Quantifying the Bias of Transformer-Based Language Models for African American English in Masked Language Modeling

Flavia Salutari[1,2], Jerome Ramos[1], Hossein A. Rahmani[1], Leonardo Linguaglossa[2], and Aldo Lipani[1]

[1] University College London, United Kingdom
{jerome.ramos.20, hossein.rahmani.22, aldo.lipani}@ucl.ac.uk
[2] Telecom Paris, France
{flavia.salutari,linguaglossa}@telecom.paris.fr

Abstract. In recent years, groundbreaking transformer-based language models (LMs) have made tremendous advances in natural language processing (NLP) tasks. However, the measurement of their fairness with respect to different social groups still remains unsolved. In this paper, we propose and thoroughly validate an evaluation technique to assess the quality and bias of language model predictions on transcripts of both spoken African American English (AAE) and Spoken American English (SAE). Our analysis reveals the presence of a bias towards SAE encoded by state-of-the-art LMs such as BERT and DistilBERT and a lower bias in distilled LMs. We also observe a bias towards AAE in RoBERTa and BART. Additionally, we show evidence that this disparity is present across all the LMs when we only consider the grammar and the syntax specific to AAE.

Keywords: Language Model · Transformers · Bias and Fairness · Evaluation

1 Introduction

Since their inception [8], transformers-based bidirectional encoder representations language models (LMs) have gained significant scientific interest due to their sizable improvements on a wide range of NLP tasks. The success of BERT pushed researchers to expand the state-of-the-art by introducing a plethora of model variants with differences in architecture [30], size [31,21,37] and training [24,22]. However, a growing concern in the research community has arisen: the potential societal risks coming from the pervasive adoption of these models [2]. Several studies highlight that this adoption would hinder equitable and inclusive access to NLP technologies and have real-world negative consequences in different areas, such as education, work, and politics [32]. Given the consistent emergence of new LMs trained on Web-based corpora, it is crucial to identify and measure the bias and fairness of these models.

Given the sheer size and heterogeneity of the Web, one might expect these models to be bias-free. However, even before the explosion of transformer-based LMs, a variety of biases have been identified in standard word embeddings [3]. Recently, some effort has been devoted to highlighting the presence of possible biases encoded by transformer-based LMs along gender, race, ethnicity, and disability status through techniques such as

sentiment analysis and named entity recognition tasks. In contrast, we focus on tasks used in conversational systems, where a word could be unheard or unrecognized by the automatic speech recognition system and would therefore need to be predicted.

In particular, we focus on spoken language, as it tends to have incomplete sentences, spontaneous self-corrections, and interruptions, and its register is more informal with respect to written language, which is typically more structured. We study the presence of potential bias towards English dialects spoken by underrepresented and historically discriminated groups, such as African American English (AAE). In linguistics, AAE and *mainstream* U.S. English, referred in this paper as Spoken American English (SAE), are regarded as two different languages because they are highly structured and possess their own phonological, syntactic and morphological rules [14]. In fact, AAE highlights the regional, societal and cultural environments in which individuals have learned to speak [13]. However, SAE speakers often believe that AAE is a version of SAE with mistakes and that AAE speakers belong to deficient cultures [28,36].

It is difficult to estimate the number of AAE speakers because some African Americans may speak a variety that aligns more with SAE, and not all AAE speakers are African Americans. Nevertheless, a 2019 census [29] estimates that approximately 13% of the U.S. population is African American, suggesting that there is a significant number of AAE speakers. Thus, the presence of potential linguistic biases may have discriminatory consequences towards a considerable group of individuals.

For these reasons, we set out to measure the robustness and quality of 7 transformer-based LMs in the prediction of *missed* words when the input is either SAE or AAE. We resort to two renowned corpora of spoken SAE and AAE and evaluate the LMs in a Masked Language Modeling (MLM) task. In particular, we formulate a *fill-in-the-blank* task, where we mask and predict a token, simulating its absence in every utterance. Next, we define two metrics, Probability Difference and Complementary Reciprocal Rank, to compare the likelihood that the model assigns the predicted token to the actual *masked* one and use that as a proxy of quality and fairness for the model itself.

We rigorously quantify the model bias and find that BERT, in both its cased and uncased variants, exposes a non-negligible bias towards SAE (up to 21% more accurate results with respect to AAE). Surprisingly we find that RoBERTa and BART models are biased towards AAE. We additionally observe distilled variants of these LMs to be fairer with respect to their teachers. Finally, our analysis reveals that the majority of bias resides in the AAE structural differences, specifically the particles, pronouns, and adpositions.

2 Related Work

Some of the major factors behind the success of transformer-based LMs include the large architectures and the training done on huge amounts of textual data. This recently raised the interest of the research community towards the potential societal risks linked to the employment of these models for either generating text tasks or as components of classification systems [2]. These works have studied the effects of transferring the stereotypical associations present in the training datasets to LMs, which cause an unintended bias towards underrepresented groups. Significant research efforts have been made to

identify race and gender bias embedded in large models [1,20,5,33,26]. [18] highlights the presence of topical biases in words predicted by BERT on sentences mentioning disabilities.

In addition to works on bias measurement, researchers have proposed methods to mitigate societal biases with debiasing techniques [23,34]. In regards to research on bias towards languages, most studies have focused on offensive language and hate speech detection [27,7], whereas research on bias against dialects spoken by underrepresented groups is quite recent [10]. In contrast to the above works, which mostly focus on the negative sentiment and stereotypical associations towards specific groups in BERT [8], this work focuses on quantifying the linguistic bias towards AAE for 7 different LMs: BERT, RoBERTa [24], BART [22], DistilBERT and DistilRoBERTa [31], including both their cased and uncased versions.

Previous works have proven that the large dimension of the training datasets for state-of-the-art LMs may not lead to diversity and inclusion for underrepresented groups [2]. Therefore, our analysis is essential to provide a framework to assess, reveal, and counteract the existing biases in order to improve the performance of large language models with regard to linguistic biases.

3 Methodology

To capture and provide an accurate and comprehensive account of societal biases embedded in state-of-the-art LMs, we leverage two corpora of spoken English. These are widely used in the linguistics field because linguists consider them a fair representation of their spoken language. Although there is a 15-year gap between the collection date of these corpora, we argue that the core structure of the language remains the same and that any bias captured due to the difference in periods will be minimal. Additionally, although LMs are generally trained using text data, we argue that spoken conversational agents leverage these same LMs when communicating with users. Thus, it is still appropriate to analyze the bias of LMs using spoken corpora. We also note that while this paper is not the first to study the presence of societal biases, to the best of our knowledge, this is the first to provide a thorough characterization of it for AAE across different models tested on an MLM task. We summarize LMs' performance by means of statistical metrics, which are used to characterize both the bias and the quality of the models.

3.1 Corpora for Spoken English

For SAE, we leverage the Santa Barbara Corpus of Spoken American English (SBC-SAE) [11], which is widely adopted for different applications, such as the assessment of political risk faced by U.S. firms [16], the measure of grammatical convergence in bilingual individuals [4], and the exploration of new-topic utterances in naturally occurring dialogues [25]. It includes conversations recorded in various real everyday life situations from a wide variety of people who differ in gender, occupation, and social background. All the audio recordings are also complemented with their transcriptions.

Since SBCSAE consists of speakers from several regional origins (except for the African American speakers that we preliminary filter out), we ensure that we do not craft

Table 1. Corpora summary: with and without filtering utterances ($\mathcal{U}$) based on their length. With $\langle \ell_u \rangle$ we indicate the average utterance length; with L, the length of the corpus in number of words, and; with $|\mathcal{T}|$, the number of terms (unique words).

Type	Corpus	Language	$\lvert\mathcal{U}\rvert$	$\lvert\langle \ell_u \rangle\rvert$	L	$\lvert\mathcal{T}\rvert$
Original	CORAAL	AAE	90,493	6.22	563,037	17,214
	SBCSAE	SAE	40,838	7.14	291,513	12,324
Filtered	CORAAL	AAE	63,814	8.23	525,067	16,352
	SBCSAE	SAE	25,113	8.38	210,430	10,540

the results by inducing unwanted bias when comparing AAE with a version of SAE that could be more similar to the Written American English, which is instead rather different from the spoken *mainstream* U.S. English.

For AAE, we use the Corpus of Regional African American Language (CORAAL) [19], which also provides the audio recordings along with their time-aligned orthographic transcriptions, of particular interest for this work. CORAAL includes 150 sociolinguistic interviews for over a million words. It is periodically updated and is the only publicly available corpus of AAE. As such, it has been used in the literature for a plethora of tasks, ranging from dialect-specific speech recognition [10] to cross-language transfer learning [17].

In this work, we only focus on the CORAAL:DCB portion, since it is comprised of the most recent interviews (carried out between 2015 and 2017) and contains the largest amount of data (more than 500k words). It includes conversations from 48 speakers raised in Washington, DC, a city with a long-standing African American population.

For each corpus, we define $\mathcal{U} = \{u_1, u_2, ..., u_n\}$ as the set of all the available utterances and $\mathcal{T} = \{t_1, t_2, ..., t_n\}$ as the set of all terms (unique words). Since we perform an utterance-level analysis, we first filter out noise. In particular, we discard both short utterances (composed of just one or two words) and very long ones (greater than 50 words).

In Table 1, we report a summary of the corpora statistics, both before and after having applied the filtering based on the utterance length. Even though the sizes of the two datasets are very different, not only in terms of the number of utterances $|\mathcal{U}|$, but also in terms of the total number of words L and terms $|\mathcal{T}|$, we can see that, after the filtering, the average utterance length $\langle \ell_u \rangle$ across corpora is very similar (~ 8 words per utterance).

3.2 Bias in Masked Language Modeling

In order to measure the bias in LMs we perform an MLM task. We leverage the transformer-based $\text{BERT}_{\text{base}}$ LM [8] and its recent variants, including $\text{DistilBERT}_{\text{base}}$ [31], in both their cased and uncased flavors, $\text{RoBERTa}_{\text{base}}$ [24], $\text{DistilRoBERTa}_{\text{base}}$, and $\text{BART}_{\text{base}}$ [22]. These LMs have all been pre-trained using an MLM objective, which consists of randomly masking 15% of the tokens using a special `[MASK]` token. Note that these models are trained on different corpora, such as `OpenWebText` and `BooksCorpus`.

Therefore, by directly querying the underlying MLM in each LM, we simulate the typical scenario where a conversational system has to infer a *missed* word in an utterance. In particular, we encode each utterance of the two corpora with the *tokenizer* of the LM considered. We then iteratively mask each word w_{mask} and predict the masked word by feeding the model with only a context of 10 tokens surrounding w_{mask}.

The LM provides each run with a list of possible terms to *fill-in-the-blank*. In the vocabulary set $\mathcal{T}$, we select the predicted term t_p with the highest probability $P(t_p|c)$, that is, ranks first in the list $\rho(t_p|c) = 1$, where c is the context surrounding t_p and ρ is the rank of $t|c$. In this notation, a word w is a term t in a context c ($t|c$). We next retrieve the corresponding probability $P(t_m|c)$ and the rank $\rho(t_m|c)$ for the actual masked token t_m from the vocabulary of possible terms $\mathcal{T}$. The latter provides a measure of how likely the LM will choose t_m as a candidate token to replace the masked one w_{mask}. We then employ the probabilities difference $\Delta P(t|c)$ as a proxy of the quality of the prediction for a single token, defined as:

$$\Delta P(t|c) = P(t_p|c) - P(t_m|c) = \Delta P(w). \tag{1}$$

We further define for each token $t|c$ the Complementary Reciprocal Rank (CRR) as:

$$\text{CRR}(t|c) = 1 - \rho(t_m|c)^{-1} = \text{CRR}(w). \tag{2}$$

Note that this is the difference between the reciprocal rank (RR) of the predicted token, which is always equal to 1 ($\rho(t_p|c)^{-1} = 1$), and the RR of the masked token.

We then define the probability difference for an utterance by averaging the probability difference for each token in the utterance:

$$\Delta P(u) = \frac{1}{\ell_u} \sum_{w \in u} \Delta P(w), \tag{3}$$

with ℓ_u being the length of the utterance in terms of tokens. Similarly, we define the CRR for an utterance as:

$$\text{CRR}(u) = \frac{1}{\ell_u} \sum_{w \in u} \text{CRR}(w). \tag{4}$$

Note that the metrics based on the ranks $\rho(t|c)$ generated by the LMs are necessary to fully capture the bias embedded in the models, since $\Delta P(t|c)$ alone could be insufficient. This is because the $\Delta P(t|c)$ strongly depends on how the LM assigns the probability. For example, the probability distribution of $P(t|c)$ could be more uniform and, consequently, would lead to, on average, a smaller $\Delta P(t|c)$. Conversely, a more skewed distribution would cause larger differences $\Delta P(t|c)$. Thus, CRR is used because it is unaffected by such differences in the output probability distribution of $P(t|c)$.

4 Results and Discussion

In this section, we first provide an accurate overview of the measured fairness of LMs and then further analyze the discovered biases from different viewpoints. We show how they vary when we take into account the syntactical, grammatical, and lexical patterns typical of AAE language.

Table 2. MAE and MSE of $\Delta P(u)$ and CRR(u) measured on AAE and SAE corpora: results obtained through the *fill-in-the-blank* task with different language models. † signifies that the AAE and SAE expectations are statistically significant according to Welch's two-tailed t-test (p-value < 0.05). The column d contains their effect size computed according to Cohen's d.

	MAE								MSE							
	$\Delta P(u)$				CRR(u)				$\Delta P(u)$				CRR(u)			
Model	**AAE**	**SAE**	**Δ**[%]	**d**	**AAE**	**SAE**	**Δ**[%]	**d**	**AAE**	**SAE**	**Δ**[%]	**d**	**AAE**	**SAE**	**Δ**[%]	**d**
BERT$_{\text{cased}}$	0.217	**0.171**	21 †	0.417	0.497	**0.441**	11 †	0.272	0.060	**0.040**	33 †	0.345	0.289	**0.233**	20 †	0.262
BERT$_{\text{uncased}}$	0.242	**0.198**	18 †	0.352	0.494	**0.446**	10 †	0.232	0.074	**0.053**	29 †	0.297	0.288	**0.238**	18 †	0.230
DistilBERT$_{\text{cased}}$	0.113	**0.108**	5 †	0.081	0.627	**0.589**	6 †	0.188	0.017	**0.016**	2 †	0.015	0.436	**0.385**	12 †	0.203
DistilBERT$_{\text{uncased}}$	0.126	**0.118**	6 †	0.104	0.578	**0.530**	8 †	0.222	0.021	**0.020**	1	0.007	0.380	**0.325**	15 †	0.223
RoBERTa	**0.223**	0.261	-15 †	0.368	**0.536**	0.592	-9 †	0.252	**0.061**	0.079	-23 †	0.311	**0.337**	0.396	-15 †	0.225
DistilRoBERTa	**0.143**	0.153	-7 †	0.137	**0.644**	0.668	-4 †	0.117	**0.026**	0.029	-11 †	0.112	**0.457**	0.487	-6 †	0.115
BART	**0.156**	0.193	-20 †	0.506	**0.613**	0.682	-10 †	0.346	**0.030**	0.043	-31 †	0.447	**0.418**	0.501	-17 †	0.328

4.1 Measuring the Bias of LMs

As described in Section 3, we test the fairness of transformer-based LMs by running experiments in an MLM setting. We use ΔP and CRR as metrics for measuring the quality and the fairness of the models towards the two investigated languages. We observe the expected behaviour of the LMs with respect to each utterance and consider an aggregate measure of the metrics on a per-utterance level.

Table 2 reports an overview of the results of $\Delta P(u)$ and CRR(u). Using a Welch's t-test [35], we find that the difference between the means of AAE and SAE for both $\Delta P(u)$ and CRR(u) is significant (p-value < 0.05). We then measure their effect size using Cohen's *d* [6], which is reported in the last two columns of Table 2. According to Cohen's classification, there is a *small* effect for both metrics and a *medium* effect for BART on $\Delta P(u)$ (d>0.5). We summarize the quality of the prediction in the corpora using Mean Absolute Error (MAE) and Mean Squared Error (MSE) for both $\Delta P(u)$ and CRR(u).

These error measures are used to quantify the quality of the predicted terms, where an MAE and MSE closer to 0 corresponds to an utterance having more accurately predicted terms. Therefore, in Table 2, we highlight the values leading to the smallest error between AAE and SAE. Additionally, we emphasize the presence of bias by pointing out the percentage of bias change of each LM $\mathbf{\Delta}[\%]$, which is calculated with respect to the model with the largest bias.

Three main patterns clearly emerge from Table 2. First, BERT and DistilBERT, in both their cased and uncased variants, show a bias towards SAE for all the metrics. Specifically, BERT not only presents a non-negligible bias against AAE but also is the LM which leads to the highest relative bias. In particular, we observe that the MAE$(\Delta P(u))$ for SAE is more than 20% lower than AAE, 11% lower for the MAE(CRR(u)), 33% for the MSE$(\Delta P(u))$, and 20% for the MSE(CRR(u)).

Second, DistilBERT, in both its cased and uncased flavours, and DistilRoBERTa, are the models which perform better with regards to the average probability difference $\Delta P(u)$. This is true both in terms of MAE and MSE, which are approximately half and one-third of the other LMs. On the one hand, this could seem somewhat unexpected since one could argue that DistilBERT is less accurate than BERT, achieving only 97%

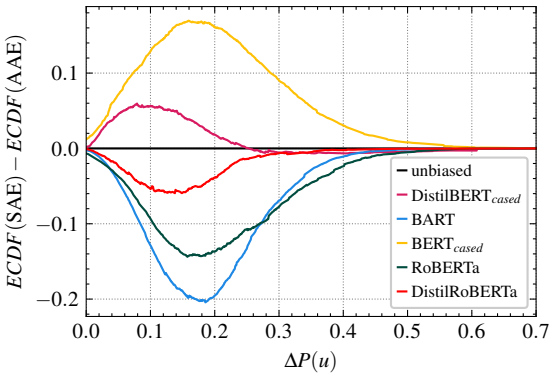

Fig. 1. The difference between the ECDFs of SAE and AAE for the $\Delta P(u)$ measure. When the values are greater than zero, the LMs are more biased towards SAE, *vice versa* otherwise.

of its performance [31]. On the other hand, this is in line with recent work [2] reporting that such LMs sometimes exceed the performance of the original ones. However, as mentioned in Section 3, it is crucial to also look at the CRR(u) because better behaviour in terms of $\Delta P(u)$ could, in practice, be tied to the fact that the model generates more uniformly distributed probabilities $P(t|c)$ with respect to the others.

Finally, we observe that BART, despite having good prediction quality for AAE (MAE$(\Delta P(u))$ and MSE$(\Delta P(u))$ are lower than BERT), shows an opposite trend with respect to BERT and DistilBERT. This unexpected bias towards AAE is also introduced by RoBERTa and DistilRoBERTa. This is somewhat surprising and could possibly be attributed to the type of datasets they have been trained on. RoBERTa and BART are pre-trained with 1000% more data than BERT. By diving into the type of data involved, we discover multiple sources, ranging from English language encyclopedias and literary works (same as BERT) to news articles and Web content. Specifically, RoBERTa, BART, and DistilRoBERTa leverage OpenWebText [12], a corpus which includes filtered Web content obtained by scraping the social media platform Reddit, which may expose the LMs to less *standard* American English. It would be interesting to explore in future studies whether there is a significant difference between shared colloquial transcriptions in OpenWebText and AAE when compared to OpenWebText and SAE.

Since Table 2 only reports a summary of the distributions of the bias metrics computed on both datasets, we also analyze the bias measured by subtracting the empirical cumulative distribution functions (ECDFs) of $\Delta P(u)$ of AAE to that of SAE, which is shown in Fig. 1. This figure includes the bias measured for the LMs and, for the sake of simplicity, only reports the *cased* variants of BERT and DistilBERT. The solid black line at $y = 0$ shows the optimal unbiased LM and visually separates what is biased against AAE (on the positive y-axis) from what is biased against SAE (on the negative y-axis). Thus, we clearly see the behaviours of the LMs leading to the two worst biases, *i.e.*, RoBERTa and BERT$_{cased}$, which are consistently biased towards one side. They both present the maximum bias when $\Delta P(u)$ is close to 0.2 and instead mitigate for larger values. A similar behaviour is observed for the CRR(u).

Table 3. A sample of AAE utterances selected based on their syntactical features and their translations to SAE. In brackets the prevalence of the feature over utterances in the AAE corpus.

Original	Translated
Double Negative (0.7%)	
• *You don't need nothing but you.*	• *You don't need anything but you.*
• *I don't know nobody over there no more.*	• *I don't know anyone over there anymore.*
Verb *be* (2.8%)	
• *And I be okay with it.*	• *And I am okay with it.*
• *All of my friends was from like DC.*	• *All of my friends were from DC.*
Contractions (4.6%)	
• *I'm'a ask you.*	• *I'm going to ask you.*
• *something gonna happen.*	• *something is going to happen.*

Table 4. Similar to Table 2 but calculated over a sample of 50 utterances of AAE and their translated version (AAE$^\intercal$) for each feature of AAE.

	MAE								MSE							
	$\Delta P(u)$				CRR(u)				$\Delta P(u)$				CRR(u)			
Model	**AAE**	**AAE$^\intercal$**	**Δ**[%]	**d**	**AAE**	**AAE$^\intercal$**	**Δ**[%]	**d**	**AAE**	**AAE$^\intercal$**	**Δ**[%]	**d**	**AAE**	**AAE$^\intercal$**	**Δ**[%]	**d**
Double Negative [50 utterances]																
BERT$_{\text{cased}}$	0.202	**0.159**	21 †	0.591	0.391	**0.334**	15 †	0.493	0.046	**0.030**	34 †	0.526	0.166	**0.125**	25 †	0.436
BERT$_{\text{uncased}}$	0.216	**0.187**	14	0.358	0.404	**0.340**	16 †	0.503	0.053	**0.041**	23	0.319	0.179	**0.130**	27 †	0.476
DistilBERT$_{\text{cased}}$	0.137	**0.106**	22 †	0.548	0.506	**0.441**	13 †	0.523	0.022	**0.014**	37 †	0.504	0.267	**0.213**	21 †	0.457
DistilBERT$_{\text{uncased}}$	0.148	**0.117**	21 †	0.485	0.479	**0.394**	18 †	0.701	0.025	**0.018**	27	0.293	0.240	**0.174**	28 †	0.611
RoBERTa	0.202	**0.181**	10	0.227	0.434	**0.383**	12	0.328	0.048	**0.042**	14	0.180	0.208	**0.175**	16	0.243
DistilRoBERTa	0.170	**0.134**	21 †	0.572	0.581	**0.498**	14 †	0.628	0.034	**0.020**	41 †	0.567	0.347	**0.272**	22 †	0.529
BART	0.164	**0.140**	15 †	0.422	0.534	**0.471**	12 †	0.469	0.030	**0.023**	22	0.368	0.297	**0.245**	18	0.392
Verb *be* [50 utterances]																
BERT$_{\text{cased}}$	0.252	**0.184**	27 †	0.691	0.589	**0.408**	31 †	1.142	0.074	**0.043**	42 †	0.622	0.373	**0.190**	49 †	1.109
BERT$_{\text{uncased}}$	0.287	**0.216**	25 †	0.642	0.595	**0.417**	30 †	1.009	0.094	**0.059**	37 †	0.520	0.383	**0.205**	46 †	0.943
DistilBERT$_{\text{cased}}$	0.134	**0.119**	11	0.273	0.703	**0.540**	23 †	0.910	0.021	**0.017**	16	0.198	0.519	**0.329**	37 †	0.893
DistilBERT$_{\text{uncased}}$	0.138	**0.118**	14 †	0.339	0.678	**0.513**	24 †	0.904	0.022	**0.017**	25	0.344	0.485	**0.302**	38 †	0.856
RoBERTa	0.246	**0.211**	14 †	0.403	0.609	**0.458**	25 †	0.800	0.069	**0.051**	26	0.380	0.405	**0.246**	39 †	0.766
DistilRoBERTa	0.169	**0.142**	16 †	0.425	0.723	**0.554**	23 †	0.947	0.032	**0.024**	25	0.389	0.549	**0.343**	38 †	0.931
BART	0.161	**0.144**	11	0.305	0.672	**0.556**	17 †	0.672	0.029	**0.024**	18	0.246	0.474	**0.344**	27 †	0.627
Contractions [50 utterances]																
BERT$_{\text{cased}}$	0.225	**0.181**	19 †	0.507	0.470	**0.347**	26 †	0.848	0.058	**0.040**	32 †	0.436	0.247	**0.136**	45 †	0.786
BERT$_{\text{uncased}}$	0.258	**0.205**	21 †	0.605	0.482	**0.355**	26 †	0.880	0.075	**0.049**	34 †	0.541	0.257	**0.143**	45 †	0.796
DistilBERT$_{\text{cased}}$	0.135	**0.114**	16	0.381	0.584	**0.463**	21 †	0.746	0.022	**0.016**	28	0.316	0.369	**0.237**	36 †	0.743
DistilBERT$_{\text{uncased}}$	0.140	**0.113**	19 †	0.477	0.538	**0.410**	24 †	0.799	0.023	**0.016**	33	0.374	0.318	**0.191**	39 †	0.761
RoBERTa	0.215	**0.193**	10	0.264	0.500	**0.402**	20 †	0.584	0.054	**0.043**	20	0.242	0.281	**0.186**	34 †	0.574
DistilRoBERTa	0.154	**0.130**	16 †	0.436	0.601	**0.488**	19 †	0.668	0.027	**0.020**	28 †	0.411	0.386	**0.268**	31 †	0.635
BART	0.143	**0.136**	5	0.117	0.567	**0.475**	16 †	0.562	0.023	**0.023**	1	0.015	0.346	**0.255**	26 †	0.520

4.2 Bias on AAE Features

Next, we investigate how results change when we acknowledge the lexical, syntactical, morphological, and phonological rules of AAE. Following AAE grammar [15], we choose to focus on three major syntactical features: (i) the use of *double* negatives, (ii)

the different usage of verb *be* and, finally, (iii) the contractions of words and groups of words.

For (i), we search for the close presence of multiple forms of grammatical negation (which in standard English are ungrammatical) in all the utterances of the AAE corpus and find that 0.7% of the utterances contain this feature. We then focus on feature (ii) and select the AAE utterances that exhibit the use of the *aspectual be* verb, typically used to denote habitual or iterative meaning (*e.g.*, *I be okay with it* in Table 3). Additionally, we filter for utterances with the verb tense in the *-ing* form where the verb is either omitted (*e.g.*, *It depends on where you going to* in Table 3) or left at the base form (*e.g.*, *they be getting mad* in Table 3), for a total of 2.8% of utterances. Finally, for (iii), we include utterances containing non-standard contractions, *e.g.*, *I'm'a, ain't* or omitting the auxiliary before *gonna*, *e.g.*, *something gonna happen* in Table 3. We do not include contractions which are popular in SAE, as *wanna, won't, aren't, etc.* We obtain 4.6% of the utterances in this class. After filtering the utterances corresponding to the specific grammar patterns, we carefully manually validate our selection by randomly picking and inspecting 1% of them. We check that the 1% random sampled utterances satisfy our criteria. From this manual labelling, we double-check our syntactical-rules-based selection strategies and find that they are 99% accurate for all three cases.

Next, we randomly choose 50 utterances from each AAE case and build a ground truth by *translating* the AAE utterances into a version compliant to SAE, which we define as $\text{AAE}^{\intercal}$. We keep the translation process as neutral as possible by preserving the contractions typical of standard English and considered in dictionaries and grammar books [9] as *short form* or *informal* and only *adjust* the selected grammar rules. Table 3 reports some examples of the utterances extracted from each AAE grammar case bucket and the corresponding translated ones.

Finally, we repeat the MLM experiments, as described in Section 3, on these 150 translated utterances $\text{AAE}^{\intercal}$ and measure the bias. We report the results in Table 4. According to Cohen's classification, there is a prevalent *medium* effect for both the metrics, with the exception of $\text{MSE}(\text{CRR}(u))$ for the *verb be* class, where it is *large*.

At first glance, we observe that the errors for the set of the AAE utterances in the verb *be* class are larger than the other two classes and the whole AAE corpus (reported in Table 2). We observe that, on average, the three classes show a less accurate average prediction with respect to the overall AAE corpus. Instead, we find that the translated utterances $\text{AAE}^{\intercal}$ are better predicted with respect to AAE, surprisingly for all seven LMs.

Notably, we observe that for the translated utterances in the *double negative* class, all four metrics are always smaller (and hence a sign of better performance) than those measured for the SAE corpus. This is somewhat unexpected since RoBERTa and BART showed a bias towards AAE. However, we note that this may be attributed to the fact that the SAE corpus, *SBCSAE*, is made up of conversations collected from people with different regional origins. Consequently, despite the effort we make in trying not to excessively standardize the utterances during the translation process, we could be generating sentences which are free from regional biases and consequently *"cleaner"* than those found in the SAE corpus.

5 Conclusion

This work proposes a methodology for evaluating the fairness of transformer-based language models. We assess and analyze the bias for two corpora, one for SAE and one for AAE. By directly querying the underlying MLM in seven LMs, we study the quality and bias of their predictions from several angles.

Results presented in this paper suggest that different models embed different biases. For example, the most popular state-of-the-art LMs, namely BERT and DistilBERT, show a non-negligible bias towards SAE, with the quality of the predictions being up to 21% more accurate than AAE. In contrast, BART, RoBERTa and DistilRoBERTa exhibit the opposite effect, with a bias leaning towards AAE. Our experiments also reveal that the distilled variants of BERT and RoBERTa are the fairest among the seven tested LMs.

Although this paper provides the first insightful snapshot of linguistic bias embedded in different LMs, it opens up a number of research questions. First, can fairer prediction outcomes be achieved with an ensemble learner of LMs embedding opposite biases, as, for instance, BERT$_{cased}$ and BART? Second, our results give insights into how the bias could be consistently mitigated with more inclusive corpora, by taking into account AAE features. Finally, special care could be put into the analysis of the distilled LMs, narrowing the gap on the causes which lead them to fairer predictions with respect to their teacher models, with a particular emphasis on the Web-based corpora used for training.

References

1. Basta, C., Costa-jussà, M.R., Casas, N.: Evaluating the underlying gender bias in contextualized word embeddings. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. pp. 33–39. Association for Computational Linguistics (Aug 2019). https://doi.org/10.18653/v1/W19-3805, https://www.aclweb.org/anthology/W19-3805
2. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the dangers of stochastic parrots: Can language models be too big? In: ACM Conference on Fairness, Accountability, and Transparency (FAccT) (2021), http://faculty.washington.edu/ebender/papers/Stochastic_Parrots.pdf
3. Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. p. 4356–4364. NIPS'16, Curran Associates Inc. (2016)
4. Cacoullos, R.T., Travis, C.E.: Bilingualism in the Community: Code-switching and Grammars in Contact. Cambridge University Press (2018)
5. Chada, R.: Gendered pronoun resolution using BERT and an extractive question answering formulation. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. pp. 126–133. Association for Computational Linguistics (Aug 2019). https://doi.org/10.18653/v1/W19-3819, https://www.aclweb.org/anthology/W19-3819
6. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates (1988)

7. Davidson, T., Bhattacharya, D., Weber, I.: Racial bias in hate speech and abusive language detection datasets. In: Proceedings of the Third Workshop on Abusive Language Online. pp. 25–35. Association for Computational Linguistics (Aug 2019). https://doi.org/10.18653/v1/W19-3504, https://www.aclweb.org/anthology/W19-3504
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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). pp. 4171–4186 (Jun 2019). https://doi.org/10.18653/v1/N19-1423, https://www.aclweb.org/anthology/N19-1423
9. Dictionary, C.E.: Cambridge english dictionary (2021), https://dictionary.cambridge.org/
10. Dorn, R.: Dialect-specific models for automatic speech recognition of African American Vernacular English. In: Proceedings of the Student Research Workshop Associated with RANLP 2019. pp. 16–20. INCOMA Ltd. (Sep 2019). https://doi.org/10.26615/issn.2603-2821.2019$_0$03, https://www.aclweb.org/anthology/R19-2003
11. Du Bois, J.W., Chafe, W.L., Meyer, C., Thompson, S.A., Martey, N.: Santa barbara corpus of spoken american english (2000), https://www.linguistics.ucsb.edu/research/santa-barbara-corpus/
12. Gokaslan, A., Cohen, V.: Openwebtext corpus (2019), http://web.archive.org/save/http://Skylion007.github.io/OpenWebTextCorpus.
13. Gorski, P.C.: Reaching and teaching students in poverty: Strategies for erasing the opportunity gap. Teachers College Press (2017)
14. Green, L.J.: Introduction, pp. 1–11. Cambridge University Press (2002). https://doi.org/10.1017/CBO9780511800306.005
15. Green, L.J.: Syntax part 1: verbal markers in AAE, p. 34–75. Cambridge University Press (2002). https://doi.org/10.1017/CBO9780511800306.005
16. Hassan, T.A., Hollander, S., van Lent, L., Tahoun, A.: Firm-level political risk: Measurement and effects. The Quarterly Journal of Economics **134**(4), 2135–2202 (2019)
17. Huang, J., Kuchaiev, O., O'Neill, P., Lavrukhin, V., Li, J., Flores, A., Kucsko, G., Ginsburg, B.: Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition. arXiv preprint arXiv:2005.04290 (2020)
18. Hutchinson, B., Prabhakaran, V., Denton, E., Webster, K., Zhong, Y., Denuyl, S.: Social biases in NLP models as barriers for persons with disabilities. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5491–5501. Association for Computational Linguistics (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.487, https://www.aclweb.org/anthology/2020.acl-main.487
19. Kendall, T., Farrington, C.: The corpus of regional african american language (2018), http://lingtools.uoregon.edu/coraal/
20. Kurita, K., Vyas, N., Pareek, A., Black, A.W., Tsvetkov, Y.: Measuring bias in contextualized word representations. In: Proceedings of the First Workshop on Gender Bias in Natural Language Processing. Association for Computational Linguistics (2019), https://www.aclweb.org/anthology/W19-3823
21. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. In: Proceedings of the 2020 International Conference on Learning Representations (2020), https://openreview.net/pdf?id=H1eA7AEtvS
22. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: 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. pp. 7871–7880. Association for Computational Linguistics (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.703, https://www.aclweb.org/anthology/2020.acl-main.703
23. Liang, P.P., Li, I.M., Zheng, E., Lim, Y.C., Salakhutdinov, R., Morency, L.P.: Towards debiasing sentence representations. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (2020), https://www.aclweb.org/anthology/2020.acl-main.488
24. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
25. Luu, A., Malamud, S.A.: Non-topical coherence in social talk: A call for dialogue model enrichment. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. pp. 118–133 (2020)
26. May, C., Wang, A., Bordia, S., Bowman, S.R., Rudinger, R.: 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). pp. 622–628. Association for Computational Linguistics (Jun 2019). https://doi.org/10.18653/v1/N19-1063, https://www.aclweb.org/anthology/N19-1063
27. Mubarak, H., Rashed, A., Darwish, K., Samih, Y., Abdelali, A.: Arabic offensive language on twitter: Analysis and experiments (2020)
28. Pullum, G.K.: African american vernacular english is not standard english with mistakes. The workings of language: From prescriptions to perspectives pp. 59–66 (1999)
29. QuickFacts, U.C.B.: United States census, QuickFacts statistics on U.S. population origin (2019), https://www.census.gov/quickfacts/fact/table/US/PST045219
30. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
31. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In: NeurIPS Energy Efficient Machine Learning and Cognitive Computing Workshop (2019)
32. Shah, D.S., Schwartz, H.A., Hovy, D.: Predictive biases in natural language processing models: A conceptual framework and overview. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5248–5264. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.468, https://www.aclweb.org/anthology/2020.acl-main.468
33. Sheng, E., Chang, K.W., Natarajan, P., Peng, N.: 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). pp. 3407–3412. Association for Computational Linguistics (Nov 2019). https://doi.org/10.18653/v1/D19-1339, https://www.aclweb.org/anthology/D19-1339
34. Utama, P.A., Moosavi, N.S., Gurevych, I.: Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance (2020)
35. Welch, B.L.: The generalization ofstudent's' problem when several different population variances are involved. Biometrika **34**(1/2), 28–35 (1947)
36. Wheeler, R., Thomas, J.: And "still" the children suffer: The dilemma of standard english, social justice, and social access. JAC pp. 363–396 (2013)
37. Xu, C., Zhou, W., Ge, T., Wei, F., Zhou, M.: BERT-of-theseus: Compressing BERT by progressive module replacing pp. 7859–7869 (Nov 2020), https://www.aclweb.org/anthology/2020.emnlp-main.633