Evaluating and Mitigating Inherent Linguistic Bias of African American English through Inference

Jamell Dacon∗ Haochen Liu Jiliang Tang
Michigan State University
{daconjam, liuhoac1, tangjili}@msu.edu

Abstract

Recent studies show that NLP models trained on standard English texts tend to produce biased outcomes against underrepresented English varieties. In this work, we conduct a pioneering study of the English variety use of African American English (AAE) in NLI task. First, we propose CODESWITCH, a greedy unidirectional morphosyntactically-informed rule-based translation method for data augmentation. Next, we use CODESWITCH to present a preliminary study to determine if demographic language features do in fact influence models to produce false predictions. Then, we conduct experiments on two popular datasets and propose two simple, yet effective and generalizable debiasing methods. Our findings show that NLI models (e.g. BERT) trained under our proposed frameworks outperform traditional large language models while maintaining or even improving the prediction performance. In addition, we intend to release CODESWITCH, in hopes of promoting dialectal language diversity in training data to both reduce the discriminatory societal impacts and improve model robustness of downstream NLP tasks.

1 Introduction

In recent years, social media has become a pivotal tool its users to express their thoughts, feelings, and opinions on similar interests (Dacon and Tang, 2021). Typically, Standard American English (SAE), a high-resource language (HRL) is often used in formal communication, whereas African American English (AAE) is primarily spoken in the United States and is often heavily and explicitly used on social media platforms such as Twitter (Field et al., 2021; Blodgett et al., 2020).

In particular, AAE is an English language variety and can be considered to be a low-resource language (LRL) that is neither spoken by all African Americans or individuals who identify as BIPOC (Black, Indigenous, or People of Color), nor is it spoken only by African Americans or BIPOC individuals (Field et al., 2021; Dacon, 2022; Bland-Stewart, 2005). However, most dominant AAE speakers reside in diglossic communities and are able to code-switch, speaking both SAE and AAE. In linguistics, code-switching also referred to as language alternation is the ability of a speaker to alternate between two or more languages or language varieties within a particular conversation (Young and Barrett, 2018; Gardner-Chloros et al., 2009; DeBose, 1992; Young, 2009; Dacon, 2022). Thus, we refer to code-switching as switching among dialects, and/or language styles. For example, bi-dialectal AAE speakers are often able to code-switch between the SAE and both phonological and morphological language features of AAE while maintaining contextual intent.

Natural Language Understanding (NLU) is a subset of NLP, which enables human-computer interaction (HCI) by attempting to understand human language data such as text or speech, and communicate back to humans in their respective languages such as English, Spanish, etc., (Schank, 1972). Hence, we will focus on inference, which is an eminent area of study of NLU. In particular, Natural language inference (NLI), a subset of NLU, also known as Recognizing Textual Entailment (RTE) is a segment-level categorization task of understanding the inferential relationships between sentence pairs and anticipating whether they are entailing, contradictory, or neutral sentences (Bowman et al., 2015; Williams et al., 2018).

Generally, the term implicit bias is used to refer
to the unconscious preferential behaviors towards a certain demographic group such as age, race, ethnicity, gender, etc. (Liu et al., 2021; Tan et al., 2020a; Ribeiro et al., 2018). However, in this study, to examine the differences in language styles from different demographic groups, we refer to this type of predisposed language style bias as inherent linguistic bias. Although, both biases are very similar, there exists a subtle difference as linguistic bias specifically refers to an analysis of every aspect of a particular language (Zhou and Bansal, 2020). The existence of these biases in large language models (LLMs) such as mask language models (MLMs) generate language bias leading to potential harmful societal impacts inconveniencing members of LRL and diglossic communities who speak both standard languages and unrepresented dialects. This may increase feelings of marginalization and disenfranchisement (Liu et al., 2020a; Blodgett et al., 2020; Field et al., 2021).

Hence, in this work, we conduct a pioneering study of robustifying MLMs to minimize false predictions by introducing dialectal language diversity in training data to determine if MLMs learn to make predictions based on demographic language features, and proposing two debias methods to enhance NLI models to mitigate the presence of linguistic bias during the training process. We posit that it is vital for production-ready MLMs improve their robustness to produce minimal systemic biases against protected attributes such as race and gender and thus, reducing discriminatory societal impacts (Hovy and Spruit, 2016; Sharma et al., 2021; Liu et al., 2020a; Tan et al., 2020a).

Specifically, we aim to answer two research questions: (1) How can we as NLP practitioners encourage dialectal language diversity in training data?; (2) Do pretrained MLMs make predictions based on demographic language features?; and (3) How can we measure fairness and mitigate such biases in order to ensure fairness in NLU.

Our contributions include:

- **CODESWITCH**, a greedy unidirectional morphosyntactically-informed rule-based translation method for data augmentation to generate intent-and-semantically equivalent AAE examples by perturbing SAE examples.
- Two intent-and-semantically equivalent NLI dataset of AAE sentence pairs with a wide range of morphological syntactic features and dialect-specific vocabulary.
- A detailed human evaluation of our human annotators to ensure contextual accuracy of adversarial sentence pairs (see Appendix D for details).
- Two simple, yet effective debiasing methods to mitigate the inherent linguistic bias in NLI models, while maintaining or even improving their prediction performance.

## 2 Preliminaries

In this section, we introduce some preliminary knowledge about the problem under study. We first present the problem statement, and then describe two popular NLI datasets used in our research.

### 2.1 Problem Statement

We aim to investigate sentence representations of two linguistic systems of different demographic groups to demonstrate the existence of constitutional linguistic bias. To address the above research questions, we define two goals:

1. The first goal is to predict inferential relationships between paired sentences i.e., the second sentence is an entailment, contradiction, or neutral with respect to the first sentence.

2. The second goal is to debias the sentence representations obtained from the words in the given sentence. Specifically, we want the sentence representation to only include the semantic information, but not the language style, whether SAE or AAE. Therefore, we want the MLM to ignore the language style of each demographic group in order to make fair predictions.

Mitigating such linguistic biases can help develop robust MLMs for LRLs and dialectal languages more easily. Our main objective is to focus on dialectal language inclusivity, while using the benefit of large pretrained MLMs in order to improve model robustness of downstream tasks of NLP technologies for LRLs and language varieties.

### 2.2 Dataset

In this subsection, we introduce two of the largest, most popular NLP datasets for textual inference, namely, the Stanford Natural Language Inference (SNLI) and Multi-Genre Natural Language Inference (MNLI) corpora.
### 2.2.1 SNLI corpus

The SNLI (Bowman et al., 2015) corpus is constructed from the Flickr30k corpus (Young et al., 2014). The original image caption is classified as the premise, whereas, the hypothesis is a human-written premise-related sentence that must satisfy one of one of three relational conditions: (1) **Entailment** – true image description, (2) **Neutral** – neutral image description, and (3) **Contradiction** – false or random image description. The SNLI corpus is a collection of 570K premise-hypothesis sentence pairs, where each pair is aligned with one of these three relational labels.

### 2.2.2 MNLI corpus

Similarly to SNLI, the MNLI corpus (Williams et al., 2018) is a closely related crowd-sourced collection of 433k sentence pairs and their relational labels. However, MNLI contains 10 distinct genre categories (i.e., Letters, Verbatim, Fiction, Face-to-face, Travel, Telephone, Travel, Oxford University Press, Slate, 9/11, and Government) written and spoken data instead of image caption data.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Premise</th>
<th>Hypothesis</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNLI</td>
<td>A land rover is being driven across a river. Children smiling and waving at camera An older man is drinking orange juice at a restaurant.</td>
<td>A vehicle is crossing a river. They are smiling at their parents Two women are at a restaurant drinking wine.</td>
<td>entailment neutral contradiction</td>
</tr>
<tr>
<td>MNLI</td>
<td>So I have to find a way to supplement that I don’t know um do you do a lot of camping The new rights are nice enough</td>
<td>I need a way to add something extra. Everyone really likes the newest benefits I know exactly.</td>
<td>entailment neutral contradiction</td>
</tr>
</tbody>
</table>

Table 1: Randomly chosen original SNLI and MNLI examples and their inferential relationships.

### 3 CODESWITCH Creation

In this section, we first describe the process of the creation of CODESWITCH, carried out in three steps: 1) data collection of morphological syntactic features and dialect-specific vocabulary, 2) candidate retrieval of simple, deterministic morphosyntactic substitutions for unidirectional translations, and 3) human evaluation to test contextual accuracy of perturbations generated by CODESWITCH.

#### 3.1 Data Collection

First, to gain an better understanding of AAE language, we engage with literature, sample text examples and mass collect morpho-syntax rules (which we adapt from the literature) (see Appendix B) (Bailey et al., 1998; Green, 2002; Bland-Stewart, 2005; Dacon, 2022; Blodgett et al., 2020; Stewart, 2014; Blodgett et al., 2016; Elazar and Goldberg, 2018). Therefore, we attempt a proactive approach in data-collection of grammatical, structural and syntactic rules of word case usage of AAE language features to understand the application of AAE in NLP downstream tasks. Next, we employ and assist 6 trained sociolinguist Amazon Mechanical Turk (AMT) workers with our collected set rules and text examples.

**Pairwise Sample Collection** We first randomly sample \( n = 5000 \) SAE premise-hypothesis sentence pairs that contain at least 8 words from both SNLI and MNLI corpora for a total of 10,000 sentence pairs. For contextual accuracy, we task the first 3 workers to obtain the AAE equivalents of our SAE samples (see Table 1), where each annotator is tasked to translate each SAE sentence pair into AAE. The full annotation guidelines can be seen in Appendix C.

#### 3.2 Candidate Retrieval

Starting from data collection, we next retrieve candidate phrases and words use cases for data augmentation from our obtained AAE equivalent sentence pairs. As Liu et al. (2021) uses a deep text classification model to illustrate that demographic language features do in fact influence models to produce false predictions on semantically equivalent SAE and AAE texts, our protocol follows simple, deterministic substitutions of English texts by dialect-specific vocabulary. To do so, we make use of both SAE and AAE sentence pairs in a pairwise fashion and construct a unidirectional informed-based translatative morpho-syntax protocol (TMsP) that enables CODESWITCH to convert any given SAE text to a text possessing adequate language features to be considered as AAE from a dominant AAE speaker. More details on TMsP can be found in Appendix B).

---

2 Each AMT worker is independent and a trained sociolinguist filtered by HIT approval rate \( \geq 96\% \), completed \( > 10,000 \) HITs and location (within the United States)
Table 2: Augmented SNLI and MNLI examples (from Table 1) following the application of CODESWITCH. Each blue highlight corresponds to the AAE equivalent from their respective SAE counterpart.

<table>
<thead>
<tr>
<th>Algorithm 1: The translative syntactic morphological method for CODESWITCH.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Input: Original SAE sequence $x$</td>
</tr>
<tr>
<td>2 Output: Translated AAE sequence $x'$</td>
</tr>
<tr>
<td>3 begin function</td>
</tr>
<tr>
<td>4 Load SAE input sequence $x$</td>
</tr>
<tr>
<td>5 $x \leftarrow$ LOWER($x$)</td>
</tr>
<tr>
<td>6 $T \leftarrow$ TOKENIZE($x$)</td>
</tr>
<tr>
<td>7 for all $i = 1, 2, ...,</td>
</tr>
<tr>
<td>8 if $i \in {\text{TMsP}}$ then</td>
</tr>
<tr>
<td>9 $T_i \leftarrow$ CODESWITCH($i$)</td>
</tr>
<tr>
<td>10 end if</td>
</tr>
<tr>
<td>11 end for</td>
</tr>
<tr>
<td>12 $x' \leftarrow$ DETOKENIZE($T$)</td>
</tr>
<tr>
<td>13 return $x'$</td>
</tr>
<tr>
<td>14 end function</td>
</tr>
</tbody>
</table>

Obtaining new texts for downstream tasks from authors of certain demographic groups is time-consuming and requires heavy human labor (Liu et al., 2021; Dacon, 2022). Therefore, we create CODESWITCH (see Algorithm 1), a greedy unidirectional morphosyntactically-informed rule-based translation method which is not only fast, but also functions as a human-in-the-loop paradigm; therefore, drastically reduces heavy human labor. Our approach for intent-and-semantically equivalent AAE data augmentation is intuitively simple and effective. Consequently, we can now explore code-switching in several NLP tasks to determine if LLMs such as MLMs learn to make predictions based on demographic/dialectal language features.

We represent each original NLI corpus as $D < P, H, L >$ with $p \in P$ as the premise, $h \in H$ as the hypothesis and, lastly, $l \in L$ as the label, and create two augmented datasets i.e., SNLI AAE and MNLI AAE, where we represent each augmented NLI dataset as $D' < P', H', L >$. Specifically, translate each premise-hypothesis pair to AAE and keep the original label unchanged to form a new instance. It is important to note that the task of CODESWITCH is to ensure both sets of datasets i.e., $D$ and $D'$ maintain their contextual accuracy, although they consist of two different language styles (see Table 2).

3.3 Human Evaluation

After an initial training of the AMT annotators with our annotation guidelines, we implement a minor calibration study by tasking the remaining 3 independent workers to test our AAE data augmentation method. We randomly sample 200 SAE/AAE sentence pair examples from each of the 4 datasets, for a total of 800 sentence pairs (or 1600 SAE/AAE sentences). The workers were asked to indicate (1) whether the AAE sentences are written by an L1 (or dominant) AAE speaker, or most likely to be machine generated (MG); and (2) whether or not their contextual accuracy is maintained. For content analysis to ensure the quality of our AAE samples and to quantify the extent of agreement between raters, we first let 3 annotators independently rate each AAE-generated sentence pair as “Native” or “MG”, then we measure the inter-annotator agreement (IAA) using Krippendorff’$\alpha$.

We calculate an inter-rater reliability of 0.82, and did not observe significant differences in agreement across the individual sentences. Qualitative analysis revealed that generated samples resembled sequences written by L1 AAE speakers, whereas few samples were classified as most likely MG. Annotators informed us of particular morpho-syntax cases, for example, constant copula deletion of the verb “be” and its variants, namely “is” and “are” is irregular and often inserted last in word order. This indicates that CODESWITCH does not account for contextual instances when generating AAE samples, hence being classified as most likely MG.

4 Empirical Study and Analysis

In this section, we conduct a preliminary study to substantiate the existence of inherent linguistic bias in NLI models. We introduce the base NLI models
and training details, and then we demonstrate our empirical results.

To illustrate inherent linguistic bias of two distinct linguistic systems, we introduce a representative MLM, namely, BERT (Devlin et al., 2018) (see Appendix A for more details).

<table>
<thead>
<tr>
<th>Models</th>
<th>Model Performance (%)</th>
<th>SNLI SAE</th>
<th>AAE</th>
<th>Diff.</th>
<th>MNLI SAE</th>
<th>AAE</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT_BASE</td>
<td>90.12</td>
<td>86.77</td>
<td>4.12</td>
<td>79.79</td>
<td>84.47</td>
<td>79.79</td>
<td>4.68</td>
</tr>
<tr>
<td>BERT_LARGE</td>
<td>90.46</td>
<td>74.55</td>
<td>15.91</td>
<td>79.79</td>
<td>84.47</td>
<td>67.35</td>
<td>17.12</td>
</tr>
</tbody>
</table>

Table 3: Model performance when tested on AAE data. The intensity of each red highlight directly corresponds to the absolute difference in accuracy disparities.

We use each original dataset i.e., SNLI and MNLI to fine-tune both BERT models on a batch size of 32 using an AdamW optimizer with a learning rate of 2e-5 and default betas ($\beta_1 = 0.9$, $\beta_2 = 0.999$) for 3 epochs. Our experiments display that pretrained MLMs “are only as good as the data they are trained on” and are unable to make fair predictions (Tan et al., 2020a). In Table 3, we see that the lack of diverse training data results in disparities in model performance in MLMs, which may be significantly be intensified as models become more complex. In Table 4, we illustrate several examples on the inherent linguistic bias on account of demographic language features, and can conclude that demographic/dialectal language features do in fact influence models to produce false predictions.

5 Debiasing Methods

In Section 4, we empirically demonstrate that popular NLI models show significant bias towards AAE by underperforming on them than SAE. A natural question arises: how can we remove the biases in NLI models towards different language styles? To solve this problem, we introduce two simple but effective debiasing strategies: (1) counterpart data augmentation (CDA); and (2) language Style disentanglement (LSD).

5.1 Counterpart Data Augmentation

The bias of NLI models originates from the training data. Since the training data contains only SAE, the NLI models trained on such data does not understand the unique vocabulary and grammar of AAE, which leads to poor performance. Thus, we propose to implement CodeSwitch to augment the original SAE training data by translating them to their AAE counterparts and in turn implement CDA strategy similar to (Zhao et al., 2018; Zmirod et al., 2019). Then, we will get a large augmented training dataset, $D^+$, which is twice the size of the original datasets (i.e., SNLI) as it contains both $D$ and $D'$.

5.2 Language Style Disentanglement

For two texts with the similar intent and semantic content of different language styles (e.g. SAE v.s. AAE), an NLI model may tend to make biased predictions towards one style. The immediate reason is that the NLI prediction are based on the language style features, instead of relying solely on the semantic features of the texts. Based on this consideration, we propose LSD, an in-processing debiasing method, which tries to disentangle the language style features from the semantic features in text representations and forces the NLI model to make inference on the pure semantic representations.

5.2.1 The LSD Framework

To achieve disentanglement, we adopt the idea of adversarial learning. Figure 1 illustrates the overall framework of LSD. We view the framework as three parts: (1) the BERT model that encodes a premise-hypothesis pair as a fixed-dimensional representation $E_{CLS}$; (2) a feed-forward neural (FFN) classifier $C$ that takes $E_{CLS}$ as input to predict the inferential relationship between the premise and the hypothesis; and (3) a FFN discriminator $D$ that predicts whether the sentence pair is SAE or AAE based on $E_{CLS}$. Via adversarial learning, our goal is to build a BERT model that can produce an accurate semantic representation of the text pair so that the classifier $C$ can make correct predictions based on it, while the representation is free from the language style features of the texts, so that the discriminator $D$ cannot distinguish whether the texts are from $D$ or $D'$.

5.2.2 An Optimization Method

We present our optimization algorithm for the LSD framework in Algorithm 2. We train the framework on the augmented training dataset obtained via our CodeSwitch method as we do in CDA. In the training data $\mathcal{T} = \{<p_i, h_i, l_i, s_i>\}_{i=1}^{7}$, each instance consists of a premise $p$, a hypothesis $h$, a label $l$, and a binary language style label $s \in \{\text{SAE, AAE}\}$. At the beginning, we first load pretrained BERT parameters, and initialize the pa-
<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypothesis</th>
<th>Label</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Dis</em> church choir sings <em>ta</em> <em>da</em> masses as <em>dey</em> song.</td>
<td><em>Da</em> church filled <em>wit</em> song.</td>
<td>Entailment</td>
<td>Neutral</td>
</tr>
<tr>
<td><em>Dis</em> church choir sings <em>ta</em> <em>da</em> masses as <em>dey</em> song.</td>
<td><em>Da</em> church filled <em>wit</em> song.</td>
<td>Entailment</td>
<td>Neutral</td>
</tr>
<tr>
<td><em>Dis</em> church choir sings <em>ta</em> <em>da</em> masses as <em>dey</em> song.</td>
<td><em>Da</em> church has cracks in <em>da</em> ceiling.</td>
<td>Neutral</td>
<td>Contradiction</td>
</tr>
<tr>
<td>A woman <em>wit</em> a green headscarf, blue shirt <em>n</em> a very big grin.</td>
<td><em>Da</em> woman very happy.</td>
<td>Entailment</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

Table 4: An illustrative example on the inherent linguistic bias of a NLI models. Each blue highlight corresponds to the AAE equivalent from their respective SAE counterpart (see Appendix B)

Figure 1: An illustration of the language-style disentanglement model.

Algorithm 2: The optimization method for the LSD framework.

1. **Input**: Training data \( \mathcal{T} = \{ < P_i, H_i, L_i, S_i > \}_{i=1}^{|\mathcal{T}|} \) and Validation data \( \mathcal{V} = \{ < P_i, H_i, L_i, S_i > \}_{i=1}^{|\mathcal{V}|} \)
2. **Output**: BERT parameters \( W^{BERT} \), classifier parameters \( W^{C} \)
3. Load pre-trained parameters \( W^{BERT} \)
4. Initialize \( W^{C} \) and \( W^{D} \)
1. for \( N \) epochs do
2. for \( M \) batches do
3. Obtain a mini-batch of training data \( B \) from \( \mathcal{T} \)
4. Update \( W^{D} \) by optimizing \( L^{D} \) in Equation 1
5. Update \( W^{BERT} \) and \( W^{C} \) by optimizing \( L \) in Equation 2
6. end for
7. Run the BERT model and the classifier \( C \) on validation data \( \mathcal{V} \)
8. Save parameters \( W^{BERT} \) and \( W^{C} \) if achieving the best validation performance so far.
9. end for

where \( L \) is the set of labels of the NLI task. \( S = 0, 1, 2 \) represent for entailment, contradiction, and neutral, respectively. \( p^{D}_j \) indicates the predicted probability for the \( j \)-th label from the classifier \( C \). Minimizing \( L^{C} \) will force \( C \) to make correct predictions. To ensure that the BERT model produces a text representation that can fool the discriminator, when training, we consider another entropy loss:

\[
L^{D} = -(p^{D}_0 \log p^{D}_0 + p^{D}_1 \log p^{D}_1)
\]

\( L^{D} \) is the entropy of the predicted distribution \( p^{D} \) from the discriminator. Minimizing it makes \( p^{D} \) close to an even distribution, preventing \( D \) from making correct predictions. We update the BERT model and the classifier by minimizing the following combined loss (line 5):

\[
L = L^{C} + L^{D}
\]
At the end of each epoch, we run the BERT model and the classifier on the validation data, and save their parameters if they achieve the best validation performance.

5.3 Experimental results

In Table 5, we show the performances of the two debiasing methods on two datasets in terms of two BERT models. In Table 3, the results of the debiased models CDA, LSD and that of the original models were compared. Note that our two debiasing methods reduce the gap between the performances on SAE and AAE significantly. The original BERT models perform well on SAE test data but exhibit a decrease in performance when they are tested on AAE data. However, the BERT models trained under CDA or LSD debiasing strategies achieve similar model performance on SAE and AAE, which demonstrates the effectiveness of the two debiasing methods to mitigate bias in NLI models.

<table>
<thead>
<tr>
<th>Models</th>
<th>Model Performance (%)</th>
<th>SNLI</th>
<th>MultiNLI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>SAE</td>
<td>AAE Diff.</td>
</tr>
<tr>
<td>CDA_BASE</td>
<td>89.77</td>
<td>89.76</td>
<td>0.01</td>
</tr>
<tr>
<td>LSD_BASE</td>
<td>90.35</td>
<td>90.49</td>
<td>0.14</td>
</tr>
<tr>
<td>CDA_LARGE</td>
<td>90.48</td>
<td>90.36</td>
<td>0.12</td>
</tr>
<tr>
<td>LSD_LARGE</td>
<td>90.60</td>
<td>90.53</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Table 5: Model performances of two debiased NLI models. The intensity of each green highlight directly corresponds to the absolute difference in accuracy.

Furthermore, our debiased models not only improve the performance on AAE data, but also maintain similar performance on SAE data as the original model. This is due to either the introduction of additional AAE training data which is not always available, and the disentanglement between the semantic and language style features of texts enhancing the model’s capability of understanding natural language. Lastly, we find that LSD generally outperforms CDA on both SAE and AAE data. In addition, LSD is an adversarial learning debiasing method that filters out irrelevant language style information towards the NLI task. In fact, LSD is also generalizable for more effective and architecturally similar models such as DeBERTa (He et al., 2020), XLNet (Yang et al., 2019), and T5 (Raffel et al., 2019) to ensure fairness as well as robustifying larger language models.

6 Related Work

Previous works focus on AAE in the context of racial bias as a result of systemic biases in model performance. For example, Blodgett et al. (2018) focus on dependency parsing social media AAE to analyze the impacts of performance disparities between AAE and SAE tweets. Other works undertake AAE within the scope of detecting and mitigating the presence of racial bias in areas of offensive and abusive language detection (Liu et al., 2020a; Sap et al., 2019), sentiment analysis (Groenwold et al., 2020) and hate speech detection (Davidson et al., 2019; Sap et al., 2019). However, these influential works do not engage with AAE literature, utilize a human-in-the-loop paradigm nor employ the humans who create such data. Thus, these pivotal works fail to understand AAE’s phonological and morphological language features—thereby simply treating AAE as another non-Penn Treebank English variety (Blodgett et al., 2020).

Fairness in NLP. As social and racial disparities have become a compelling issue within the NLP community, focal topics of fairness, accountability, ethics, sustainable development, etc., have gained momentous attention in recent years (Hovy and Spruit, 2016). Recent work on fairness has primarily been focused on racial and gender biases in distributed word representations (Bolukbasi et al., 2016; Zhao et al., 2018; Zmigrod et al., 2019), coreference resolution (Rudinger et al., 2018), sentence encoders (May et al., 2019), machine translation (Tan et al., 2020b; Prates et al., 2018), and dialogue generation (Liu et al., 2020a,b).

Adversarial learning in NLP. Adversarial examples were initially explored in computer vision by Szegedy et al., where these examples were intended to influence models to produce false predictions. However, in NLP, adversarial examples can occur at a phonetic, phonological, morphological, syntactic, semantic, or pragmatic level (Tan et al., 2020a; DeBose, 1992; Gardner-Chloros et al., 2009; Young and Barrett, 2018). Liu et al. (2020a) displays that dialogue systems are prone to produce offensive responses when fed AAE language features in comparison to SAE, whereas Liu et al. (2020b) propose a novel adversarial learning framework which directly addresses the issue of gender bias in dialogue models while maintaining their performance. Both Alzantot et al. (2018) and Joshi et al. (2019) exploit the notion of adversariality by utilizing word embeddings to find the k nearest
synonymic examples.

**Summary.** These influential works demonstrate novel adversarial learning methodologies on a character and/or word-level in order to address bias issues surrounding protected attributes such as race and gender by improving model robustness. Similarly, our work utilizes a human-in-the-loop paradigm by employing humans who create such data, to create a novel morphosyntactic method to perturb language styles on a syntactic-level to highlight the need for dialectal language diversity in training data.

7 Conclusion and Future Works

To address compelling fairness, accountability, transparency, and ethical concerns surrounding the sustainability of language use in NLP applications, we claim that the addition of diverse dialectal language in training data will improve model robustness and generalizability. Our findings show that our proposed debiasing methods not only improves the performance on AAE data but effectively reduces the performance gap between SAE and AAE significantly, while maintaining or even improving the prediction performance on SAE data. Therefore, training under these two debiasing strategies aids in the mitigation of linguistic bias in NLI models.

We conclude that though similar, the two language styles, SAE and AAE are not identical, and thus, should not solely be evaluated against each other, but compared to as a basis of model performance minimize the existence of inherent linguistic bias in language models. In the future, we intend to release CODESWITCH a morphosyntactically-informed rule-based translation method for unidirectional data augmentation for generating intent-and-semantically-equivalent AAE examples as a public python package, to encourage further computational linguistic research into debiasing various NLP systems. We actively intend on updating CODESWITCH s.t. it can include new or regional-specific lingo. In this way, CODESWITCH can constitute potential groundwork on ways that AAE can effectively be integrated in NLP systems to improve future language models during their development and employment.

8 Limitations And Ethical Considerations

All authors must warrant mentioning that the increased performance for underrepresented dialects in NLP systems has the potential to enable automated discrimination based on the use of non-standard dialects. Although, we attempt to highlight the need for dialectal inclusivity for impactful speech and language technologies, we do not intend for increased feelings of marginalization of an already stigmatized community.

We have established our method’s effectiveness for data augmentation for generating intent-and-semantically-equivalent AAE examples and believe that CODESWITCH could be further improved by addressing the following limitations:

1. Currently, CODESWITCH is a unidirectional data augmentation method and cannot be used in reverse as a deterministic text normalization/preprocessing system which can convert all text to SAE.

2. CODESWITCH operates on simple, deterministic substitutions for morphosyntactically-informed translations rules found in Appendix B rather than that of real L1 and L2 AAE speakers, which may result in the lack of several formal/informal phrases, expressions, idioms, cultural and regional-specific lingo, and slang-related words (Blodgett et al., 2020). For example, “I **shall** was finna ask who money dat is”, where “**shall**” refer to the replacement of the word “**sure**”.

3. Although CODESWITCH possesses several simple, deterministic morphosyntactically-informed translation rules it does account for contextual instances of accurate copula deletion. This may lead to a discrepancy between actual text written by L1 and/or L2 AAE speakers and our proposed data augmentation method.

In the future, we intend to address these limitations and ethical considerations by partnering with AAE diglossic communities in hopes of robustifying CODESWITCH to be probabilistic rather than deterministic to capture different AAE variants of the same SAE term (for example, the AAE equivalents to “what’s” → “**waz*/”wus*”*/”wats”. In addition, we will investigate inherent linguistic bias in other NLP applications.

References

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang.


we will now give details of each pretrained BERT
mantic equivalence, and quality of a text. Please
textual accuracy i.e., original structure, intent, se-
from SAE to AAE while maintaining con-
Here we present a set of 20 linguistic phonetic
B Translative Morpho-syntax Protocol
mographic groups e.g. African Americans. Now,
differences in language styles from different de-
tions between sentence pairs by examining the
for our task of understanding the inferential rela-
formly, and thus, it serves as a appropriate model
In summary, BERT optimizes its two objectives uni-
jectives: (1) Masked Language Modeling (MLM)
exploits an attention mechanism to learn contextual
aches state-of-the-art results in a wide variety of
NLP tasks. BERT is trained on a huge Books Cor-
achieves state-of-the-art results in a wide variety of
mological linguistic AAE features (which we adapt from AAE literature). Our
deterministic translative morpho-syntax protocol (TMSP) and its cases are as follows:
1. Consonant (’t’) deletion (Special case) : e.g.
   “just” → “jus”; “must” → “mus”
2. Contractive (’all) gain: “You all” → “Y’all”
3. Contractive negative auxiliary verbs replace-
   “doesn’t” → “don’t”
4. Contractive (’re) loss: e.g. “you’re” → “you”;
   “we’re” → “we”; “they’re” → “they”
5. Contractive word replacement: e.g. “isn’t” →
   “ain’t”; “wasn’t” → “ain’t”
6. Copula deletion: Deletion of the verb “be”
   and its variants, namely “is” and “are” e.g.
   “He is on his way” → “He on his way”; “You
   are right” → “You right”
7. Gerund consonant (’g’) deletion and retain-
   • Consonant (’g’) deletion: e.g. “coming”
     → “comin”; “going” → “goin”
   • Consonant (’g’) retention (Exception case): e.g. “–ing”
8. Homophonic word replacement: e.g. “whine”
   → “wine”; “you’re” → “your”
9. Indefinite article replacement: e.g. “an” →
   “a”
10. Indefinite pronoun replacement: e.g. “anyone”
     → “anybody”; “everyone” → “everybody”
11. Interdental fricative loss: e.g. “this” → “dis”;
    “that” → “dat”; “than” → “dan”; “their” →
    “they (dey)”; “the” → “da”
12. Negative concord replacement: e.g. “Don’t
    say anything” → “Don’t say nothing”
13. Phrase reduction (present/ future tense) ⇒
    word e.g. “going to” → “gonna”; “want to”
    → “wanna”; “trying to” → “tryna”; “what’s
    up” → “wassup”; “fixing to” → “finna”
14. Possessive (’s) removal: e.g. “He’s mad at me”
    ⇒ “He mad at me”

A Implementation Details
A.1 Details of the Base Model
BERT – Bidirectional Encoder Representations
from Transformers (BERT) (Devlin et al., 2018) is
a Transformer-based ML technique for NLP
that exploits an attention mechanism to learn contextual
relationships between words and optimizes two ob-
jectives: (1) Masked Language Modeling (MLM)
and (2) Next Sentence Prediction (NSP), and has a
vocabulary size of 30,522.
A.2 Details of Experimental Settings
In summary, BERT optimizes its two objectives uni-
formly, and thus, it serves as a appropriate model
for our task of understanding the inferential rela-
tionships between sentence pairs by examining the
differences in language styles from different de-

tographic groups e.g. African Americans. Now,
we will now give details of each pretrained BERT
model below:
1. BERT-base-uncased - Trained on raw English
text, and consists of 12-layers, 768-hidden,
12-heads, 110M parameters.
2. BERT-large-cased - Trained on raw lower-
cased English text, and consists of 24-layer,
1024-hidden, 16-heads, 335M parameters.
Trained on cased English text.
B Translative Morpho-syntact Protocol
Here we present a set of 20 linguistic phonetic
and morphological text rules that are used to code-
switch from SAE to AAE while maintaining con-
textual accuracy i.e., original structure, intent, sem-
antic equivalence, and quality of a text. Please
note that these are only a few examples of the most
commonly used morphological linguistic AAE fea-
tures (which we adapt from AAE literature). Our

deterministic translative morpho-syntax protocol
(TMSP) and its cases are as follows:

Xiang Zhou and Mohit Bansal. 2020. Towards robusti-
yzing NLI models against lexical dataset biases. In
Proceedings of the 58th Annual Meeting of the Asso-
ciation for Computational Linguistics, pages 8759–
8771, Online. Association for Computational Lin-
guistics.
Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and
Ryan Cotterell. 2019. Counterfactual data augmentation
for mitigating gender stereotypes in languages
with rich morphology. In Proceedings of the 57th
Annual Meeting of the Association for Computational
Linguistics, pages 1651–1661, Florence, Italy. Assoc-
iation for Computational Linguistics.

Ryan Cotterell. 2019. Counterfactual data augmenta-
tion NLI models against lexical dataset biases. In
Proceedings of the 58th Annual Meeting of the Asso-
ciation for Computational Linguistics.
15. Present tense possession replacement: e.g. “John has two apples” → “John got two apples”; “The neighbors have a bigger pool” → “The neighbors got a bigger pool”

16. Remote past “been” + completive (‘done’): “I’ve already done that” → “I been done that”

17. Remote past “been” + completive (‘did’): “She already did that” → “She been did that”

18. Remote past “been” + Present tense possession replacement: “I already have food” → “I been had food”; “You already have those shoes” → “You been got those shoes”

19. Term-fragment deletion: e.g. “brother” → “bro”, “sister” → “sis”, “your” → “ur”, “suppose” → “pose”; “more” → “mo”


C Annotation Guidelines

You will be given a phrase that is written in Standard American English (SAE), your task is to correctly identify if the translatable vocabulary rules in Appendix B are accurate in order to translate SAE text to AAE text. Furthermore, while reviewing the rules, be sure to mention that these rules and/or morpho-syntax word cases in the sampled premise-hypothesis sentence pairs maintain their contextual accuracy i.e., original structure, intent, semantic equivalence and quality.

SAE to AAE Protocol

1. Are you a dominant AAE speaker?

2. If you responded “yes” above, are you bidental?

3. If you responded “yes” above, are you capable of code-switching by alternating between SAE and AAE frequently on a daily basis in a single conversation or situation?

4. Given TMsP above in Appendix B, are these main grammatical, structural and syntactic rules of word case usage of AAE linguistic features?

5. If you responded “no” above, can clarify which rule is insufficient? In addition, if possible, can you provide a grammatical, structural or syntactic rule that is not detailed in Appendix B?

D Contextual accuracy Protocol

Given a table of SAE-AAE sentence pairs examples, determine whether or not their contextual accuracy is maintained.

<table>
<thead>
<tr>
<th>SAE</th>
<th>AAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>i will go back to the house</td>
<td>imma go back ta da house</td>
</tr>
<tr>
<td>i don’t want to go to bed</td>
<td>ion wanna go ta bed</td>
</tr>
<tr>
<td>he isn’t my friend, but he’s a king</td>
<td>he ain’t my friend, but he a king</td>
</tr>
<tr>
<td>she is being weird to me</td>
<td>she been weird ta me</td>
</tr>
<tr>
<td>you all are annoying</td>
<td>yall annoyin</td>
</tr>
<tr>
<td>he isn’t coming anymore</td>
<td>he ain’t comin no mo</td>
</tr>
<tr>
<td>a woman is trying to walk</td>
<td>a woman tryna walk</td>
</tr>
<tr>
<td>this bag and that shoe are mine</td>
<td>dis bag n dat shoe mine</td>
</tr>
<tr>
<td>their kids are laughing</td>
<td>they kids laughin</td>
</tr>
<tr>
<td>john and kates have two dogs</td>
<td>john n kates hav two dogs</td>
</tr>
<tr>
<td>you are going through something</td>
<td>u goin thru sum'n</td>
</tr>
<tr>
<td>what are you doing</td>
<td>wat u doin</td>
</tr>
<tr>
<td>what’s the temperature</td>
<td>wus da temperature</td>
</tr>
<tr>
<td>they have a better car than us</td>
<td>dey hav a betta car dan us</td>
</tr>
<tr>
<td>so you’re going to the party</td>
<td>so your gonna go ta da party</td>
</tr>
<tr>
<td>they are singing but they can’t sing</td>
<td>dey singing but dey can’t sing</td>
</tr>
<tr>
<td>you could of have it all</td>
<td>u coulda hav it all</td>
</tr>
<tr>
<td>he would’ve had it if he was here</td>
<td>he woulda had it if he was here</td>
</tr>
<tr>
<td>we should have been first in line</td>
<td>we shoulda been first in line</td>
</tr>
<tr>
<td>he should of had the last bite</td>
<td>he shoulda had da last bite</td>
</tr>
</tbody>
</table>

Table 6: SAE examples and their AAE equivalents (after using CODESWITCH).

1. As you responded “yes” a previous question, ... are you capable of code-switching by alternating between SAE and AAE frequently on a daily basis in a single conversation or situation?

We will now provide 20 lower-cased test sentences is Table 6.

2. Have you ever seen any of these words in a particular sentence in Table 6, for example, on social media such as Twitter?

3. If you responded “yes” above, For each SAE sentence, does each plausible AAE sentence resemble adequate AAE morphological language features from a dominant AAE speaker after applying CODESWITCH?

4. If you responded “yes” above, do these pairs maintain their contextual accuracy i.e., original structure, intent, semantic equivalence and quality?
5. For dialectal (morphological and phonological) purposes, are these particular words spelt how would you say or use them? For example, texting or posting on social media?

6. If you responded “no” above, can you provide a different spelling along with its SAE equivalent?