Sociolectal Analysis of Pretrained Language Models

Sheng Zhang[1,2], **Xin Zhang**[1], **Weiming Zhang**[1], **Anders Søgaard**[2]
[1]National University of Defense Technology
[2]University of Copenhagen
zhangsheng@nudt.edu.cn, shinezhang_nudt@qq.com,
wmzhang@nudt.edu.cn, soegaard@di.ku.dk

Abstract

Using data from English cloze tests, in which subjects also self-reported their gender, age, education, and race, we examine performance differences of pretrained language models across demographic groups, defined by these (protected) attributes. We demonstrate wide performance gaps across demographic groups and show that pretrained language models systematically disfavor young non-white male speakers; i.e., not only do pretrained language models learn social biases (stereotypical associations) – pretrained language models also learn sociolectal biases, learning to speak more like some than like others. We show, however, that, with the exception of BERT models, *larger* pretrained language models reduce some the performance gaps between majority and minority groups.

1 Introduction

While speakers of English generally understand each other, our linguistic preferences may differ, depending on the linguistic varieties we were exposed to, and how susceptible we were to them at the time (Tagliamonte and D'Arcy, 2009). Linguistic varieties form a multi-dimensional continuum of dialects and sociolects (McCormack et al., 2011).[1]

Such linguistic variation presents a real challenge: The linguistic preferences of English pretrained language models may align better with the linguistic preferences of some groups in society than with those of others. Group disparities constitute a fairness problem (Sokol et al., 2020): If our technologies provide end users with new opportunities, group disparities mean unequal opportunities across groups. Moreover, if the groups are defined in terms of protected attributes, our technologies may discriminate between end users in ways that violate regulations.

After waiting three hours, Cal whined and started to ___.		
	Human	
cry (0.50)	complain (0.11)	leave (0.08)
squirm (0.05)	pout (0.05)	fidget (0.04)
yell (0.04)	pace (0.03)	argue (0.02)
	Machine	
run (0.12)	bark (0.08)	pace (0.07)
cry (0.07)	laugh (0.04)	growl (0.04)
eat (0.03)	move (0.03)	rise (0.03)

Table 1: An example of a cloze (fill-in-the-gap) task with human and model predictions.

We evaluate the sociolectal biases of a range of pre-trained English language models. Unlike previous work on biases in pre-trained language models, we do not consider representational biases (Sun et al., 2019), but performance disparities; moreover, we do not consider downstream performance differences after fine-tuning for downstream tasks such as coreference resolution (Rudinger et al., 2018) or machine translation (Stanovsky et al., 2019), but performance differences across demographics of the language models themselves on cloze (fill-in-the-gap) problems. Since the cloze task is how these pre-trained language models are trained, we can evaluate models directly without introducing biases from probes or downstream tasks.

Note that some sociolinguistic variables are salient (e.g., *this*→*dis*), others are not (Jaeger and Weatherholtz, 2016). One strategy to evaluate the robustness of pre-trained language models across groups would be to identify salient variables and evaluate language models in the context of those (Demszky et al., 2021). While such evaluations are easier to interpret than evaluations of performance parity, they typically only cover a small set of variables or lectal features and therefore run the risk of only scratching the surface of sociolectal variation. In contrast, we will *not focus on salient lectal features*, but on *error rates across demographics* (precision at k and mean reciprocal rank).

[1] Dialects are mainly defined in terms of geography; sociolects in terms of demographics (Trudgill, 2003).

Contributions We align the lexical preferences of 17 commonly used pretrained language models for English with fill-in-the-gap experimental data across 16 demographics (defined by four binary variables for gender, age, race, and education). The language models systematically disfavor *young non-white male speakers*. Other groups that are poorly aligned with language models include *older white speakers*. For ELECTRA and GPT-2, bigger models are *more* fair; while for BERT, DistilBERT, and ALBERT, bigger models are *less* fair.

2 Experimental Setup

Dataset We use a publicly available cloze-style word prediction dataset[2] for our experiments. The dataset consists of fill-in-the-gap (cloze-style) example sentences with (always) the last word is removed. Table 1 presents an example sentence. The sentences are generally narrative and open-ended, and do not have standard answers. The data collectors asked the annotators to complete the sentence with what was from their own experience, the most likely one word continuation. At the same time, the annotators were asked (on a voluntary basis) to provide their demographic information (including age, gender, race, educational background). The data has been anonymized by replacing unique user IDs with generated IDs. The dataset consists of 3,085 different sentences, and the average sentence length is 10. Each sentence is annotated by 104 different annotators on average, providing 35 different continuations on average. In 40% of the sentences, the most common continuation is provided by more than half of the annotators.

The statistics of the annotators is shown in Table 2. In total, the dataset includes 307 annotators. We focus on four protected attributes: age, gender, education, and race. We binarize each attribute to obtain roughly balanced groups, binning the annotators in a total of 16 different demographic groups. For brevity, we use `emojis` to represent the 16 groups. Note that for each attribute, the two group sizes never sum to 307; this is because a few annotators did not report this information. The number of annotators in each of the 16 groups can be found in Table 3.

Pre-processing Each annotator was provided with about 1049 sentences on average. Some examples were left unanswered. The data collectors manually corrected for typos and agreement. We ignore multi-word completions.

Attribute	**Group1**	**Count**	`emojis`	**Group2**	**Count**	`emojis`
Age(yrs)	<38	147	👩🏻	>=38	159	👵🏻
Gender	Female	165	👩🏾	Male	138	👨🏾
Education(yrs)	<16	151	👨🏻	>=16	153	👨🏻‍🎓
Ethnicity	White	256	🧙🏻‍♀️	Non-White	47	🧙🏾‍♀️
Total 16 Groups	👩🏻 👨🏻 👵🏻 👴🏻 👩🏻‍🎓 👨🏻‍🎓 🧙🏻‍♀️ 🧙🏻‍♂️ 👩🏾 👨🏾 👵🏾 👴🏾 👩🏾‍🎓 👨🏾‍🎓 🧙🏾‍♀️ 🧙🏾‍♂️					

Table 2: The four protected annotator attributes. The split points (38 and 16) of numerical attributes were chosen for approximately balanced binarization.

Pre-trained language models The probed models [3] are listed here:

1. BERT (Devlin et al., 2019) language models are trained with a masked language modeling and next sentence prediction objective. The models probed cover different sizes, cased or uncased, English or multilingual: bert-base-cased, bert-base-uncased, bert-large-cased, bert-large-uncased, bert-large-uncased and bert-base-multilingual-cased.

2. The DistilBERT (Sanh et al., 2019) (distilbert-base-cased) model is distilled from original BERT model by adopting knowledge distillation. The model is 40% smaller but 60% faster than a BERT model.

3. ALBERT (Lan et al., 2020) also reduces the number of parameters in the BERT architecture, by using embedding matrix factorization and cross-layer parameter sharing. We use albert-base-v2, albert-large-v2 and albert-xxlarge-v2 below.

4. Liu et al. (2019) found that the BERT model is undertrained. They improved the pre-training by removing next sentence prediction task and obtained better results by adjusting the parameters. The model, called RoBERTa, achieves better performance in downstream tasks. We used two different sizes of RoBERTa model, which are roberta-base and roberta-large.

5. ELECTRA (Clark et al., 2020) uses a jointly trained discriminator network to distinguish the masked tokens from candidates suggested by the generator, avoiding costly inference over the full vocabulary. We use the generator models, google/electra-small-generator and google/electra-large-generator, which are suitable for the cloze-style word prediction.

6. Finally, we also include instances of the unidirectional architecture proposed in (Radford et al., 2019) (GPT). Since GPT-3 (Brown et al., 2020) is not currently open source, we probe gpt2, gpt2-medium, gpt2-large and gpt2-xl models below.

Metrics We follow Shin et al. (2020) in using precision (P@1) and mean reciprocal rank (MRR)

[2] `https://github.com/jpeelle/sentence-prediction`

[3] Model names listed in this paper are consistent with the name in 🤗 Transformers packages (`https://github.com/huggingface/transformers`). All models can be downloaded at `https://huggingface.co/models`

to evaluate the extent to which pretrained language models are aligned with annotator preferences. Given a incomplete sentence $v_1 \ldots v_n \underline{\hspace{1em}}$, and $W = [w_1, w_2, \cdots, w_r]$ the r-most frequent continuations of $v_1 \ldots v_n$ (within a group of human annotators). Our probed language model ranks candidate words by their model likelihood $C = [c_1, c_2, \cdots, c_p]$. The P@1 of the language model is then defined as:

$$\text{P@1} = \mathbb{1}[c_1 \in W] \tag{1}$$

where $\mathbb{1}[\cdot]$ is the indicator function, and MRR, as:

$$\text{MRR} = \max_{i \in [1,p]} \frac{1}{Rank_i^W} \tag{2}$$

where $Rank_i^W$ is the rank of c_i in W and equals ∞ if c_i is not in W. We report average P@1 and MRR scores.

	n	**avg**	**max**	**min**	**range**	**std**
👨🏾	2	75.4	80.0	70.8	9.2	4.6
👵🏻	29	68.6	80.0	54.0	26.0	5.9
👩🏻‍🎓	35	67.5	80.0	50.0	30.0	6.3
👨🏻‍🎓	29	67.4	80.0	50.6	29.4	6.6
👩🏻	51	66.7	82.0	52.0	30.0	6.4
👵🏾	11	66.5	74.0	58.0	16.0	4.7
🧙🏻‍♀️	20	66.0	81.0	34.3	46.7	9.9
👨🏻	28	65.7	76.4	51.4	25.0	5.1
👩🏾	4	65.6	68.3	60.0	8.3	3.3
👴🏾	4	64.9	80.0	50.9	29.1	10.7
🧙🏻‍♂️	41	64.7	75.6	40.0	35.6	8.3
👴🏻	19	64.4	75.0	45.7	29.3	7.1
🧙🏾‍♂️	10	62.0	69.2	55.9	13.3	3.8
🧙🏾‍♀️	9	61.4	75.3	37.1	38.2	11.1
👩🏾‍🎓	4	59.4	68.1	42.1	26.1	10.3
👨🏾‍🎓	3	58.1	69.3	37.0	32.3	14.9

Table 3: Statistics of P@1 scores in different demographic groups. All floating point numbers are expressed as percentages.

3 Q1: Outlier demographics?

Before comparing pretrained language models with human annotations, we first consider how continuations differ across demographic groups. We compare demographic groups by computing the average P@1 scores for individuals in each group relative to the overall majority vote (across all groups). Note that we can also compute the variance within groups by computing the P@1 scores for each annotator.

Table 3 shows group-level P@1 scores for each demographic group (**avg**) of n annotators, as well as the variance across annotators in each group: **max** is the highest average annotator P@1 within the group, and **min** the lowest. We also report the **range** (**max-min**) and the standard deviation (**std**). The gap in group P@1 values is about 17%, and we observe several outlier groups: less-educated young non-white male annotators (👨🏾), educated non-white annotators (🧙🏾‍♂️, 🧙🏾‍♀️, 👩🏾‍🎓 and 👨🏾‍🎓).

model	**max**	**min**	**avg**	**range**	**std**
bert-base-cased	50.5/29.2	45.7/24.3	47.7/26.8	**4.9**/4.9	1.3/1.2
bert-base-uncased	52.1/32.2	47.2/25.9	49.3/27.6	**4.9**/6.3	**1.2**/1.7
bert-base-multilingual -cased	20.7/9.0	14.5/6.0	15.9/7.2	6.2/**3.0**	1.4/**0.8**
bert-large-cased	**58.2**/34.2	**52.5/30.0**	54.4/32.0	5.7/4.2	1.3/1.1
bert-large-uncased	57.6/**35.6**	51.5/29.6	**54.6/32.1**	6.1/6.0	1.4/1.4
distilbert-base -uncased	45.5/24.1	39.6/20.9	42.5/22.1	5.9/**3.2**	1.4/**0.8**
albert-base-v2	36.3/17.7	31.7/14.5	33.5/15.6	**4.6/3.2**	**1.1**/0.9
albert-large-v2	49.7/27.8	43.8/22.8	45.6/24.2	6.0/5.1	1.6/1.3
albert-xxlarge-v2	**56.7/32.2**	**50.2/28.5**	**52.8/30.1**	6.5/3.7	1.5/1.0
roberta-base	59.2/35.3	54.3/31.8	57.3/33.6	**4.9/3.5**	**1.2/1.0**
roberta-large	**65.2/43.3**	**58.4/36.3**	**61.9/38.0**	6.7/7.0	1.5/1.6
google/electra-small-generator	41.7/23.2	31.6/15.8	33.9/17.1	10.0/7.5	2.3/1.8
google/electra-large-generator	**54.2/30.2**	**45.2/23.6**	**47.3/25.6**	**9.0/6.6**	**2.0/1.5**
gpt2	42.4/23.0	37.6/19.3	39.9/21.0	4.9/3.7	1.2/1.0
gpt2-medium	51.3/28.6	46.4/24.8	48.4/26.6	4.9/3.8	1.2/1.0
gpt2-large	52.2/29.8	47.3/26.4	50.3/28.0	4.8/3.4	1.4/0.9
gpt2-xl	**54.3/31.5**	**50.6/28.3**	**53.1/29.5**	**3.6/3.2**	**0.9/0.8**

Table 4: Statistics of "P@1/MRR" scores of each pretrained language models. Similar types of models are clustered together by horizontal lines, and optimums in each cluster are shown in **bold**.

4 Q2: Unfair language models?

We evaluate the pretrained language models on the cloze examples and obtain the logits vectors of the last layers corresponding to the masked tokens (gaps), perform softmax normalization to obtain the top-10 most likely candidate words and compute the fairness of the pretrained language models based on these predictions. Our fairness metric is an instance of multi-group ϵ-fairness (Donini et al., 2018), sometimes referred to as min-max Rawlsian fairness (Zafar et al., 2017), and says a model is ϵ-fair if the risk across any two groups is approximately the same. To this end, we compare the pretrained language models' range (**range**) of P@1 and MRR scores across groups. For each pretrained language model, we compute the maximum performance difference across any two groups (ϵ or **range**). If a language model m is ϵ-fair, for some value of ϵ, and no other language models are ϵ-fair, m is the most fair language model in our batch.[4]

Table 4 lists the performance of our pretrained language models across the 16 groups. We see

[4] Alternatively, we can think of the language model with the lowest divergence across groups as the most fair language model; such a definition differs from standard Rawlsian fairness, but has been advocated in Kamishima et al. (2012); Ghassami et al. (2018).

Models	Demographics Alignment															
bert-base-cased	👩🏾‍🎓	👨🏻	👴🏾	👩🏻‍🎓	🧙🏾‍♂️	🧙🏾‍♀️	👩🏻	🧙🏻‍♂️	👴🏻	👵🏻	👨🏻‍🎓	👵🏾	🧙🏻‍♀️	👩🏾	👨🏾	👨🏾‍🎓
bert-base-uncased	🧙🏾‍♀️	👩🏻‍🎓	👨🏻	👴🏾	👩🏻	🧙🏻‍♂️	👴🏻	👩🏾‍🎓	🧙🏻‍♀️	👵🏻	🧙🏾‍♂️	👨🏻‍🎓	👵🏾	👩🏾	👨🏾	👨🏾‍🎓
bert-base-multilingual-cased	👩🏾‍🎓	👨🏻	👩🏻‍🎓	👨🏻‍🎓	👩🏻	👩🏾	🧙🏻‍♀️	🧙🏾‍♂️	👵🏻	👴🏾	🧙🏾‍♀️	🧙🏻‍♂️	👴🏻	👨🏾	👵🏾	👨🏾‍🎓
bert-large-cased	👨🏻	👩🏾‍🎓	👴🏾	👩🏻‍🎓	👨🏾	🧙🏻‍♂️	👨🏻‍🎓	👵🏾	👵🏻	🧙🏾‍♂️	👴🏻	👩🏻	🧙🏾‍♀️	👩🏾	🧙🏻‍♀️	👨🏾‍🎓
bert-large-uncased	🧙🏾‍♀️	👩🏾	👴🏾	👩🏾‍🎓	👨🏻	👩🏻‍🎓	👵🏾	🧙🏻‍♂️	👴🏻	🧙🏾‍♂️	👩🏻	👨🏻‍🎓	🧙🏻‍♀️	👵🏻	👨🏾	👨🏾‍🎓
distilbert-base-uncased	🧙🏾‍♀️	👨🏻	👩🏾‍🎓	👩🏻	👩🏾	🧙🏻‍♀️	👩🏻‍🎓	👵🏾	🧙🏻‍♂️	🧙🏾‍♂️	👴🏻	👵🏻	👨🏻‍🎓	👴🏾	👨🏾	👨🏾‍🎓
albert-base-v2	👵🏾	🧙🏾‍♂️	👨🏻	👩🏻‍🎓	🧙🏻‍♀️	👩🏾‍🎓	👴🏾	👵🏻	👴🏻	🧙🏻‍♂️	🧙🏾‍♀️	👩🏻	👩🏾	👨🏻‍🎓	👨🏾	👨🏾‍🎓
albert-large-v2	🧙🏾‍♀️	👨🏻	🧙🏾‍♂️	👩🏾‍🎓	👵🏾	👩🏻‍🎓	👩🏾	👵🏻	👩🏻	👴🏾	🧙🏻‍♂️	👨🏻‍🎓	👴🏻	🧙🏻‍♀️	👨🏾‍🎓	👨🏾
albert-xxlarge-v2	👩🏾‍🎓	🧙🏾‍♂️	🧙🏾‍♀️	👴🏾	🧙🏻‍♂️	👵🏾	👨🏻	👴🏻	👩🏾	🧙🏻‍♀️	👩🏻	👨🏾‍🎓	👨🏻‍🎓	👵🏻	👩🏻‍🎓	👨🏾
roberta-base	🧙🏾‍♀️	👩🏾‍🎓	👨🏻	👨🏾‍🎓	👵🏾	👨🏾	👩🏻‍🎓	🧙🏻‍♂️	🧙🏾‍♂️	👨🏻‍🎓	👴🏻	👵🏻	👴🏾	👩🏻	🧙🏻‍♀️	👩🏾
roberta-large	🧙🏾‍♀️	👩🏾	👨🏻	👴🏻	🧙🏻‍♂️	👩🏾‍🎓	👨🏻‍🎓	🧙🏾‍♂️	👴🏾	👩🏻‍🎓	👵🏾	👩🏻	🧙🏻‍♀️	👵🏻	👨🏾‍🎓	👨🏾
google/electra-large-generator	👨🏻	🧙🏾‍♀️	👩🏾	👩🏻‍🎓	👴🏾	👩🏻	👩🏾‍🎓	🧙🏻‍♀️	🧙🏾‍♂️	👴🏻	👨🏻‍🎓	🧙🏻‍♂️	👵🏾	👵🏻	👨🏾	👨🏾‍🎓
google/electra-small-generator	🧙🏾‍♀️	👨🏻	👩🏻‍🎓	🧙🏻‍♀️	👴🏾	🧙🏾‍♂️	👩🏾‍🎓	👨🏾	👩🏻	👵🏾	🧙🏻‍♂️	👵🏻	👨🏻‍🎓	👴🏻	👩🏾	👨🏾‍🎓
gpt2	🧙🏾‍♀️	👨🏻	🧙🏾‍♂️	👩🏻‍🎓	👩🏾‍🎓	👵🏾	👩🏻	🧙🏻‍♀️	👨🏻‍🎓	👴🏻	👴🏾	👵🏻	🧙🏻‍♂️	👨🏾	👩🏾	👨🏾‍🎓
gpt2-medium	🧙🏾‍♀️	🧙🏾‍♂️	👵🏾	👨🏻	👩🏾‍🎓	👨🏻‍🎓	👴🏾	👩🏻‍🎓	👨🏾	👵🏻	🧙🏻‍♂️	👩🏻	👩🏾	🧙🏻‍♀️	👴🏻	👨🏾‍🎓
gpt2-large	🧙🏾‍♀️	🧙🏾‍♂️	👨🏾‍🎓	👩🏾‍🎓	👨🏾	👨🏻	👩🏾	👵🏾	👴🏻	👩🏻‍🎓	🧙🏻‍♂️	👨🏻‍🎓	👵🏻	🧙🏻‍♀️	👩🏻	👴🏾
gpt2-xl	🧙🏾‍♀️	👩🏾‍🎓	🧙🏾‍♂️	👨🏻	👨🏾	👵🏾	👩🏻‍🎓	👵🏻	👨🏻‍🎓	👨🏾‍🎓	🧙🏻‍♂️	👴🏻	👴🏾	👩🏾	👩🏻	🧙🏻‍♀️
Group	👨🏻	🧙🏾‍♀️	👩🏾‍🎓	🧙🏾‍♂️	👩🏻‍🎓	👵🏾	👴🏾	🧙🏻‍♂️	👩🏻	👩🏾	👴🏻	👨🏻‍🎓	🧙🏻‍♀️	👵🏻	👨🏾	👨🏾‍🎓
Mean Rank	3.1	3.4	4.0	6.1	6.1	8.1	8.1	9.2	9.8	9.9	10.3	10.3	10.8	11.1	12.0	13.8

Table 5: The alignment between different pre-trained models and demographic groups.

that roberta-large has the best overall performance across groups (across both P@1 and MRR). Using P@1 as our performance metric, gpt2-xl is most fair; using MRR, bert-base-multilingual-cased is most fair.[5] We make the following general observations: Larger models perform better, but are not necessarily more fair. For ELECTRA and GPT-2, *fairness* increases with model size. For BERT, AlBERT, and RoBERTa, the opposite is true: Group disparity increases with model size. Generally, though, the ELECTRA models are significantly more sensitive to protected attributes. google/electra-small-generator is, across both metrics, the least fair model in our batch.

5 Q3: Demographics of models?

Finally, we explore how specific pretrained language models align with group-level preferences. In other words, are different pretrained language models aligned with different demographics, or are they all biased in similar ways? Table 5 illustrates the alignment between pre-trained language models and demographic groups, and its lower part shows the mean rank of each group across all models. The correlations of different models' rankings are shown as a heat map in Figure 1 in the Appendix. We make the following observation: The language models systematically disfavor *young non-white male speakers*. Other groups that are poorly aligned with language models include *older white speakers*. google/electra-small-generator, which was, across both metrics, the least fair model in our batch, favors 👨🏻. 👨🏻 is generally the demographic that our pretrained language models align best with. This is also a demographic known to contribute the most to crowdsourced resources such as Wikipedia and social media (Hargittai and Shaw, 2015; Barthel et al., 2016), which pretrained language models are often trained on. Interestingly, the second most aligned demographic is 🧙🏾‍♀️. The two most fair models favor 🧙🏾‍♀️ (GPT-2) and 👩🏾‍🎓 (mBERT).

6 Conclusion

We compared pretrained language models to human cloze tests, and showed that models align much better with some groups of participants than others. For ELECTRA and GPT-2, larger models are more fair; for BERT, ALBERT, and RoBERTa, the opposite is true. ELECTRA models are least fair. Generally, models disfavor young non-white men the most. Previous work has explored social biases in how pretrained language models represent concepts (May et al., 2019; Kurita et al., 2019), but this is, to the best of our knowledge, the first work on *whose* language pretrained language models reflect, i.e., what sociolects models align best with. Work on personalized language modeling (Garimella et al., 2017; Welch et al., 2020) is loosely related, e.g., Stoop and van den Bosch (2014) present interesting to work to make word prediction sociolect-aware, showing significant keystroke savings by conditioning on sociolect.

[5] Both models remain the most fair under the standard deviation definition of fairness.

Ethics Statement

Our paper considers the sensitivity of pretrained language models to protected attributes. The data used in our experiments was collected using Amazon Mechanical Turk by researchers in psycholinguistics. The protected attributes are self-reported on a voluntary basis, and annotators were payed equally regardless of whether they reported this information. We find that pretrained language models *are* sensitive to protected attributes and hence biased toward some groups. For practical reasons, we use emojis as a short-hand to represent these groups. We consulted with two experts on emojis and cultural identity to make sure our use of emojis did not reinforce stereotypes, and they both assessed they would not.

References

Michael Barthel, Galen Stocking, Jesse Holcomb, and Amy Mitchell. 2016. Nearly Eight-in-Ten Reddit users get news on the site. Technical report, Pew Research Center.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pretraining text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Dorottya Demszky, Devyani Sharma, Jonathan H. Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features.

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.

Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295, Copenhagen, Denmark. Association for Computational Linguistics.

AmirEmad Ghassami, Sajad Khodadadian, and Negar Kiyavash. 2018. Fairness in supervised learning: An information theoretic approach.

Eszter Hargittai and Aaron Shaw. 2015. Mind the skills gap: the role of internet know-how and gender in differentiated contributions to wikipedia. *Information Communication and Society*, 18(4):424–442.

T Florian Jaeger and Kodi Weatherholtz. 2016. What the heck is salience? how predictive language processing contributes to sociolinguistic perception. *Frontiers in Psychology*, 7:1115.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. 2012. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, Berlin, Heidelberg. Springer Berlin Heidelberg.

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.

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 arXiv:1907.11692*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

J. McCormack, M. Pratt, and A.R.A. Rolls. 2011. *Hexagonal Variations: Diversity, Plurality and Reinvention in Contemporary France*. Faux Titre. Editions Rodopi.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Kacper Sokol, Alexander Hepburn, Rafael Poyiadzi, Matthew Clifford, Raúl Santos-Rodríguez, and Peter A. Flach. 2020. FAT forensics: A python toolbox for implementing and deploying fairness, accountability and transparency algorithms in predictive systems. *J. Open Source Softw.*, 5(49):1904.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Wessel Stoop and Antal van den Bosch. 2014. Using idiolects and sociolects to improve word prediction. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 318–327, Gothenburg, Sweden. Association for Computational Linguistics.

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.

Sali A. Tagliamonte and S. A. A. D'Arcy. 2009. Peaks beyond phonology: Adolescence, incrementation, and language change. *Language*, 85:108 – 58.

Peter Trudgill. 2003. *A glossary of sociolinguistics*. Edinburgh University Press, Edinburgh.

Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. Compositional demographic word embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online. Association for Computational Linguistics.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2017. Fairness beyond disparate treatment & disparate impact. *Proceedings of the 26th International Conference on World Wide Web*.

Appendix

A Comparison of Models

Table 6 shows the detailed information of 17 models used in this paper, including the name of the model, the number of parameters, the size of vocabulary, the tokenizer used for segmentation, and the pretraining task adopted. For more detail, we refer to the corresponding paper.

Name	Model	#param	#vocab	Tokenizer	Pretraining Task
bert-base-cased	BERT (Devlin et al., 2019)	108M	28,996	WordPiece	Masked Language Modeling (MLM) + Next Sentence Prediction (NSP)
bert-base-uncased		110M	30,522		
bert-base-multilingual-cased		178M	119,547		
bert-large-cased		334M	28,996		
bert-large-uncased		335M	30,522		
distilbert-base-uncased	DistilBERT (Sanh et al., 2019)	66M	30,522	WordPiece	MLM+NSP+Distillation
albert-base-v2	ALBERT (Lan et al., 2020)	11M	30,000	Sentence-Piece	MLM+NSP
albert-large-v2		16M	30,000		
albert-xxlarge-v2		206M	30,000		
roberta-base	RoBERTa (Liu et al., 2019)	124M	50,265	BPE	MLM
roberta-large		355M	50,265		
google/electra-small-generator	ELECTRA (Clark et al., 2020)	14M	30,522	WordPiece	MLM + Token Discrimination
google/electra-large-generator		51M	30,522		
gpt2	GPT-2 (Radford et al., 2019)	124M	50,257	Byte-Pair Encoding (BPE)	Unidirectional Language Modeling
gpt2-medium		355M	50,257		
gpt2-large		774M	50,257		
gpt2-xl		1,558M	50,257		

Table 6: The comparison of different models.

B Rank Correlations of Models

Figure 1 shows the correlation of the alignment rankings of different models. The correlation is obtained by calculating Kendall's Tau correlation coefficients of the ranking sequences in Table 5 pairwise, and displayed in a heat map. The blue-lined squares box the correlation of the same type of models. Blocks in red means that ranks of two models are highly related. On the contrary, blocks in blue means they are less related or even negatively correlated. It can be seen that the same type of models in the blue box usually have higher similarity. In addition, albert-large-v2 and gpt2 models have relatively high similarities with other models, while the gpt2-large and roberta-base models are less correlated to other models.

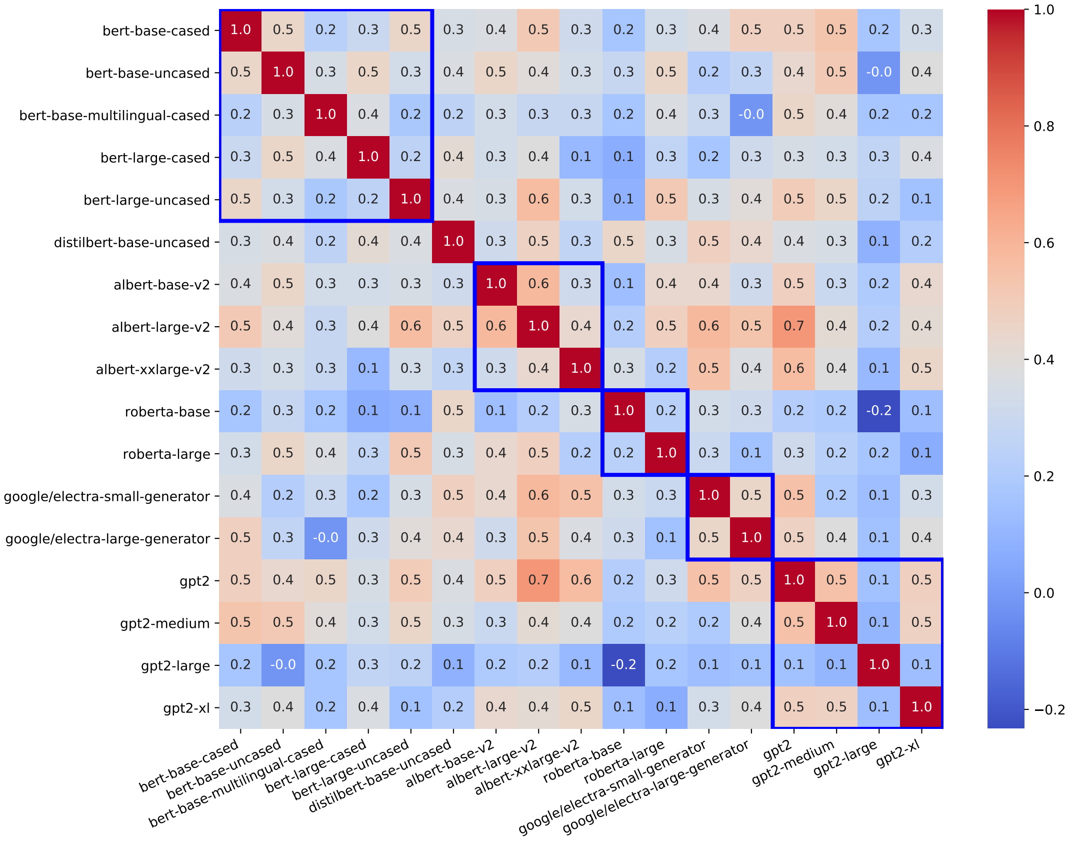

Figure 1: The heat-map of Kendall's Tau correlation coefficients between different pre-trained language models. The same type of models are in blue squares