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

Flavia Salutari, Jerome Ramos, Hosein A Rahmani, Leonardo Linguaglossa, Aldo Lipani

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

Flavia Salutari[1,2], Jerome Ramos[1], Hosein A. Rahmani[1], Leonardo Linguaglossa[2] and Aldo Lipani[1]
[1] University College London (United Kingdom), `first.last@ucl.ac.uk`
[2] Telecom Paris, LTCI, Institut Polytechnique de Paris (France), `last@telecom-paris.fr`

***Abstract*—In the last three years we witnessed the proliferation of innovative natural language processing (NLP) algorithms attempting at solving different tasks and designed for the most diverse applications. Despite groundbreaking transformer-based language models (LMs) have been proposed and widely adopted, 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 the bias of the predictions of these LMs on transcripts of both spoken African American English (AAE) and Standard American English (SAE). Our analysis reveals the presence of a bias towards SAE encoded by state-of-the-art LMs, like BERT and DistilBERT, a lower bias in distilled LMs and an opposite bias 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.**

I. Introduction[1]

Since their inception [10], transformers-based bidirectional encoder representations language models (LMs) gained lots of 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 the architecture [34], the size [35, 23, 42] and the training [26, 24]. This resulted in a growing concern of the research community to discuss the potential risks coming from the pervasive adoption of these models [3]. Indeed, several studies highlight that this would hinder an equitable and inclusive access to NLP technologies and have real-world negative consequences in different areas, as education, work and politics [36]. In this context, given the consistent emergence of new LMs trained on Web-based corpora, it is crucial to define to which extent such models are fair and not instead prone to bias.

Actually, given the sheer size and heterogeneity of the Web, one could expect these models to be bias-free. However, already before the explosion of transformer-based LMs, a variety of biases have been identified in standard word embeddings [4, 5]. Recently, some effort has been devoted to highlight the presence of possible biases encoded by transformer-based LMs along gender, race, ethnicity, and disability status. Yet, whereas the study of such biases is commonly tackled via sentiment analysis and named entity recognition tasks, in this paper we take a different approach. Inspired by the frequent scenario occurring in conversational systems, where a word could be unheard or unrecognized by the Automatic Speech Recognition system and would therefore need to be predicted, we measure how token predictions change based on their context.

In this work we focus on the study of potential bias towards English dialects spoken by underrepresented and historically discriminated groups, such as African American English (AAE). Particularly, AAE slightly differs from *mainstream* English, also known as Standard American English (SAE). In linguistics, these two variants are regarded as two different languages because highly structured with their own phonological, syntactic and morphological rules [15]. However, SAE speakers often believe that AAE is a version of SAE with mistakes and that AAE speakers belong to deficient cultures [32, 41]. While, instead, AAE highlights the regional, societal and cultural environments in which individuals have learned to speak [14].

It is difficult to estimate the number of AAE speakers, since some African Americans may speak a variety that aligns more with SAE and besides, not all AAE speakers

[1] This version is accepted for publication at PAKDD 2023.

are African Americans. Nevertheless, a 2019 census [33] estimates that approximately 13% of the U.S. population is currently African American. This suggests that the fraction of population speaking AAE could be large. Hence, the presence of potential linguistic biases would have discriminatory consequences towards a considerable group of individuals.

For these reasons, we set out to measure the robustness and the 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. This is a *fill-in-the-blank* task, where we mask and predict a token simulating its absence in every utterance. We next define two metrics to compare the likelihood that the model assigns to the predicted token and to the actual *masked* one, that we use as a proxy of quality and fairness for the model itself.

Specifically, 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 this bias to be reversed for RoBERTa and BART models. We additionally observe distilled variants of these LMs to be fairer with respect to their teachers. Finally, our analysis reveals how most of the bias resides in the AAE structural differences, and identifies the particles, the pronouns and the adpositions as principal parts of speech sources of bias.

II. Related Work

The success of transformer-based LMs is down to several factors, among which it is worth mentioning 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 [3]. These works have studied the effects of transferring the stereotypical associations present in the training datasets to LMs, which cause unintended bias towards underrepresented groups. A significant research effort has been done to show race and gender bias embedded in large models [43, 38, 2, 22, 7, 37, 28]. **(author?)** highlights the presence of topical biases in the words predicted by BERT on sentences mentioning disabilities.

In addition to bias measurement works, researchers have proposed methods to mitigate societal biases with debiasing techniques [25, 39, 20]. As for the bias towards languages, most studies have focused on offensive language and hate speech detection [29, 30, 9], while assessing the bias against dialects spoken by underrepresented groups is quite recent [11]. Whereas the above works mostly focus on the negative sentiment and stereotypical associations towards specific groups in BERT [10], in this work we quantify the linguistic bias towards AAE for 7 different LMs: BERT, RoBERTa [26], BART [24], DistilBERT and DistilRoBERTa [35], including both their cased and uncased versions.

These works have proven that the large dimension of the training datasets for state-of-the-art LMs is not synonymous of diversity and, as a consequence, of inclusion [3]. Therefore, in this regard, our analysis is essential to provide a framework to assess, reveal and counteract the existing biases, which we hope will contribute in enriching the scientific community knowledge on this domain.

III. 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 by the linguists because considered a fair representation of their spoken language. We note that, while this paper is not the first in studying 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 a 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.

A. Corpora for Spoken English

For SAE, we leverage the Santa Barbara Corpus of Spoken American English (SBCSAE) [12], which has been already widely adopted for different applications, as the assessment of political risk faced by U.S. firms [17], the measure of grammatical convergence in bilingual individuals [6] and the exploration of new-topic utterances in naturally occurring dialogues [27].

The SBCSAE is the only existing large-scale corpus of naturally occurring spoken interactions from people with different regional origins in USA. It includes conversations from a wide variety of people, differing in gender, occupation and social background, recorded in various real everyday life situations. All the audio recordings are complemented with their transcribed counterparts, which are the ones we use in this work.

The fact that SBCSAE consists of speakers from several regional origins prevents us from crafting the results and unintentionally inducing a bias by comparing

Corpus	Language	$\lvert\mathcal{U}\rvert$	$\langle\ell_u\rangle$	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
Preprocessed					
CORAAL	AAE	63,814	8.23	525,067	16,352
SBCSAE	SAE	25,113	8.38	210,430	10,540

TABLE I

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).

Model	**Training Data**
BERT, DistilBERT	BOOKSCORPUS and English Wikipedia (16GB)
RoBERTa, BART	BERT data + CC-NEWS, OPENWEBTEXT and STORIES (160GB)
DistilRoBERTa	OPENWEBTEXT (38GB)

TABLE II

TRAINING DATA FOR THE TESTED LMS.

AAE with an *academic* version of SAE, which is instead rather different from the commonly spoken English and, hence, far from the purpose of this work. Therefore, we filter out Hispanic and African American speakers (1092 AAE utterances, a negligible number *w.r.t.* to the size of the corpus) and obtain a corpus of SAE language.

For AAE, we leverage the Corpus of Regional African American Language (CORAAL) [21], which also provides the audio recordings along with their time-aligned orthographic transcription, 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 literature for a plethora of tasks, ranging from dialect specific speech recognition [11] to cross-language transfer learning [18].

In this work, we only focus on the CORAAL:DCB portion, since it is the one comprising the most recent interviews (carried out between 2015 and 2017) and 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. Particularly, we discard both short utterances (composed by just one or two words) and very long ones (greater than 50 words). Therefore, we only keep utterances having a number of words ranging from 3 to 50.

In Tab. I we report a terse 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 number of utterances $|\mathcal{U}|$, but also in terms of total number of words L and terms $|\mathcal{T}|$, we can see that, after the filtering, the average utterance length $\langle\ell_u\rangle$ is very similar ($\sim$ 8 words per utterance).

B. Bias in Masked Language Modeling

In order to measure the bias in LMs we perform a MLM task. We leverage the transformer-based BERT$_{\text{base}}$ LM [10] and its recent variants, including DistilBERT$_{\text{base}}$ [35], in both their cased and uncased flavors, RoBERTa$_{\text{base}}$ [26], DistilRoBERTa$_{\text{base}}$ and BART$_{\text{base}}$ [24]. These LMs have all been pre-trained using a MLM objective, which consists in randomly masking 15% of the tokens using a special `[MASK]` token. Note that these models are trained on different corpora, summarized in Tab. II.

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. Specifically, we encode each utterance of the two corpora with the *tokenizer* of the LM considered, then, in turn, we mask each word w_{mask} and finally predict it by feeding the model with only a context of 10 tokens surrounding the masked one w_{mask}. Tab. III shows an example, illustrating how the experiment is carried on: (i) we let the LM encode the original utterance $\mathbf{u}$ (the one reported in the table has a length lower than 10 tokens so there is no need for the window), (ii) we mask and predict the first token w_1, (iii) we iteratively repeat this process until the last token of the utterance is masked.

The LM provides for each run a list of possible terms to *fill-in-the-blank*. In this vocabulary set ($\mathcal{T}$) we select the predicted term t_p having the highest probability $P(t_p|c)$ and, as such, ranking first in the list $\rho(t_p|c) = 1$, where c is the context surrounding t_p and ρ is the rank of $t|c$ provided by the model. In this notation, a word w is a term t in a context c $(t|c)$. We next retrieve from the vocabulary of possible terms $\mathcal{T}$ the corresponding probability $P(t_m|c)$ and the rank $\rho(t_m|c)$ for the actual masked token t_m. 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}. It is then natural to employ the probabilities difference $\Delta P(t|c)$ as a proxy of the quality of the prediction for a single token, so

original utterance (u)	And I be okay with it .
u with $\mathbf{w_1}$ masked	`[MASK]` I be okay with it .
u with $\mathbf{w_2}$ masked	And `[MASK]` be okay with it .
	...
u with $\mathbf{w_7}$ masked	And I be okay with it `[MASK]`

TABLE III
EXAMPLE SHOWING THE MASKED TOKEN EXPERIMENT.

defined:

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

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, as the $\Delta P(t|c)$ alone could be insufficient. This because, the $\Delta P(t|c)$ strongly depends on how the LM assigns the probability. Indeed, the probability distribution of $P(t|c)$ could be more uniform, and consequently would lead, on average, to a smaller $\Delta P(t|c)$, or more skewed, causing instead larger differences $\Delta P(t|c)$. Instead, this effect is not present in CRR that remains unaffected by such differences in the output probability distribution of $P(t|c)$.

IV. Results & Discussion

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

A. Measuring the Bias of LMs

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

Tab. IV reports an overview of the results of $\Delta P(u)$ and CRR(u). After having assessed that the difference between the means of AAE and SAE for both $\Delta P(u)$ and CRR(u) with a Welch's t-test [40] is significant (p-value < 0.05), we measure their effect size using the Cohen's d [8]. This is reported in the last two columns of Tab. IV. According to Cohen's classification there is a *small* effect for both the metrics, and a *medium* effect for BART on $\Delta P(u)$ (d>0.5).

We summarize the quality of the prediction in the corpora by means of two error measures. We report the Mean Absolute Error (MAE) for each of the two distributions:

$$\text{MAE}(\Delta P(u)) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} |\Delta P(u)|, \tag{5}$$

$$\text{MAE}(\text{CRR}(u)) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} |\text{CRR}(u)|. \tag{6}$$

We also report the Mean Squared Error (MSE), defined as:

$$\text{MSE}(\Delta P(u)) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \Delta P(u)^2, \tag{7}$$

$$\text{MSE}(\text{CRR}(u)) = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \text{CRR}(u)^2. \tag{8}$$

Indeed, these error measures can be used to quantify the quality of the predicted terms. MAE and MSE closer to 0 correspond to an utterance having more accurately predicted terms. Therefore, in Tab. IV we highlight the values leading to the smallest error between AAE and SAE. We additionally emphasize the presence of bias by pointing out the percentage of bias change of each LM $\mathbf{\Delta}[\%]$. This is always calculated with respect to the model with the largest bias, and when positive the model is biased towards SAE, *vice versa* otherwise.

Three main patterns clearly emerge from Tab. IV. 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 it is the LM which leads to the highest relative bias. Specifically, notice that the MAE$(\Delta P(u))$ for SAE is more than 20% lower than

	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

TABLE IV

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 THE WELCH'S TWO-TAILED T-TEST (P-VALUE < 0.05). THE COLUMN D CONTAINS THEIR EFFECT SIZE COMPUTED ACCORDING TO THE COHEN'S D.

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 flavors, and DistilRoBERTa, are the models which perform better as regards 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% of its performance [35]. On the other hand, this is in line with recent work [3] reporting that such LMs sometimes exceed the performance of the original ones. However, as mentioned in Sec. III, it is crucial to also look at the CRR(u), since a better behavior in terms of $\Delta P(u)$ could in practice just 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 leading to a decent quality of the prediction 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 reverse unexpected bias towards AAE is also introduced by RoBERTa and DistilRoBERTa. This is somewhat surprising and could probably be ascribable to the type of datasets they have been trained on. Indeed, as shown in Tab. II, RoBERTa and BART are pre-trained with 1000% more data than BERT. Particularly, by delving into the type of data involved, we discover multiple sources, ranging from English language encyclopedia and literary works (same as BERT), to news articles and Web content. Specifically, RoBERTa, BART and DistilRoBERTa leverage OPENWEBTEXT [13], a corpus which includes filtered Web content obtained by scraping the social media platform Reddit, possibly exposing the LMs to a less *standard* American English.

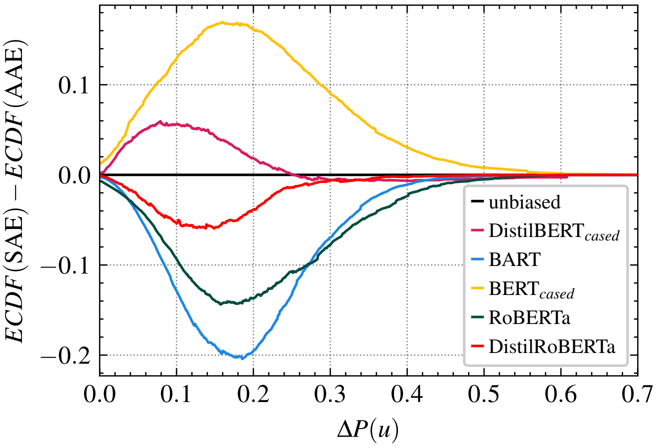

Fig. 1. The difference between the ECDFs of SAE and AAE for the $\Delta P(u)$ measure. When the values are greater than zero the LMs are more biased towards SAE, *vice versa* otherwise.

Since Tab. IV reports only a summary of the distributions of the bias metrics computed on both the datasets, for a better understanding, we show in Fig. 1 the bias measured by subtracting the empirical cumulative distribution functions (ECDFs) of $\Delta P(u)$ of AAE to that of SAE. This figure includes the bias measured for the LMs, reporting, for the sake of simplicity, for BERT and DistilBERT only their *cased* variants. The solid black line at $y = 0$ shows the optimal unbiased LM and, hence, visually separates what is biased against AAE (on the positive *y*-axis) from what instead is biased against SAE (on the negative *y*-axis). In this way, we clearly see the behaviors of the LMs leading to the two worst biases, *i.e.*, RoBERTa and BERT$_{cased}$: they are consistently biased towards one side (BERT$_{cased}$ is always positive, whilst RoBERTa is instead always negative). They both present the maximum bias when $\Delta P(u)$ is close to 0.2 and instead mitigate for larger values. A similar behavior is observed for the CRR(u) (available in the Appendix).

Original	**Translated**
Double Negative (0.7%)	
• *You don't need nothing but you.* • *I wasn't no lifeguard cause I couldn't swim.* • *Don't never try to chase another person happiness.* • *I don't know nobody over there no more.*	• *You don't need anything but you.* • *I wasn't a lifeguard because I couldn't swim.* • *Never try to chase another person's happiness.* • *I don't know anyone over there anymore.*
Copula *be* (2.8%)	
• *And I be okay with it.* • *It depends on where you going to.* • *All of my friends was from like DC.* • *Okay, we having a baby.*	• *And I am okay with it.* • *It depends on where you are going to.* • *All of my friends were from DC.* • *Okay, we are having a baby.*
Contractions (4.6%)	
• *I'm'a ask you.* • *I ain't coming home.* • *something gonna happen.* • *you gonna be there for a couple of hours.*	• *I'm going to ask you.* • *I'm not coming home.* • *something is going to happen.* • *you will be there for a couple of hours.*

TABLE V

A SAMPLE OF AAE UTTERANCES SELECTED BASED ON THEIR SYNTACTICAL FEATURES AND THEIR TRANSLATIONS TO SAE. IN BRACKETS THE PREVALENCE OF THE FEATURE OVER THE UTTERANCES IN THE AAE CORPUS.

B. *Bias on AAE Features*

We next investigate how results change when we acknowledge the lexical, syntactical, morphological and also phonological rules of AAE. Following AAE grammar [16], we choose to focus on three major syntactical features: (i) the use of *double* negatives, (ii) the different usage of copula *be* and, finally, (iii) the contractions of words and groups of words.

As for (i), we search for the close presence of multiple forms of grammatical negation (which in Standard English are instead understood to resolve to a positive) in all the utterances of the AAE corpus, and find, that 0.7% of the utterances contains such a feature. Concerning (ii), we select the AAE utterances exhibiting the use of the *aspectual be* verb, typically used to denote habitual or iterative meaning (*e.g.*, *I be okay with it* in Tab. V). Additionally, we also filter on utterances with the verb tense in the *-ing* form where the copula is either omitted (*e.g.*, *It depends on where you going to* in Tab. V) or left at the base form (*e.g.*, *they be getting mad* in Tab. V), for a total of 2.8% of utterances. Finally, for (iii) we include those utterances containing not-standard contractions, *e.g.*, *I'm'a, ain't* or omitting the auxiliary before *gonna*, *e.g.*, *something gonna happen* in Tab. V. 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 having properly filtered the utterances corresponding to the specific grammar patterns, we carefully manually validate our selection, by random picking and inspecting 1% of them. We check that the 1% random sampled utterances are actually satisfying the criteria we were looking for. From this manual labeling we double check our selection strategies based on syntactical rules and find that for both the 3 cases these are 99% accurate.

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, that we define as AAE$^{\intercal}$. We keep the translation process as neutral as possible, by preserving the standard officially recognized contractions and by only *adjusting* the selected grammar rules. Tab. V 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 III, on these 150 translated utterances AAE$^{\intercal}$ and measure the bias. We report the results in Tab. VI. 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 *copula* class, where it is *large*.

At a first glance, we observe that the errors for the set of the AAE utterances in the *copula* class are larger than both the other two classes and the whole AAE corpus (reported in Tab. IV). More in general, we observe that, on average, both the three classes, and therefore, all the 150 AAE utterances, come with a less accurate average prediction with respect to the overall AAE corpus. We observe instead that the translated utterances AAE$^{\intercal}$ are better predicted with respect to AAE surprisingly for all the seven LMs.

Notably, we observe that for the translated utterances in the *double negative* class, the four metrics are always smaller (and hence sign of better performance) than those measured for the SAE corpus. This is somewhat unexpected since we observed for RoBERTa and BART an opposite bias on SAE. However, we remind 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 bias and consequently *"cleaner"* than those found in the SAE corpus.

C. *Bias on Part-of-Speech*

Finally, we investigate to which extent the POS tags are tied to the measured bias towards AAE or SAE. To produce these results, we preliminary tag the tokens independently generated by each language model with the NLTK [1] POS-tagger. Next, we group by the 12 main tags of the universal tagset [31] and compute the MAE and the MSE on the term-level measurements $\Delta P(t)$ and CRR(t).

	MAE								MSE							
	$\Delta P(u)$				CRR(u)				$\Delta P(u)$				CRR(u)			
Model	**AAE**	**AAE$^\mathsf{T}$**	**Δ**[%]	**d**	**AAE**	**AAE$^\mathsf{T}$**	**Δ**[%]	**d**	**AAE**	**AAE$^\mathsf{T}$**	**Δ**[%]	**d**	**AAE**	**AAE$^\mathsf{T}$**	**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
								Copula *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

TABLE VI

SIMILAR TO TABLE. IV BUT CALCULATED OVER A SAMPLE OF 50 UTTERANCES OF AAE AND THEIR TRANSLATED VERSION (AAE$^\mathsf{T}$) FOR EACH FEATURE OF AAE.

Indeed, rather than averaging across the tokens in one utterance, we consider all the terms t belonging to a given POS tag. Tab. VII reports the results obtained for the top-3 POS featuring the highest cumulative bias, computed by summing the absolute bias $|\mathbf{\Delta}[\%]|$ introduced by each LM and measured with the MAE(CRR(t)): the particles (*e.g.*, *to, up, out, etc.*), the pronouns (*e.g.*, *you, it, my, etc.*) and the adpositions (*e.g.*, *like, of, with, etc.*). The results for the rest of the POS are available in the Appendix. In order to trust the results of the POS-tagger we manually check the correctness of 100 tokens for each class and language. We find that the accuracy is 100% for the *pronouns*, 99% for the *adpositions* and 92% for the *particles*. Also in this case, we measure the effect for both the metrics, and find that, according to the 6-grade Cohen's classification scale, it is *very small*.

Interestingly, for the *particles* class, one can notice the same pattern reported in Tab. IV. Particularly, DistilBERT$_{cased}$ is the LM which performs better in terms of $\Delta P(t)$ and, DistilRoBERTa the one that leads to the lowest bias. Conversely, BERT is the model that shows the highest bias towards SAE: it is up to 29% more accurate with respect to AAE for MAE$(\Delta P(t))$. BART presents the opposite largest bias in favor of AAE: 23% (18%) more for the MAE of $\Delta P(t)$ (CRR(t)) on the *particles* class. It is also interesting to note that DistilBERT also at a token-level analysis presents better values for $\Delta P(t)$ rather than CRR(t).

Quite surprisingly, we discover a bias presented by all the tested LMs towards AAE in the *pronouns* class. This holds for both the $\Delta P(t)$ and the CRR(t) and is revealed with both the error measures, with the exception of BERT and DistilBERT *cased* for the MSE of $\Delta P(t)$. This result deserves further investigation.

V. CONCLUSION

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

In a nutshell, results presented in this paper suggest that different models embed diverse biases. Particularly, the most popular state-of-the-art LMs, namely BERT and DistilBERT, show a non-negligible bias towards SAE

	MAE								MSE							
	$\Delta P(t)$				CRR(t)				$\Delta P(t)$				CRR(t)			
Model	**AAE**	**SAE**	**Δ**[%]	**d**	**AAE**	**SAE**	**Δ**[%]	**d**	**AAE**	**SAE**	**Δ**[%]	**d**	**AAE**	**SAE**	**Δ**[%]	**d**
Particles [16k (AAE) 6k (SAE) terms], $\sum\lvert\mathbf{\Delta}[\%]\rvert = 66$																
BERT$_{\text{cased}}$	0.113	**0.081**	29 †	0.143	0.212	**0.192**	9 †	0.054	0.071	**0.040**	44 †	0.176	0.177	**0.157**	11 †	0.062
BERT$_{\text{uncased}}$	0.126	**0.090**	29 †	0.145	0.212	**0.191**	10 †	0.059	0.086	**0.049**	44 †	0.182	0.176	**0.155**	12 †	0.066
DistilBERT$_{\text{cased}}$	0.079	**0.070**	12 †	0.062	0.328	**0.310**	6 †	0.045	0.030	**0.025**	16 †	0.051	0.272	**0.259**	5 †	0.036
DistilBERT$_{\text{uncased}}$	0.090	**0.073**	19 †	0.099	0.313	**0.294**	6 †	0.046	0.040	**0.028**	29 †	0.102	0.262	**0.248**	5 †	0.038
RoBERTa	**0.101**	0.114	-11 †	0.058	**0.208**	0.238	-13 †	0.083	**0.056**	0.062	-10 †	0.039	**0.168**	0.194	-13 †	0.081
DistilRoBERTa	**0.099**	0.108	-8 †	0.049	**0.354**	0.369	-4 †	0.037	**0.043**	0.047	-9 †	0.035	**0.300**	0.312	-4 †	0.032
BART	**0.079**	0.102	-23 †	0.139	**0.261**	0.317	-18 †	0.144	**0.032**	0.043	-26 †	0.107	**0.212**	0.261	-19 †	0.142
Pronouns [84k (AAE) 31k (SAE) terms], $\sum\lvert\mathbf{\Delta}[\%]\rvert = 59$																
BERT$_{\text{cased}}$	**0.182**	0.186	-2 †	0.017	**0.349**	0.379	-8 †	0.078	0.101	**0.098**	3 †	0.017	**0.268**	0.288	-7 †	0.062
BERT$_{\text{uncased}}$	**0.186**	0.203	-8 †	0.061	**0.326**	0.367	-11 †	0.110	**0.110**	0.116	-5 †	0.027	**0.246**	0.278	-12 †	0.100
DistilBERT$_{\text{cased}}$	**0.139**	0.141	-1	0.011	**0.554**	0.592	-6 †	0.103	0.051	**0.049**	4 †	0.018	**0.447**	0.480	-7 †	0.094
DistilBERT$_{\text{uncased}}$	**0.090**	0.104	-14 †	0.086	**0.404**	0.453	-11 †	0.124	**0.034**	0.039	-12 †	0.041	**0.319**	0.361	-12 †	0.117
RoBERTa	**0.176**	0.187	-6 †	0.045	**0.351**	0.368	-5 †	0.044	**0.096**	0.102	-7 †	0.034	**0.271**	0.284	-5 †	0.039
DistilRoBERTa	**0.116**	0.123	-5 †	0.036	**0.466**	0.481	-3 †	0.037	**0.047**	0.051	-7 †	0.030	**0.382**	0.393	-3 †	0.028
BART	**0.124**	0.166	-25 †	0.233	**0.444**	0.520	-15 †	0.188	**0.046**	0.067	-32 †	0.188	**0.362**	0.428	-16 †	0.178
Adpositions (prepositions and postpositions) [50k (AAE) 18k (SAE) terms], $\sum\lvert\mathbf{\Delta}[\%]\rvert = 55$																
BERT$_{\text{cased}}$	0.227	**0.199**	13 †	0.105	0.507	**0.447**	12 †	0.140	0.129	**0.105**	18 †	0.108	0.442	**0.380**	14 †	0.153
BERT$_{\text{uncased}}$	0.251	**0.222**	11 †	0.097	0.499	**0.447**	11 †	0.122	0.153	**0.127**	17 †	0.107	0.435	**0.381**	13 †	0.134
DistilBERT$_{\text{cased}}$	**0.103**	0.104	-0.3	0.002	0.779	**0.753**	3 †	0.073	0.034	**0.033**	4 †	0.012	0.730	**0.7**	3 †	0.065
DistilBERT$_{\text{uncased}}$	0.140	**0.135**	4 †	0.029	0.598	**0.562**	6 †	0.084	0.057	**0.053**	8 †	0.032	0.532	**0.493**	7 †	0.095
RoBERTa	0.199	**0.195**	2	0.014	0.447	**0.408**	9 †	0.090	0.108	**0.108**	0.4	0.002	0.385	**0.344**	10 †	0.100
DistilRoBERTa	**0.139**	0.143	-3 †	0.022	0.584	**0.542**	7 †	0.099	**0.057**	0.060	-6 †	0.026	0.523	**0.474**	9 †	0.118
BART	0.154	**0.154**	0.02	0.000	0.552	**0.525**	5 †	0.063	**0.062**	0.063	-2	0.008	0.485	**0.455**	6 †	0.072

TABLE VII

SIMILAR TO TABLE. IV BUT CALCULATED FOR t RATHER THAN u, FOR THREE POS CLASSES.

(quality of the predictions up to 21% more accurate than AAE). Instead, BART, RoBERTa and DistilRoBERTa exhibit an opposite bias. Our experiments reveal also that the distilled variants of BERT and RoBERTa are the fairest among the seven tested LMs.

Yet, despite this paper provides a first insightful snapshot of linguistic bias embedded in different LMs, it opens 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 on how the bias could be consistently mitigated with more inclusive corpora, by taking into account AAE features. Finally, a special care could be put in 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] Nltk: The natural language toolkit. ETMTNLP '02, page 63–70. Association for Computational Linguistics, 2002.

[2] Christine Basta, Marta R. Costa-jussà, and Noe Casas. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39. Association for Computational Linguistics, August 2019.

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[4] Su Lin Blodgett and Brendan O'Connor. Racial disparity in natural language processing: A case study of social media african-american english. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017)*, 2017.

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to home-

maker? debiasing word embeddings. NIPS'16, page 4356–4364. Curran Associates Inc., 2016.

[6] Rena Torres Cacoullos and Catherine E Travis. *Bilingualism in the Community: Code-switching and Grammars in Contact*. Cambridge University Press, 2018.

[7] Rakesh Chada. Gendered pronoun resolution using BERT and an extractive question answering formulation. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 126–133. Association for Computational Linguistics, August 2019.

[8] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 1988.

[9] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35. Association for Computational Linguistics, August 2019.

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

[11] Rachel Dorn. Dialect-specific models for automatic speech recognition of African American Vernacular English. In *Proceedings of the Student Research Workshop Associated with RANLP 2019*, pages 16–20. INCOMA Ltd., September 2019.

[12] John W Du Bois, Wallace L Chafe, Charles Meyer, Sandra A Thompson, and Nii Martey. Santa barbara corpus of spoken american english, 2000.

[13] Aaron Gokaslan and Vanya Cohen. Openwebtext corpus, 2019.

[14] Paul C Gorski. *Reaching and teaching students in poverty: Strategies for erasing the opportunity gap*. Teachers College Press, 2017.

[15] Lisa J. Green. *Introduction*, pages 1–11. Cambridge University Press, 2002.

[16] Lisa J. Green. *Syntax part 1: verbal markers in AAE*, page 34–75. Cambridge University Press, 2002.

[17] Tarek A Hassan, Stephan Hollander, Laurence van Lent, and Ahmed Tahoun. Firm-level political risk: Measurement and effects. *The Quarterly Journal of Economics*, 134(4):2135–2202, 2019.

[18] Jocelyn Huang, Oleksii Kuchaiev, Patrick O'Neill, Vitaly Lavrukhin, Jason Li, Adriana Flores, Georg Kucsko, and Boris Ginsburg. Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition. *arXiv preprint arXiv:2005.04290*, 2020.

[19] Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in NLP models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501. Association for Computational Linguistics, July 2020.

[20] Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716. Association for Computational Linguistics, July 2020.

[21] Tyler Kendall and Charlie Farrington. The corpus of regional african american language, 2018.

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

[23] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of the 2020 International Conference on Learning Representations*, 2020.

[24] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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. Association for Computational Linguistics, July 2020.

[25] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[26] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv*

preprint arXiv:1907.11692, 2019.

[27] Alex Luu and Sophia A Malamud. 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*, pages 118–133, 2020.

[28] Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 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. Association for Computational Linguistics, June 2019.

[29] Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. Hate speech detection and racial bias mitigation in social media based on bert model. volume 15, pages 1–26. Public Library of Science, 08 2020.

[30] Hamdy Mubarak, Ammar Rashed, Kareem Darwish, Younes Samih, and Ahmed Abdelali. Arabic offensive language on twitter: Analysis and experiments, 2020.

[31] Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2089–2096. European Language Resources Association (ELRA), May 2012.

[32] Geoffrey K Pullum. African american vernacular english is not standard english with mistakes. *The workings of language: From prescriptions to perspectives*, pages 59–66, 1999.

[33] U.S. Census Bureau QuickFacts. United States census, QuickFacts statistics on U.S. population origin, 2019.

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

[35] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS Energy Efficient Machine Learning and Cognitive Computing Workshop*, 2019.

[36] Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 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*, pages 5248–5264. Association for Computational Linguistics, 2020.

[37] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 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. Association for Computational Linguistics, November 2019.

[38] Yi Chern Tan and L. Elisa Celis. Assessing social and intersectional biases in contextualized word representations. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13209–13220, 2019.

[39] Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. Mind the trade-off: Debiasing nlu models without degrading the in-distribution performance, 2020.

[40] Bernard L Welch. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35, 1947.

[41] Rebecca Wheeler and Julia Thomas. And ”still” the children suffer: The dilemma of standard english, social justice, and social access. *JAC*, pages 363–396, 2013.

[42] Canwen Xu, Wangchunshu Zhou, Tao Ge, Furu Wei, and Ming Zhou. BERT-of-theseus: Compressing BERT by progressive module replacing. pages 7859–7869, November 2020.

[43] Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. Hurtful words: Quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, page 110–120. Association for Computing Machinery, 2020.