Evaluating and Mitigating Inherent Linguistic Bias Against African American Vernacular English Dialect in Token-Corruption Trained Discriminative Large Language Models

Hayley Zorkic
HLZ227
hayley.zorkic@utexas.edu

Abstract

This project investigates if the BERT-based language model, ELECTRA comprehends English by investigating how well it can perform Natural Language Inference tasks on non-standard English dialects- in particular African American Vernacular English (AAVE). First, the model's ability to generalize to AAVE is evaluated using a challenge test set of AAE premise-hypothesis pairs. Subsequently, potential errors are addressed through pretraining on both Standard American English (SAE) and AAVE training datasets. Unfortunately, we found only a trivial increase in accuracy is achieved on NLI tasks when AAVE data is included in the pre-training process. The discussion section provides insights into future research directions for further improvement.

1 Introduction

1.1 Motivation

Natural Language Processing (NLP) tools are often trained and evaluated on predominant language variants, such as Standard American English (SAE). This results in a significant decline in the performance when applied to non-SAE dialects such as African American Vernacular English (AAVE), Cajun Vernacular English, and Chicano English to name a few. Studies indicate that SAE models tested on AAVE encounter difficulties in language identification (Jurgens et al., 2017). As we continue to integrate Large Language Models (LLMs) into daily life, it is critical these models can comprehend more casual and diverse linguistic patterns. If we do not ensure language models can understand nuanced patterns of English, there is substantial potential for generating harmful and biased output that exhibit undesirable behaviors. Furthermore, it is critical to identify these patterns and correct them.

In comparison to SAE, AAVE possesses distinctive grammatical structures, vocabulary, and syntactical patterns. The patterns learned during the model's training on SAE data result in difficulties during language identification tasks and other natural language tasks when applied to AAVE or other dialects. This project specifically focuses on the task of Natural Language Inference (NLI) where a model determines whether the given "hypothesis" logically follows from a "premise". More explicitly, NLI aims to determine if a hypothesis is true based only on the premise provided. We would like to use NLI as a proxy to identify and correct model that were originally trained on mostly SAE. We expect SAE trained models will have poor performance when completing similar NLI tasks when the test data is AAVE.

1.2 Methods Overview

First, we propose investigating whether major NLP models even recognize AAVE in the first place. The simplest way to do this is to use a commonly used pre-trained model, such as ELECTRA-small pretrained on the MNLI dataset, and then test it on a small subset of AAVE NLI data. We decided to choose MNLI instead of SNLI because of the nature of the data in each- MNLI data is train on many more casual language bases like Twitter, whereas SNLI is trained on more formal language bases like Wikipedia. If a model is trained on formal sources like Wikipedia and Project

Gutenberg, it has seen very few examples of more casual language patterns which effect performance on conversational SAE and *especially* AAVE as there are different linguistic patterns from SAE all together. On the other hand, if the data are trained on social media data, AAVE is much more likely to appear.

2 Methodology

This project aimed to mirror the methods and results outlined in the DADA paper from Liu, et al. (2023). Due to time constraints and this being a solo project, we were only able to get through stage 1 of the outlined experimentation, however, we do suggest continuing to attempt to replicate the results from that paper (more on this in the discussion).

2.1 Datasets

For pretraining our ELECTRA-small model, we used the **SAE** *MNLI* dataset as outlined in the paper. The Multi-Genre Natural Language Inference (MultiNLI, MNLI) corpus is a collection of 433k crowd-sourced sentence pairs annotated with textual entailment information. Modeled after the SNLI corpus, MNLI differs as it covers a range of genres of spoken and written text, supporting a distinctive "cross-genre generalization evaluation." Because MNLI consists of a broader range of genres and writing styles, making it more representative of various real-world scenarios, we suspect any accuracy improvements from the baseline of an **SAE** MNLI pretrained model would represent more significant gains than improvements on SNLI pretrained models as the gap from SNLI to **AAVE** is much larger than the gap from MNLI to **AAVE**. Because our task at hand is deals with more "casual" language pattens, MNLI gives our model the most appropriate general understanding of English.

For our analysis and further fine-tuning, we needed to find an NLI task dataset in **AAVE**, however this proved to be *incredibly difficult*. At first, we assessed the plausibility of prompting a language model to convert examples from the MNLI or SNLI dataset from **SAE** to **AAVE**, but in my research, we found many "SAE-to-AAVE" translator tools to create incredibly ignorant and naïve translations rooted a misunderstanding that AAVE is SAE with mistakes. So, we took to the literature to find an AAVE NLI dataset created by reputable sources. Ideally, this data would be generated by people who spoke AAVE, however, this is not publicly available. So, we searched for labs who created synthetic datasets based on very specific linguistic patterns. Luckily, we were able to find a robust synthetic **AAVE** NLI dataset from the authors of the DADA paper via their HuggingFace account. The data are common NLI examples that have been synthetically adapted from other NLI datasets using a series of 10+ linguistic rules outlined in many studies. Though not ideal, this is the best AAVE NLI data we could find.

2.2 Pretrain Model on SAE NLI data

Pretraining on SAE MNLI can provide the model with a general understanding of natural language inference tasks, which may include some transferable knowledge to AAVE. For our baseline model, we are Pre-training ELECTRA small on **SAE** MNLI data. We will perform testing on a small sample of the SAE NLI dataset, and testing on the AAVE validation set. From here, we will investigate common prediction errors in the model performance.

2.3 Pretrain Model onAAVE+SAE NLI data

For our model adaptation, we will use the ELECTRA-small model, but pretrain on a composite dataset of AAVE and SAE NLI examples to expose the model to some AAVE patterns. Then, we will test on SAE NLI dataset and the AAVE NLI validation set. We supplement the training data with examples from the challenge set to provide the model with a more comprehensive understanding of the task at hand in hopes that the model performance on AAVE tasks would improve.

The DADA paper outlines minimal improvements in model test accuracy scores for both AAVE and SAE datasets, so to expand on their analyses, we will attempt to modify model parameters to achieve better performance gains. Furthermore, we train the models using 3, 5, and 10 epochs.

2.4 Pretrain Model on AAVE NLI data

For another model adaptation experiment, we will use the ELECTRA-small model, but pretrain on just AAVE NLI examples which will undoubtedly expose the model to AAVE

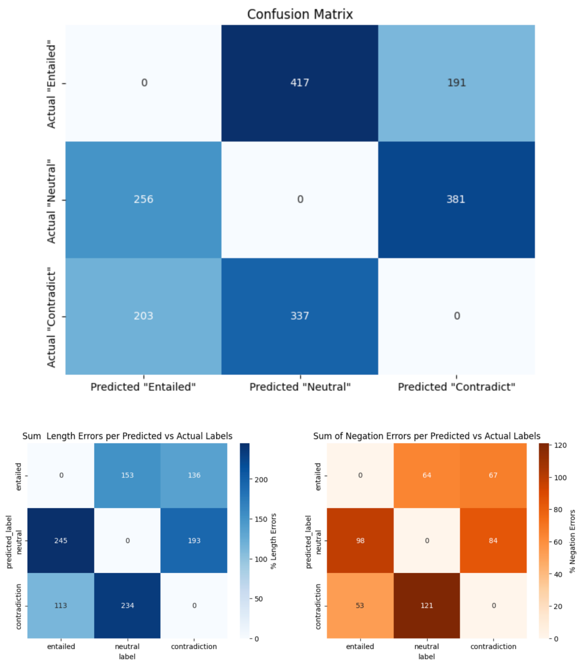

Figure 1: (a) A heatmap showing Predicted vs Actual labels of INCORRECT **SAE test results**. (b) Heatmaps divide this heatmap into two common error types- negation and length related errors. Length imbalances tend to lead to neutral pairs being classified as contradictory and entailed pairs to be considered neutral. Negation errors tend to cause many errors in correctly predicting neutral pairs- in most cases, neutrals are incorrectly labeled as contradictions.

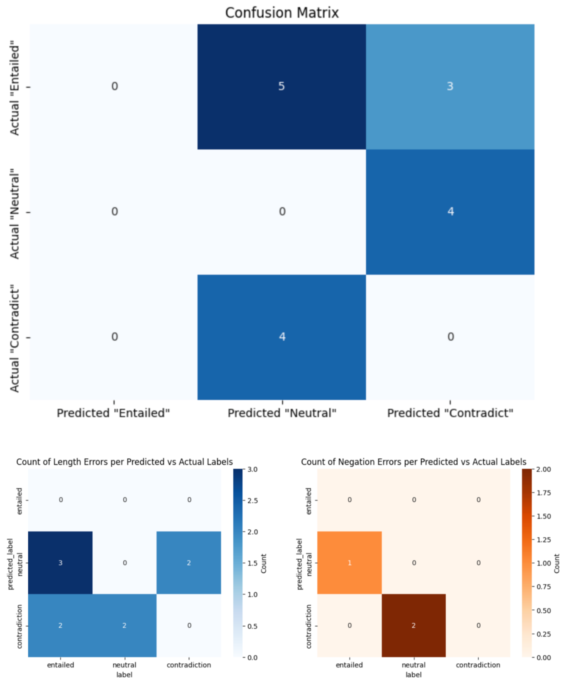

Figure 2: (a) A heatmap showing Predicted vs Actual labels of INCORRECT **AAVE test results**. Interestingly, it seems the model does not predict any entailments incorrectly. (b) Heatmaps divide this heatmap into two common error types- negation and length related errors. Length imbalances tend to cause more of an impact than negation.

patterns. Then, we will test on SAE NLI dataset and the AAVE NLI validation set.

3 Analysis

3.1 Investigating Dataset Artifacts

It looks like our baseline model had two main error types- length imbalance and negation. Both categories made about ~70-90% of the errors in the subset of data that was labeled incorrectly in both the SAE and AAVE datasets. Figure 1 and Figure 2 show heatmaps of the ERRORS from these evaluations. These heatmaps do NOT include correct predictions. There were FAR less AAVE pairs than SAE pairs, and the density of the heatmaps showcase that. For AAVE and SAE respectively, we had an accuracy of 81.6% and 81.8%, which is decent!

Figure 1 shows the incorrect SAE test results. The larger heatmap shows the model had some difficulties correctly predicting neutrals. The sub-heatmaps for negation and length related errors show length imbalances tend to lead to neutral pairs being classified as contradictory and entailed pairs to be considered neutral. Negation errors tend to cause many errors in correctly predicting neutral pairs- in most cases, neutrals are incorrectly labeled as contradictions.

Figure 2 is a heatmap showing Predicted vs Actual labels of INCORRECT AAVE test results. Though the test set was small, it appears the model does not predict any entailments *incorrectly*! Heatmaps divide this heatmap into two common error types- negation and length related errors. Length imbalances tend to cause more of an impact than negation on AAVE errors.

3.2 Mitigating Dataset Artifacts

In attempts to improve accuracy of the models, we tried a few things. First, we downloaded the ELECTRA-small model and pretrained it on SAE data with 3 epochs- this is our baseline model. Next, we pretrained ELECTRA-small on a composite dataset made up of SAE+AAVE NLI pairs for 3, 5, and 10 epochs. Finally, we used the ELECTRA-small model we pretrained on SAE

Dialect Adaptation Details				Test Accuracy		
Backbone	Method	Training Data	epochs	AAVE	SAE	SNLI
ELECTRA small	Pretrain	SAE	3	81.6%	81.8%	77.2%
	Pretrain	SAE + AAVE	3	81.6%	81.8%	76.9%
	Pretrain	SAE + AAVE	5	78.1%	82.5%	77.0%
	Pretrain	SAE + AAVE	10	74.7%	81.7%	77.1%
ELECTRA Pretrained on SAE	Continued Pretraining	AAVE	3	81.6%	81.0%	76.5%

Table 1: Test Data Accuracy Results. This chart shows a variety of model accuracies. We tried pretraining ELECTRA small on SAE data and an SAE + AAVE data hybrid. Increasing the number of training epochs only maintained or minimized the performance on the AAVE test set.

data and then pretrained it again on AAVE data for just 3 epochs.

4 Results

Unfortunately, we did not observe significant gains in the model performance when attempting to pretrain on adversarial data, such as the AAVE challenge sets directly or the SAE+AAVE composites (Liu et al., 2019; Zhou and Bansal, 2020; Morris et al., 2020). We do believe the implementation of the models and training were correct as we didn't see dramatic decreases in accuracy- at worst the model was not learning anything new. We suspect the poor results could be due to a few reasons.

1. The quality of the AAVE dataset. Recall the AAVE NLI data was synthetic and quite small in relation to the SAE data.
2. Fine-tuning the model could have been fine tuned on the challenge set itself rather than pretrained. This would've allowed the model to adapt to the complexities present in the specific linguistic style of the challenge set.
3. Amalgamating the challenge data. We believe that initializing several model adapted for unique linguistic rules in the dialect and then combining their predictions through ensemble learning could improve the models generalizability and robustness.

5 Ethics Statement

There is an inherent risk in using synthetic datasets in NLP projects- especially when training as it risks perpetuating biases and inaccuracies. Upon scouring the internet for human generated, publicly available NLI datasets on AAVE, none were to be found, so for the sake of this project, improvisions had to be made. Any further work in this subject area should utilize ethical approaches which engage directly with AAVE speakers to ensure the authenticity and legitimacy of the examples. Furthermore, the difficulty in coming across such a dataset beckons the creation of NLI dataset generation for English dialects.

Acknowledgments

This is a final project for The University of Texas at Austin's Natural Language Processing course taught by Dr. Durrett. Special thanks to Dr. Durrett and all of his wonderful TA's for an i*ncredibly* well run, helpful, informative, and engaging semester.

References

Jurgens, D., Tsvetkov, Y., & Jurafsky, D. (2017). *Incorporating Dialectal Variability for Socially Equitable Language Identification*. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 51–57. Vancouver, Canada: Association for Computational Linguistics.

Liu, Y., Held, W., & Yang, D. (2023). DADA: Dialect Adaptation via Dynamic

Aggregation of Linguistic Rules. *arXiv preprint arXiv:2305.13406*.

Liu, N. F., Schwartz, R., & Smith, N. A. (2019). *Inoculation by fine-tuning: A method for analyzing challenge datasets*. 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)*, 2171–2179. Minneapolis, Minnesota, June: Association for Computational Linguistics.

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). *Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP*.

Zhou, X., & Bansal, M. (2020). *Towards robustifying NLI models against lexical dataset biases*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8759–8771. Online, July: Association for Computational Linguistics.