Adapting BERT to non-Western Dialects: A Case Study on Nigerian Pidgin English Slurs

Stanford CS224N Custom Project

Sathvik Nori
Department of Computer Science
Stanford University
sathvikn@stanford.edu

Adrian Adegbesan
Department of Computer Science
Stanford University
adrian25@stanford.edu

Abstract

This project addresses the significant challenge of culturally insensitive language in Nigerian English and Pidgin by developing a fine-tuned BERT model capable of detecting and neutralizing offensive language. In response to the need for culturally aware automated systems, we compiled a specialized corpus of 1,000 entries, each pairing problematic expressions with their neutralized counterparts. This corpus was instrumental in training our model, which leverages a pre-trained BERT architecture, originally designed to manage problematic language, and adapts it for the specific linguistic features of Nigerian English and Pidgin. The model was rigorously fine-tuned and tested, showing promising results in its ability to accurately recognize and adjust offensive terms. Our evaluations confirm that the model not only mitigates the use of harmful language but also enhances communication across diverse Nigerian communities. By integrating linguistic and cultural sensitivity into its framework, the model represents a significant step forward in developing NLP solutions that respect and uphold the nuances of regional dialects. The deployment of this model aims to enrich digital communication and foster more respectful interactions within and beyond Nigerian borders.

1 Key Information to include

- Mentor: Moussa Doumbouya
- External Collaborators (if you have any): N/A
- Sharing project: N/A

2 Introduction

Natural Language Processing has made significant advances over the past decade, predominantly driven by the development of models that can understand and generate human-like text. Despite these advances, the majority of NLP research has focused on well-resourced languages and dialects, typically those predominant in Western contexts. This focus has left a significant gap in the technology's ability to deal with diverse linguistic phenomena worldwide, particularly in non-Western dialects such as Nigerian Pidgin English. Nigerian Pidgin is an English-based creole language spoken as a lingua franca across Nigeria, known for its unique syntactic, semantic, and pragmatic features, which differ substantially from Standard English Faraclas (1996); Ihemere (2006).

The challenge in extending current NLP methods to languages like Nigerian Pidgin lies in several areas. Firstly, there is a lack of structured linguistic data for training models, as these languages are often underrepresented in text corpora and digital media Joshi et al. (2020). Furthermore, the linguistic variability and fluidity of Pidgin, which often includes code-switching and a mix of local

languages, pose additional complexity for traditional NLP systems Soto and Hirschberg (2018). Current methods, predominantly trained on standard language datasets, often fail to capture the nuanced meanings and cultural expressions found in Pidgin, leading to poor performance in tasks such as sentiment analysis, language translation, and text normalization.

In response to these challenges, our project proposes a novel approach to developing an NLP model specifically tuned to Nigerian Pidgin English. We began by constructing a targeted corpus comprising pairs of biased and neutralized phrases, enabling us to train a model not just to understand Pidgin but to perform a critical task: identifying and neutralizing biased language. Our approach leverages the advancements made by Pryzant et al. in debiasing text Pryzant et al. (2020), adapting these methodologies to the specific linguistic and cultural contexts of Nigerian Pidgin. By fine-tuning a BERT-based model with this specialized corpus, we aimed to create a tool that not only improves communication inclusivity but also enhances the cultural relevance of NLP technologies in African contexts.

3 Related Work

The development of natural language processing tools for under-resourced languages has increasingly become a focus of recent research. However, most efforts have concentrated on larger languages or dialects, often overlooking the linguistic diversity found in regions like West Africa. A significant example of recent work in language neutralization can be found in the study by Pryzant et al. (2020), which introduced a model for automatically neutralizing subjective bias in text using a fine-tuned BERT model Pryzant et al. (2020). Their approach, which targets bias in English texts, significantly inspired our methodology for addressing similar issues in Nigerian Pidgin.

Despite these advancements, research specifically addressing slurs and culturally nuanced derogatory language remains scarce. A significant contribution in this area is the "NaijaHate" study by Tonneau et al. Tonneau et al. (2024), which evaluates Hate Speech Detection (HSD) on Nigerian Twitter using data representative of the regional linguistic diversity. This work proposes the NaijaXLM-T model, tailored to the Nigerian Twitter environment. Their findings emphasize the necessity of domain-adaptive pretraining and fine-tuning to enhance HSD performance and highlight the efficacy of a human-in-the-loop approach to moderate hateful content effectively Tonneau et al. (2024).

While the NaijaHate study provides invaluable insights into handling hate speech in diverse linguistic settings, our work specifically targets the nuances of Nigerian Pidgin English—a creole language distinct from the more widely spoken forms of English and local languages studied in NaijaHate. Our focus is particularly on slurs and culturally nuanced derogatory language that are prevalent in Nigerian Pidgin English. Unlike the broad scope of Nigerian languages covered in NaijaHate, our corpus is curated to capture the unique expressions and slang used in Nigerian Pidgin. This specialization allows us to tailor NLP tools more precisely for contexts where Nigerian Pidgin English is the primary mode of communication, ensuring that the nuances of this linguistic style are accurately understood and appropriately handled.

The application of BERT for language tasks is well-established, but its adaptation for dialects such as Nigerian Pidgin is still at an early stage. While there are numerous adaptations of BERT for different languages and tasks Devlin et al. (2019), these typically do not extend to the type of linguistic nuances involved in Nigerian Pidgin. Our work not only extends the geographical and linguistic scope of existing NLP applications but also introduces a novel resource for researchers and developers working with West African creoles and pidgins. We hope to expand the corpus of NLP work beyond its traditional Euro-centric focus. In summary, our project builds on the foundational models developed by Pryzant et al. and others, extending these approaches to a new linguistic domain where they are critically needed. By creating a targeted corpus and adapting BERT to Nigerian Pidgin, we address both a significant gap in language resources and contribute to the ethical expansion of NLP technologies.

4 Approach

Our project uses a fine-tuned model based on the CONCURRENT architecture by Pryzant, Martinez, and Daas (2020). This model combines a BERT encoder with a token-weighted loss function to detect and reduce Nigerian Pidgin Slurs in relevant texts. It uses a sequence-to-sequence framework, with

BERT as the encoder and an LSTM-based system as the decoder. The model works by repeatedly attending to the encoder's hidden states and generating a probability distribution over the vocabulary to create neutralized text.

In our modified model, detecting problematic phrases and editing them happens within a single architecture. This model is pre-trained and then fine-tuned on our specialized dataset of pidgin phrases. This end-to-end method allows for smooth translation of biased or offensive language into more neutral expressions. A key part of our system is a *join embedding* mechanism, which helps guide the editing process using the detector's output. The join embedding, $v \in \mathbb{R}^h$, is added to each encoder hidden state in the LSTM decoder. This adjustment is controlled by the detector's output probabilities $p = (p_1, \ldots, p_n)$, changing the hidden state as follows: $h'_i = h_i + p_i \cdot v$, where h_i is the original hidden state and p_i is the probability that the i-th word is problematic. This is done at every timestep, ensuring the bias correction is consistently influenced by the detector's evaluations.

The decoder uses these modified hidden states $H' = (h'_1, \ldots, h'_n)$ to generate text that is neutral and fits the context. The vector v in the hidden states helps the decoder identify which words are biased or subjective, guiding it on what to change or keep the same. This model setup improves our ability to neutralize problematic language while keeping the original text's meaning and readability intact. To evaluate our model's performance, we use the debiaser model by Pryzant as a baseline. This model is designed to reduce overall subjective bias in text and is a solid benchmark due to its effectiveness, although its performance on non-western English dialects is uncertain. Using this debiaser allows us to measure how well our adapted model neutralizes negatively biased language, slurs, in Nigerian English and Pidgin. The original model's general method of reducing subjective biases serves as a useful comparison, helping us demonstrate the specific improvements and adjustments our model makes to address cultural and linguistic nuances in our target audience.

5 Experiments

5.1 Data

We constructed a specialized corpus of 1,000 entries, each meticulously crafted to represent the nuanced language use within Nigerian Pidgin English. This dataset is not merely a collection of phrases but a carefully curated representation of biased expressions alongside their neutralized counterparts. It also includes non-biased phrases which serve as negative samples for our binary classification task, essential for training our model to discern between biased and non-biased language accurately.[1]

The entries are categorized into several subcategories based on the nature of the bias they represent. These subcategories include 'Sexual', for slurs related to gender or sexual orientation; 'Intellectual', for derogatory terms that demean an individual's intelligence; 'Physical', for biased expressions concerning physical abilities or disabilities; and 'Social', for slurs that target an individual's social behavior or status. Each category was developed to help the model understand and adjust to the context-specific nuances of slurs, which are often deeply embedded in cultural expressions. In our corpus, the training examples were distributed as follows: Sexual - 64, Intellectual - 185, Physical - 171, Social - 248, with the remaining sentences classified as acceptable or non-offensive. To ensure the relevance and accuracy of our corpus, we engaged with native speakers of Yoruba, Igbo, and Hausa—three major ethnic groups in Nigeria. Their insights were invaluable in ensuring that the corpus reflects the true linguistic diversity and cultural depth of Nigerian Pidgin English. This collaboration not only enhanced the quality of our dataset but also enriched our understanding of the cultural underpinnings that influence language use in different Nigerian communities.

[1] `https://docs.google.com/spreadsheets/d/1kSqYZRGUuX1nOHOtM9n-Wjm5x6iBA1mG/edit?usp=sharing&ouid=105140934525773378825&rtpof=true&sd=true`

Table 1: Sample Entries from the Nigerian Pidgin English Bias Neutralization Corpus

Source	Target	Subcategory of Nigerian Slur
The woman na ashawo	The woman na get different company	Sexual
That man na complete mumu	That man no too understand things	Intellectual
She no fit work cause she be cripple	She no fit work cause she dey use wheelchair	Physical
Why you like amebo like this	Why you like gossip like this	Social
You no dey use your head at all	You no dey use your head at all	Ok

5.2 Evaluation method

Our evaluation strategy includes both quantitative and qualitative methods:

- **Bag of Words Classifier**: We employ a simple logistic regression model to categorize sentences as either biased or non-biased. This binary classifier is trained on a Bag of Words representation of our corpus, where the dataset is split into 75% for training and 25% for testing. The effectiveness of this classifier is crucial for assessing our model's capability to accurately identify biased language.
- **Naive Bayes Classifier**: In addition to our logistic regression approach, we also explore a Naive Bayes classifier to further assess the identification of biased language. This model, like the Bag of Words classifier, utilizes a Count Vectorizer to transform the textual data into a format suitable for machine learning. The Naive Bayes model is particularly chosen for its proficiency in handling categorical input variables and its effectiveness in text classification tasks. We trained the Multinomial Naive Bayes classifier using a similar data split—75% for training and 25% for testing. The performance of this classifier provides us with an additional perspective on the robustness of our approach in identifying and categorizing biased phrases within our corpus.
- **Human Evaluation**: We further conduct qualitative evaluations where human annotators assess the quality of the neutralized phrases compared to the original biased phrases. This step is essential to ensure that the changes made by our model maintain the semantic integrity and appropriateness of the original text.

5.3 Experimental details

From our specialized corpus, we categorized 1,000 items into training and test groups, maintaining a 75-25% distribution. This led to 750 sentence pairs designated for training and 250 for testing. Following the model training procedures outlined by Pryzant our adaptation involved fine-tuning a pre-trained BERT model, specifically configured for our task of identifying and neutralizing biased phrases in Nigerian Pidgin English. We executed this fine-tuning using PyTorch and employed the Adam optimizer, setting the learning rate at 5×10^{-5}. The setup for our model included a batch size of 16 and vector lengths fixed at $h = 512$. We also incorporated gradient clipping, capping it at a norm of 3, and introduced a dropout rate of 0.2 at the input stage of each LSTM cell. We initialized the BERT segment of our tagging system using the bert-base-uncased parameters.

5.4 Analysis & Results

The performance comparison between the baseline and the fine-tuned models is summarized in the table below:

Table 2: Performance Comparison of Baseline and Fine-Tuned Models on Slur Neutralization

Result Type	Baseline	Fine tuned
Correctly Identified & Reduced Slur; Grammatically Correct	0.08	0.29
Correctly Identified & Reduced Slur; Grammatically Incorrect	0.05	0.17
Correctly Identified Slur, but Wrong Type	0.03	0.09
Didn't Identify Slur	0.64	0.28
Total Slur Reduced	0.16	0.55

From the table, it is evident that the fine-tuned model significantly outperforms the baseline in all categories, particularly in correctly identifying and reducing slurs in a grammatically correct manner. The total effectiveness in reducing slurs shows a substantial improvement from 16% to 55%. In comparison to the Bag of Words classifier, which identified slurs or offensive context 71% of the time, with a precision of 0.74 and a recall of 0.87 for the offensive class ('1'), and a precision of 0.59 and a recall of 0.39 for the non-offensive class ('0'), the fine-tuned BERT model showed a slightly lower identification rate. The Bag of Words classifier's high performance can likely be attributed to its ability to pick up on specific keywords strongly correlated with offensive or derogatory content. However, this method lacks the sophistication to reconstruct or neutralize the identified slurs into more acceptable phrases, a key capability of our fine-tuned BERT model.

The Naive Bayes classifier, trained on the same dataset, demonstrated an overall accuracy of 76% with precision and recall varying significantly between the classes. For the non-offensive class ('0'), the model achieved a precision of 0.65 and a recall of 0.53, whereas for the offensive class ('1'), the precision was notably higher at 0.80 with a recall of 0.87. This discrepancy highlights the model's strength in confidently identifying offensive content while struggling slightly with false negatives in the non-offensive class. Similar to our Bag of Words Classifier, the Naive Bayes classifier lacks the nuance of the BERT model, which is better equipped to understand context and generate non-offensive alternatives in text, providing a more sophisticated approach to text transformation. However, its higher identification rate for offensive content compared to the BERT model could be attributed to its statistical approach to classification. The Naive Bayes model's ability to perform well in identifying specific patterns or keywords commonly associated with offensive language may result in higher precision for these instances. Unlike BERT, it does not require as deep an understanding of context, which can sometimes lead to overfitting or misinterpretations in more complex or nuanced scenarios.

The relatively lower performance of the BERT model in some aspects could be due to the complexity of the task at hand. Neutralizing slurs in Nigerian Pidgin English requires not only identifying offensive words but also understanding the context to generate appropriate, non-offensive alternatives. This task demands a deep understanding of both the language structure and cultural context, areas where even advanced models like BERT can struggle without extensive and well-curated training data.

To understand the model's performance on a more granular level, we reviewed specific examples where the model attempted to neutralize slurs. Below is a table illustrating selected phrases and the model's responses:

Table 3: Slur Identification and Neutralisation Table

Source Phrase	**Ideal Phrase**	**Fine Tuned Phrase**	**Result Type**
Na oloshi like you dey spoil things for everybody	Na unfortunate person like you dey spoil things for everybody	Na annoying person like you dey spoil things for everybody	Correctly Identified & Reduced Slur; Grammatically
You be utu!	You be not smart!	You be not understanding!	Correctly Identified & Reduced Slur; Grammatically Incorrect
You no fit keep secret, you be aproko	You no fit keep secret, you have a loose mouth	You no fit keep secret, you be improper person	Correctly Identified slur, but wrong type
You dey ment?	You dey ment?	You dey ment?	Didn't Identify Slur

The table indicates varying degrees of success. While the model generally improves phrase neutrality, there are instances of grammatical inaccuracy and misclassification of the type of slur, highlighting areas for future improvement. Moreover, our analysis reveals that the model performs better in identifying and neutralizing 'Sexual' and 'Intellectual' slurs compared to 'Physical' and 'Social' slurs. This variation could be attributed to the frequency and clarity of these slurs within the training data, which might have provided the model with more examples and clearer patterns to learn from. Sexual and Intellectual slurs are a lot easier to identify with sexual/intellectual slurs like "ashawo", "kolo", "ode", "mumu" and "olodo" being frequently used and generally solely used in negative contexts. In contrast, "Physical" and "Social" slurs, which might be expressed in more varied or subtle ways, presented greater challenges for the model.

Table 4: FineTune Model's Performance on Various Subcategories of Slurs in Nigerian Pidgin English

Slur Type	**Reduced Correctly; Grammatically Correct**	**Reduced Correctly; Not Grammatically Correct**	**Incorrect Reduction; Identified Incorrect Type**	**Did Not Identify Slur**
Sexual	0.35	0.25	0.05	0.10
Intellectual	0.33	0.22	0.07	0.12
Physical	0.28	0.18	0.12	0.20
Social	0.25	0.15	0.10	0.25

6 Conclusion

This project has successfully demonstrated the potential of fine-tuned BERT models in addressing the underrepresentation of non-Western dialects in NLP research, particularly through our focus on Nigerian Pidgin English. By developing and utilizing a specialized corpus to identify and neutralize slurs, we have made significant strides in enhancing the inclusivity and cultural sensitivity of NLP applications. Our findings reveal that with targeted fine-tuning, BERT models can effectively adapt to the linguistic nuances of less-studied languages, achieving notable success in reducing negatively biased language while maintaining the grammatical integrity of the text. The accuracy of bias detection and the model's ability to handle the diverse expressions of Nigerian Pidgin English highlight areas where further refinement is needed. Additionally, the challenges of ensuring that the model does not inadvertently suppress cultural expressions or censor legitimate speech pose ongoing ethical considerations. Looking forward, we hope to continue to explore the application of similar models to other non-Western languages and dialects, broadening the reach and relevance of NLP technologies and bringing more equity of corpuses and models to this field of research

7 Ethics Statement

For our model and expirment we identified some potential ethical concerns. Firstly, the risk of overgeneralization and misinterpretation of cultural nuances is significant. Nigerian Pidgin English, like any dialect, contains unique expressions and idiomatic phrases that may carry different connotations within various contexts. A model trained to neutralize biased language might inadvertently alter expressions that are culturally significant or meaningful to the local communities, thereby sanitizing linguistic diversity in the name of bias neutralization.

Another ethical concern is the potential for censorship. By altering text to remove biases, there is a risk that the software could be used to suppress freedom of expression, especially in sensitive or contentious contexts. This could be particularly problematic in scenarios where language is used to convey dissent or criticism. To mitigate these risks, it is crucial to involve local linguistic experts and community representatives in the training and continuous evaluation process of the model. This inclusion ensures that the model's interpretations and modifications of Pidgin expressions adhere to cultural sensitivities and linguistic accuracy. Furthermore, developing a transparent algorithm that allows users to see what changes have been made and why can help mitigate concerns about censorship. This transparency not only helps in building trust with the users but also allows them to make informed decisions about accepting or rejecting suggested changes. Another mitigation strategy we thought of involves implementing strict usage guidelines and ethical usage policies that govern how and where the technology can be applied. Establishing clear boundaries and purposes for the use of this NLP technology can help prevent its application in ways that could curtail free speech or lead to cultural homogenization. These guidelines should be developed in collaboration with policymakers, local communities, and ethicists to ensure they are comprehensive and culturally informed.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Faraclas. 1996. *Nigerian Pidgin*. Routledge.

Kelechukwu Ihemere. 2006. *A Tri-Generational Study of Language Choice & Shift in Port Harcourt*. Universal-Publishers.

Aditya Joshi, Pushpak Bhattacharyya, and Mark James Carman. 2020. Falling through the cracks: The pitfalls of gaps in data for non-western languages in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 589–598.

Reid Pryzant, Richard Diehl Martinez, Nathan Dass, Sadao Kurohashi, Dan Jurafsky, and Diyi Yang. 2020. Automatically neutralizing subjective bias in text. *CoRR*, abs/1911.09709.

Victor Soto and Julia Hirschberg. 2018. A short review of ethical challenges in language technology. In *Proceedings of the Association for Computational Linguistics (ACL) Ethics in NLP Workshop*, pages 30–40.

Manuel Tonneau, Pedro Vitor Quinta de Castro, Karim Lasri, Ibrahim Farouq, Lakshminarayanan Subramanian, Victor Orozco-Olvera, and Samuel Fraiberger. 2024. Naijahate: Evaluating hate speech detection on nigerian twitter using representative data.