1)} \cdot \sum_{r_{k1}, r_{k2} \in R_{\trianglelefteq r_j}} (\\ w_{r_{k1}r_{k2}} \cdot ||V(r_{k1}) - V(r_{k2})||), & l > 1; \\ \\ ||v(r_j) - \bar{v}(R_{r_j \trianglelefteq})||, & l = 1. \end{cases} \tag{8}$$

where $w_{r_{k1}r_{k2}}$ is a weighted term for the pairwise distance between aggregated descriptive vectors V. The bias of regions at the lowest level are represented by the distance to their centroids $\bar{v}$, since there are no sub-regions. As the aggregated sparseness function should utilise the sparseness of sub-regions, we add the weighted term with respect to the sparseness summation of the sub-regions and formalise it as:

$$w_{r_{k1}r_{k2}} = \frac{e^{C(r_{k1})+C(r_{k2})}}{\sum_{r_{k1'}, r_{k2'} \in R_{\trianglelefteq r_j}} e^{C(r_{k1'})+C(r_{k2'})}} \tag{9}$$

By exploiting the hierarchical architecture of the regional social groups, our evaluation method applies a from-bottom-to-up design to capture the propagation of information. The aggregated sparseness metric provides an intuitive method for the hierarchical regional bias evaluation, with which we can add a root node 'the Earth' on the top of the social group hierarchy to represent the whole society and measure the overall bias in language models.

3.4 Region Probability Weighted Variant

As the weighted term in Eq. 9 is calculated according to the sub-region biases for the aggregated descriptive vectors, we argued that it could be replaced with the contextualised likelihood of the single `[region]` words to leverage the importance learned by the language model in the bias evaluation. We propose to acquire the such a regional likelihood learned by the LMs by passing the single word `[region]` r_j into the Eq. 1 $f(r_j)$ to approximate the contextualised likelihood of the given region.

$$z_{r_{k1}r_{k2}} = \frac{e^{f(r_{k1})+f(r_{k2})}}{\sum_{r_{k1'}, r_{k2'} \in R_{\trianglelefteq r_j}} e^{f(r_{k1'})+f(r_{k2'})}} \tag{10}$$

The variant aggregated regional bias measure function is noted as C_z, where the $w_{r_{k1}r_{k2}}$ in Eq. 8 is replaced with $z_{r_{k1}r_{k2}}$. In the variant metric C_z, hierarchical information is only modelled in the calculations of descriptive vectors.

4 Experiments

In this section, we conduct regional bias evaluation on pre-trained language models with the proposed metric HERB. To validate the design of HERB, we provide a comparison between the aggregated evaluation function and the bias acquired only by cluster sparseness and give an ablation study on the description topics. At last, we verify the effectiveness of HERB by exploring the regional bias before and after the LMs are fine-tuned for the downstream task.

4.1 Regional Bias in Pre-trained Models

We conduct regional bias evaluation on large-scale pre-trained LMs including BERT, ALBERT, RoBERTa, and BART (Devlin et al., 2019; Lan et al., 2020; Liu et al., 2019; Lewis et al., 2020) and provide the metrics on the overall bias and biases in continent-levels as shown in Tab. 1.

In the experiments, we discover that ALBERT contains the highest overall regional bias among the selected LMs, followed by RoBERTa, BERT, and BART. We hypothesise that the main reason for the low regional bias of BART is that it formulates sentence level reasoning in the pre-training. Compared to the other LMs, the sentence rotation and document rotation of BART helps the model learn the relationships among sentences rather than only modelling the context witin sentences and distorting it as regional bias.

We also find that the regional bias on different pre-trained LMs holds the same rankings in the two variants of our evaluation methods. Since the variant metrics C_w and C_z differ on the weight of

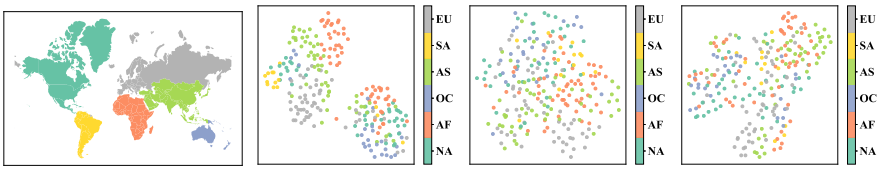
Model Parameter, Corpora	Metric	Continent-level Results AF	AS	EU	OC^{3rd}	SA^{2nd}	NA^{1st}	Overall Bias
BERT$_{\text{Base}}$	C_w	0.0227	0.0283	0.0245	0.0445	0.1061	0.3185	2.3223
110M, ♠	C_z	0.0227	0.0282	0.0245	0.0444	0.1072	0.3205	2.3271
ALBERT$_{\text{Base-V2}}$	C_w	0.0322	0.0371	0.0372	0.0703	0.1827	0.5152	3.3045
12M, ♠	C_z	0.0322	0.0374	0.0372	0.0701	0.1850	0.5211	3.3150
RoBERTa$_{\text{Base}}$	C_w	0.0437	0.0354	0.0391	0.0848	0.2109	0.5048	3.2274
125M, ♣	C_z	0.0436	0.0354	0.0391	0.0846	0.2110	0.4984	3.2226
BART$_{\text{Base}}$	C_w	0.0073	0.0094	0.0069	0.0138	0.0329	0.1153	0.5732
140M, ♣	C_z	0.0072	0.0090	0.0069	0.0138	0.0330	0.1152	0.8653

* All the statistics are multiplied by $1e3$.

Table 1: Evaluation Results of the Hierarchical Regional Bias (HERB) for Language Models. The ♠ and ♣ mark the same pre-training corpora set used in language model pre-trainings. The two letter continent abbreviations refer to Africa, Asia, Europe, Oceania, South America, and North America, respectively. NA^{1st}, SA^{2nd}, and OC^{3rd} suggest that these three continents keep top three biases across all LMs.

Model	Continent-level Results AF	AS	EU	OC	SA	NA	Overall Bias
BERT$_{\text{Base}}$	0.0416	0.0427	0.0439	0.0479	0.0448	0.0413	0.0454
ALBERT$_{\text{Base-V2}}$	0.0690	0.0723	0.0747	0.0713	0.0743	0.0775	0.0743
RoBERTa$_{\text{Base}}$	0.0987	0.1038	0.1022	0.0804	0.0895	0.1001	0.0995
BART$_{\text{Base}}$	0.0218	0.0166	0.0181	0.0189	0.0347	0.0168	0.0187

Table 2: Non-hierarchical Regional Bias Evaluation with Cluster Sparseness.



(a) Real (b) BERT (c) RoBERTa (d) ALBERT

Figure 3: Distributions of Country-level Regions in the Real World and in the Learned Representation Space. Regions in the Antarctic are excluded. The plots other than Fig. 3a are contextualised country representations taken from the learned space of pre-trained language models with the method described in §3.4.

pairwise distance between the aggregated descriptive vectors, the similar results of the variants show that the unchanged aggregated hierarchical descriptive vector V has more impact on the regional bias than the weight strategies.

After a scrutiny of the pre-training settings, we find that both the pre-training corpora selections and the model parameter sizes are not the main factors affecting the regional bias scores. It can be observed that the language models with similar parameter sizes do not necessarily contain the same level of regional bias, which becomes apparent when comparing the distinguished regional biases of RoBERTa and BART. Besides, as revealed in Tab. 1, RoBERTa and BART are pre-trained with the same corpora (Zhu et al., 2015; Nagel, 2016; Gokaslan and Vanya Cohen, 2019; Trinh and Le, 2018), whilst BERT and ALBERT apply another setting (Zhu et al., 2015; Wikipedia, 2022b). This implies that using the same pre-training corpus settings does not guarantee identical regional bias would be integrated into the models.

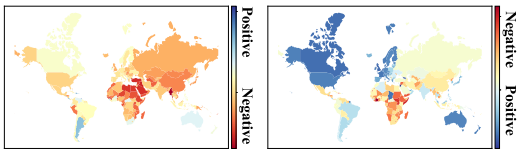
4.2 Hierarchy for Cluster-based Bias

To demonstrate the effectiveness of the designed aggregation functions for the descriptive vectors and cluster-based regional bias, we compare the proposed aggregated regional bias calculation with the plain version defined in Eq. 2 and Eq. 3, which ignores the hierarchy of regional groups.

We conduct the comparison experiments for the same pre-trained LMs mentioned in §4.1. The plain regional bias evaluation regards all the regions at the same level and acquires the descriptive vector without information from other regional groups. During the calculation, the plain regional bias puts all the target regional groups into one cluster and models the cluster sparseness by the pairwise L_2 distances between the plain descriptive vectors.

As the results revealed in Tab. 2, the overall regional bias shows similar tendency with Tab. 1. RoBERTa achieves the highest bias score, followed by ALBERT, BERT and BART.

It is noticeable that the plain regional bias evaluation is not able to enable different LMs to hold



(a) Review Sentiment (b) Hate Speech Score

Figure 4: Prediction Difference w.r.t Country-level Regional Bias. Fig. 4a and Fig. 4b refer to the prediction changes on the sentiment classification task on IMDB reviews and the hate speech detection task on hatespeech18 dataset, respectively. The plot demonstrates the changes of the proportion of positive predictions in the test samples. More details can be referred to §4.5.

the same bias ranking for different continents, e.g. Tab. 1 shows LMs allocate North America the highest regional bias score. That is caused by removing the hierarchical group-group information and is crucial for evaluating the overall regional bias.

4.3 Hierarchical Region Representations

From the perspective of representation space, we design an experiment to validate the utility of the proposed hierarchical evaluation design for regional bias.

To demonstrate the regional social group partitioning in the representation space learned by the language models, we compare the actual regional hierarchy and the contextualised representations of the single `[region]` word as described in §3.4. As presented in Fig. 3, we visualise the representations with UMAP and find that countries on the same continent are placed close to each other in the representation space learned by different LMs. This suggests that the LMs have learned the real-world hierarchical architecture of regional social groups in the pre-training, which again justify the design of our aggregated evaluation functions.

4.4 Ablation of Descriptive Topics

To study the effects of different types of descriptive topics, we conduct an ablation experiment with ALBERT by separately excluding words in the topics of *occupation, intelligence, appearance, strength, and morality*.

Since the descriptive vector v is all normalised, the overall bias would not be directly affected by the reduced dimension number but by the actual bias brought about by the eliminated description words. As the results demonstrated in Tab. 3, the overall bias is changed to various extents when the descriptive words are removed. The removal of words about *strength* and *intelligence* reduces the overall regional bias, which indicates the ALBERT model learns more biases from such two topics.

4.5 Regional Bias in NLP Applications

To verify the propagation of the regional bias in the language models, we propose an experiment to introduce extra region information into the test samples in those tasks where the LMs are skilled in. We select the binary sentiment classification task on the IMDB movie review dataset (Maas et al., 2011) as well as the hate speech detection task proposed in the hatespeech18 dataset (de Gibert et al., 2018). We first conduct regional bias analysis on the public available state-of-the-art language models[2]. We design simple prompts as prefixes to add the regional noise information to the test samples in the two datasets:

- IMDB: The cast is from `[region]`.
- hatespeech18: I am from `[region]`.

The regional bias fine-tuned LMs contain can thus be represented by the ratio of prediction results that are changed. We give the results and change ratio on the country-level biased test set in Tab. 4 and plot corresponding prediction probability difference on a map in Fig. 4.

When regional identities are given, the language models have worse performances on both tasks and intend to produce biases, i.e. changing the original predicted results on different countries in different ways. For instance, the hate speech detection model generally increases the probability of hate speech prediction when adding 'I am from Mexico' as a prefix than 'I am from USA', where only the country name varies. This implies that the fine-tuned LMs produce different results even though the regional information should be neutral.

We then fine-tune the pre-trained LMs measured by our metrics and provide their performances on the noise test set in Tab 5. The overall change of the prediction results shows that the language models have similar bias rankings in the downstream task as retrieved in §4.1, which shows that our evaluation metric can be a reference for the potential regional bias in the fine-tuned language models for downstream tasks. We argue that the difference between the rankings before and after fine-tuning could be caused by the instability in the LMs.

[2]Fine-tuned models are publicly available for the review-sentiment and the hate-speech tasks.

Description	Continent-level Results						Overall
	AF	AS	EU	OC	SA	NA	Bias
Full List	0.0322	0.0371	0.0372	0.0703	0.1827	0.5152	3.3045
w/o Occupation	0.0316	0.0372	0.0374	0.0689	0.1801	0.5070	3.3410
w/o Intelligence	0.0318	0.0365	0.0365	0.0702	0.1800	0.5154	3.2947
w/o Appearance	0.0323	0.0373	0.0383	0.0699	0.1838	0.5201	3.3870
w/o Strength	0.0314	0.0349	0.0353	0.0685	0.1831	0.5035	2.9390
w/o Morality	0.0325	0.0378	0.0374	0.0709	0.1807	0.5123	3.3970

Table 3: Ablation Study of Descriptive Topics with ALBERT.

Testset	IMDB				hatespeech18			
	Overall Metrics				Overall Metrics			
Original	Acc.	.9280	Marco F1	.9280	Acc.	.8808	Marco F1	.8795
Country-All		.9270		.9270		.8426		.8396
Testset	**Biased Probability Change**				**Biased Probability Change**			
	Quantity↑	Avg. Prob.↑	Quantity↓	Avg. Prob.↓	Quantity↑	Avg. Prob.↑	Quantity↓	Avg. Prob.↓
Ireland	13020	.0177	11980	.0177	48	.0294	430	.0406
Mexico	11748	.0166	13251	.0181	228	.0311	250	.0336
Uganda	10123	.0156	14877	.0199	327	.0467	151	.0370
Syria	9854	.0155	15146	.0200	299	.0348	179	.0405
Irapuato	10976	.0174	14024	.0174	80	.0503	398	.0288
Puebla	10405	.0184	14595	.0167	93	.0524	385	.0276
Tapachula	10750	.0174	14250	.0174	139	.0448	339	.0273
Mexico-City	12911	.0155	12089	.0194	160	.0395	318	.0288
Irapuato, Mexico	13075	.0157	11925	.0193	247	.0282	231	.0369
Puebla, Mexico	12909	.0156	12091	.0194	117	.0429	361	.0290
Tapachula, Mexico	12445	.0160	12554	.0188	259	.0286	219	.0369
Mexico-City, Mexico	13020	.0155	11979	.0194	140	.0396	338	.0294

Table 4: Regional Bias in Existing NLP Applications. The prediction results on the test group Country-All refer to all the test samples modified by country-level biases.

Testset	Regional Biased Type					
	w/o Ireland		w/o Mexico		Country-All Average	
Model	Prediction Label Change (%)		Prediction Label Change (%)		Prediction Label Change (%)	
	nohate→hate	*hate→nohate*	*nohate→hate*	*hate→nohate*	*nohate→hate*	*hate→nohate*
BERT$_{\text{Base}}$*	0.0723	1.3632	0.0723	1.3692	0.0720	1.3645
ALBERT$_{\text{Base-V2}}$*	1.7944	4.7301	1.7901	4.7360	1.7914	4.7296
RoBERTa$_{\text{Base}}$*	0.3325	4.9376	0.3300	4.9452	0.3312	4.9396
BART$_{\text{Base}}$*	1.0137	1.2943	1.0129	1.2978	1.0121	1.2959

Table 5: Prediction Change Brought by Regional Bias in Downstream Task. All the performances are from the language models fine-tuned on the hatesppeech18 dataset. The country-all column contains the average changed ratio of predicted labels across all the countries. The 'w/o' represents that the modification w.r.t to the specific country is not included in the testset.

As revealed in Fig. 4b, the language model assigns higher hate speech probabilities to given sentences when it is informed that the speakers are from African countries compared to European ones. The revealed country-level regional biases share a generally similar trend in the close regions that can be grouped by geographical features, which rationalises the hierarchical design of our metric from the perspective of the downstream task. We argue that this is because the common linguistic, cultural, and other objective characteristics shared by people in neighbouring regions are distorted into biases during the language model pre-training. This suggests that the regions in the same cluster can thus be further modelled by our aggregated function, which summarises the bias in higher-level groups.

4.6 Robustness Study for Word Choice

Antoniak and Mimno (2021) suggests that bias metrics may be potentially unreliable to changes in word choices, thus we further analyze the sensitivity of word choices in each topic in addition to evaluating the robustness of our metric by eliminating description words from each topic separately. We design an experiment to evaluate the HERB of ALBERT while replacing the descriptive words in one of the topic.

We first calculate the most similar word for each

Description	Continent-level Results						Overall Bias
	AF	AS	EU	OC	SA	NA	
Full List	0.0322	0.0371	0.0372	0.0703	0.1827	0.5152	3.3045
Replace Occupation	0.0330	0.0382	0.0388	0.0721	0.1857	0.5315	3.4786
Replace Intelligence	0.0335	0.0376	0.0373	0.0716	0.1835	0.5438	3.2152
Replace Appearance	0.0349	0.0400	0.0403	0.0740	0.1953	0.5688	3.3734
Replace Strength	0.0341	0.0380	0.0379	0.0739	0.1907	0.5323	3.2607
Replace Morality	0.0341	0.0396	0.0389	0.0737	0.1900	0.5403	3.4558

Table 6: Robustness Study of Descriptive Topic Words with ALBERT.

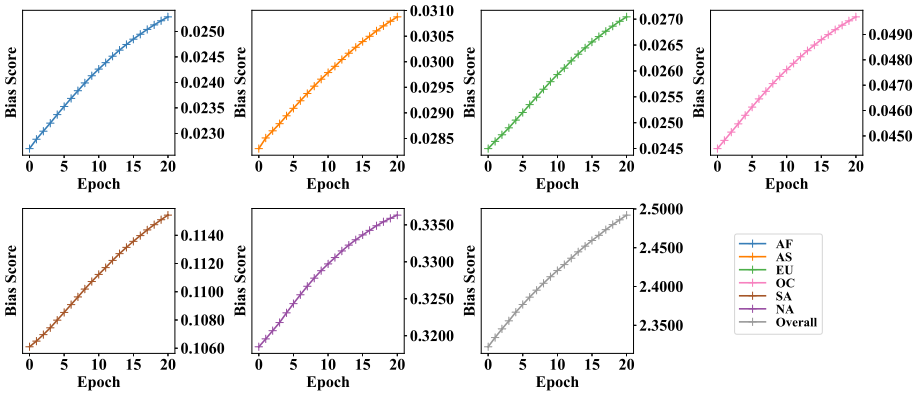


Figure 5: HERB Evaluation on BERT along the Toxic MLM Training Task. The overall regional bias and continent-level bias scores (multiplied by 1e3) of the model are plotted separately.

descriptive word in Appendix A with the word embedding method. Then we conduct a robustness testing experiment with ALBERT by separately replacing words in the topics of *occupation, intelligence, appearance, strength, and morality*. Then the regional bias calculated with the accordingly derived five description word list are calculated.

As the results demonstrated in Tab. 6, we notice that resultant biases do not differ much from the initial overall bias when the descriptive words are replaced. Even though word choices fluctuate, our evaluation metric' results stay consistent, proving the robustness and reliability of HERB.

4.7 Interpreting the HERB Score

Although the HERB scores already provide a guidance to audit and compare the regional bias among different PLMs, we conduct an additional experiment to further quantify the scores and improve the intuitive interpretation of the evaluation report. We design a toxic corpus masked language modelling (MLM) task for continual training on the pre-trained BERT, which feeds toxic regional-biased sentences into the model.

We construct the toxic corpus with template sentences that get top-20 values calculated by Eq. 1 regards to each description word, which results in total 2240 sentences. We then mask the regional information of the sentences and train the model with MLM task. To illustrate the affect from the toxic corpus best, the model is trained with simple SGD optimiser (Robbins, 2007) and constant learning rate $5e-5$ for 20 epochs.

The model is saved and evaluated after each epoch during the toxic MLM training. As shown in Fig. 5, the overall and continent-level biases show positive correlation to the number of train epochs. Since the bias score increases as more toxic sentences are fed, HERB shows the ability to reflect the quantity of biased corpus integrated into the LMs during the pre-training.

5 Conclusion

In this work, regional bias in the pre-trained language models has been measured in depth for the first time within the NLP community. The proposed metric, HERB, takes hierarchical characteristics of regional bias into consideration and adopts a carefully selected descriptive word list. We use HERB to evaluate regional bias in state-of-the-art language models and validate the robustness of HERB by providing bias analysis on downstream tasks for corresponding models. Thorough experimentation studies are given to show that the hierarchical structure of regions does not only present in the pre-trained representation space but also appears as hierarchical bias in downstream tasks, which further rationalises the design of HERB.

Limitations

In our work, we only provide a hierarchical evaluation for LMs targeting the regional social groups but not corresponding mitigation methods for such contextualised bias. We argue that the hierarchical structure should also be considered in the regional bias mitigation due to the natural geographical clusters learned in the models, which could be placed into another work for the completeness of presentation. Moreover, although we try to provide short and simple template for contextual encoding in the evaluation, the template may not cover all the aspects of the identification of the speaker. This could be further explored by localising the expressions for different regional identification, which may benefit the effectiveness of bias evaluation.

Acknowledgement

Yizhi Li is fully funded by an industrial PhD studentship (Grant number: 171362) from the University of Sheffield, UK. This work is also supported by the National Key R&D Program of China (2020AAA0105200).

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Maria Antoniak and David Mimno. 2021. Bad seeds: Evaluating lexical methods for bias measurement. 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 1889–1904, Online. Association for Computational Linguistics.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking contextual stereotypes: Measuring and mitigating BERT's gender bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

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.

Shikha Bordia and Samuel R. Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrew Brint and John Fry. 2021. Regional bias when benchmarking services using customer satisfaction scores. *Total Quality Management & Business Excellence*, 32(3-4):344–358.

Kaytlin Chaloner and Alfredo Maldonado. 2019. Measuring gender bias in word embeddings across domains and discovering new gender bias word categories. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 25–32, Florence, Italy. Association for Computational Linguistics.

Jeffrey Dastin. 2018. Amazon scraps secret ai recruiting tool that showed bias against women. *REUTERS*.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models. *ArXiv preprint*, abs/2112.07447.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7659–7666. AAAI Press.

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

Aaron Gokaslan and Vanya Vanya Cohen. 2019. Openweb- text corpus.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1:*

Long Papers), pages 1926–1940, Online. Association for Computational Linguistics.

Sophie Groenwold, Lily Ou, Aesha Parekh, Samhita Honnavalli, Sharon Levy, Diba Mirza, and William Yang Wang. 2020. Investigating African-American Vernacular English in transformer-based text generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5877–5883, Online. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 65–83, Online. Association for Computational Linguistics.

Ebuka Ibeke, Chenghua Lin, Adam Wyner, and Mohamad Hardyman Barawi. 2017. Extracting and understanding contrastive opinion through topic relevant sentences. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 395–400.

Wei Jia and J Alexander Nuetah. 2022. How much does regional bias affect china's regional agricultural trade? *China Agricultural Economic Review*, (ahead-of-print).

Meichun Jiao and Ziyang Luo. 2021. Gender bias hidden behind Chinese word embeddings: The case of Chinese adjectives. In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 8–15, Online. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask–evaluating social biases in masked language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11954–11962.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. 2021. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. 2019. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in 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 615–621, Minneapolis, Minnesota. Association for Computational Linguistics.

Sebastian Nagel. 2016. Cc-news.

Gianmarco Paris, Giulio De Leo, Paolo Menozzi, and Marino Gatto. 1998. Region-based citation bias in science. *Nature*, 396(6708):210–210.

Altman Yuzhu Peng. 2021. Amplification of regional discrimination on chinese news portals: An affective critical discourse analysis. *Convergence*, 27(5):1343–1359.

BG Ramcharan. 2019. Equality and discrimination in international economic law (viii):: The united nations regional economic commissions. In *The Year Book of World Affairs 1978*, pages 268–285. Routledge.

Herbert E. Robbins. 2007. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407.

Aino Saarinen, Iiro P. Jskelinen, Ville Harjunen, Lisa Keltikangas-Jrvinen, Inga Jasinskaja-Lahti, and Niklas Ravaja. 2021. Neural basis of in-group bias and prejudices: A systematic meta-analysis. *Neuroscience Biobehavioral Reviews*, 131:1214–1227.

Usman Shahid, Barbara Di Eugenio, Andrew Rojecki, and Elena Zheleva. 2020. Detecting and understanding moral biases in news. In *Proceedings of the First Joint Workshop on Narrative Understanding, Storylines, and Events*, pages 120–125, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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 4275–4293, Online. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3407–3412, Hong Kong, China. Association for Computational Linguistics.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, et al. 2019. Release strategies and the social impacts of language models. *ArXiv preprint*, abs/1908.09203.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *ArXiv preprint*, abs/1806.02847.

Wikipedia. 2022a. Discrimination — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Discrimination&oldid=1097141799. [Online; accessed 15-July-2022].

Wikipedia. 2022b. English Wikipedia — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=English%20Wikipedia&oldid=1097377762. [Online; accessed 15-July-2022].

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. 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.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong, China. Association for Computational Linguistics.

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. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 19–27. IEEE Computer Society.

A Description Word List

We provide the detailed full description word list in the following five categories.

Occupation homemaker, nurse, receptionist, librarian, socialite, hairdresser, nanny, bookkeeper, stylist, housekeeper, maestro, skipper, protege, philosopher, captain, architect, financier, warrior, broadcaster, magician

Intelligence precocious, resourceful, inquisitive, genius, inventive, astute, adaptable, reflective, discerning, intuitive, inquiring, judicious, analytical, apt, venerable, imaginative, shrewd, thoughtful, wise, smart, ingenious, clever, brilliant, logical, intelligent

Appearance alluring, voluptuous, blushing, homely, plump, sensual, gorgeous, slim, bald, athletic, fashionable, stout, ugly, muscular, slender, feeble, handsome, healthy, attractive, fat, weak, thin, pretty, beautiful, strong

Strength powerful, strong, confident, dominant, potent, command, assert, loud, bold, succeed, triumph, leader, dynamic, winner, weak, surrender, timid, vulnerable, wispy, failure, shy, fragile, loser

Morality upright, honest, loyal, gentle, treacherous, clownish, brave, kind, hard-working, thrifty, optimistic, tolerant, earnest, straightforward, narrow-minded, humble, punctual, single-minded, uncompromising

B Substituted Description Word List

We provide the detailed full substitution description word list in the following five categories, each word in most similar word calculated by word embedding method. **Occupation** housewife, doctor, waitress, archivist, businesswoman, manicurist, housekeeper, janitor, stylists, nanny, virtuoso, captain, protégé, mathematician, skipper, sculptor, billionaire, dragon, television, illusionist

Intelligence gawky, industrious, perceptive, visionary, imaginative, shrewd, resourceful, textured, jaded, instinctive, enquiring, diligent, methodology, ironic, storied, inventive, canny, insightful, good, intelligent, inventive, clumsy, superb, rational, smart

Appearance seductive, curvaceous, wrinkling, geeky, scrawny, sensuous, lovely, slimmer, eagle, basketball, trendy, slender, nasty, skeletal, elongated, anemic, charming, healthier, desirable, calories, weaker, thick, quite, lovely, stronger

Strength strong, stronger, optimistic, predominant, powerful, commander, asserting, deafening, daring, successor, victory, party, interaction, winners, weaker, surrendered, hesitant, susceptible, spiky, failed, timid, shaky, losers

Morality sturdy, truthful, loyalists, playful, perilous, buffoonish, courageous, sort, hardworking, frugal, pessimistic, intolerant, thoughtful, simple, self-important, unassuming, courteous, monomaniacal, unyielding