Understanding Pidgin English Through NLP: Translation, Sentiment, and Social Media Discourse

Oluwatomiwa Baruwa, Oluwatobi Olajide, Abimbola Olorogun, Iyinoluwa Ayodele

Summer 2025

1 Introduction

Pidgin English, encompassing various regional forms such as Nigerian Pidgin, Cameroonian Pidgin, Tok Pisin (Papua New Guinea), and Solomon Islands Pijin, functions as a crucial lingua franca in multilingual societies across Africa, the Caribbean, and the Pacific. These languages, born out of historical trade, colonization, and cultural exchange, facilitate communication among diverse linguistic groups. Despite their widespread use, Pidgin languages have historically been marginalized in formal education, media, and technological applications. The advent of Natural Language Processing (NLP) offers an opportunity to bridge this gap. By developing NLP tools tailored to Pidgin languages, we can enhance machine translation, sentiment analysis, and the understanding of social media discourse in these languages. This is particularly important as Pidgin languages are extensively used on social media platforms, where they serve as vehicles for cultural expression, political discourse, and community engagement. Recent studies have begun to address these challenges. For instance, transformer-based models have been developed to generate coherent Nigerian Pidgin text, leveraging datasets like the Afriberta corpus. These models demonstrate the potential for high-quality text generation in low-resource languages. By advancing NLP applications for Pidgin English, we not only promote linguistic inclusivity but also empower communities through improved access to technology and information. This research aims to bridge the gap between Pidgin English and NLP, focusing on translation, sentiment analysis, and social media discourse to enhance communication and understanding in linguistically diverse regions.

2 Research Context and Problem Statement

Nigerian Pidgin English is a widely spoken creole language used in West Africa, particularly in Nigeria, where it serves as a primary mode of informal communication for millions of people. It appears in everyday conversations, music,

political commentary, and increasingly in social media discourse. Despite its prevalence and sociolinguistic richness, Pidgin is often excluded from formal systems of education, governance, and especially digital language technologies such as machine translation, sentiment analysis, and chatbots.

The exclusion of Pidgin from mainstream Natural Language Processing (NLP) models stems largely from its classification as a low-resource language. Unlike standardized languages with large corpora and formal grammar rules, Pidgin exhibits significant variation in spelling, syntax, and vocabulary, making it challenging to process using conventional NLP techniques. Most existing models, trained primarily on high-resource languages, are unable to accurately interpret or generate Pidgin text. This gap has real-world implications, such as misclassified sentiment, mistranslations, and the digital marginalization of Pidgin-speaking communities.

Previous work has taken important steps toward addressing this issue. The *NaijaSenti* corpus (Abdul-Mageed et al., 2022) has improved the sentiment analysis for Nigerian languages, including Pidgin, offering manually annotated social media data. However, its primary focus is on emotion classification, not translation or discourse analysis.

Projects like *Masakhane* (Nekoto et al., 2020) are foundational, but do not adequately address the dynamic and informal nature of Pidgin as used in digital spaces today. The formal tone and domain mismatch of existing corpora lead to weak generalization when models are applied to social media data. Moreover, the emotional expressiveness and cultural context embedded in Pidgin are often lost in current sentiment models, which rely on lexicons developed for other languages.

This research aims to bridge these gaps by constructing a custom dataset of real-world Nigerian Pidgin-English text pairs sourced from social media platforms, with the goal of fine-tuning multilingual models for improved translation and sentiment analysis. In addition, the study will examine how Pidgin is used to express identity, humor, protest, and community in online conversations. The broader objective is to develop more inclusive and representative NLP tools that support equitable digital access for Pidgin-speaking communities.

Research Problem: Despite growing interest in African languages in NLP, existing models and data sets are not sufficient to handle the informal, expressive, and culturally specific nature of Nigerian Pidgin English. This project addresses the problem of building accurate and culturally aware NLP tools - particularly for translation and sentiment analysis - by developing a custom, high-quality parallel corpus of Pidgin-English social media texts and adapting state-of-the-art models to better handle this underrepresented language.

3 Literature Review

Natural Language Processing (NLP) for low-resource languages has gained momentum in recent years, yet Pidgin English remains significantly underrepresented. A few foundational efforts provide groundwork for this study. The Nai-

jaSenti corpus (Muhammad et al., 2022) is a large-scale, manually annotated dataset focused on sentiment analysis in Nigerian languages, including Pidgin. With approximately 14,000 annotated Pidgin tweets, it enables improved sentiment modeling, especially when paired with adapted sentiment lexicons like the 300-token VADER enrichment developed by Oyewusi et al. (2020). While effective for emotion classification, these datasets are not designed for translation. For machine translation, the Masakhane NMT project has been instrumental in pioneering African language modeling using neural networks. Their work includes experiments with multilingual transformers such as mBART and mT5 on Pidgin, but the available parallel corpora remains limited in both size and domain variety. The JW300 corpus offers aligned religious text pairs between Pidgin and English, largely derived from Bible translations. While useful for structure and syntax alignment, its formal and spiritual tone differs drastically from casual, internet-based Pidgin used in everyday discourse. Additionally, the AfriBERTa and NLLB-200 models demonstrate that large-scale multilingual pretraining can include underrepresented languages like Pidgin, but such models are rarely fine-tuned on domain-specific content like social media posts, where most real-world Pidgin is now used. This research distinguishes itself by targeting modern, real-world Pidgin, particularly from social media platforms, to construct a custom parallel corpus and fine-tune existing NMT models for improved bidirectional translation. In doing so, it fills a critical gap between available structured datasets and informal online language use, while opening pathways for further discourse analysis and application development.

4 Expected Outcomes

Development of Pidgin-English NLP Tools: Creation or adaptation of Natural Language Processing models specifically trained to process, understand, and generate Nigerian Pidgin English, including tools for translation, sentiment analysis, and text classification.

Improved machine translation accuracy: Enhanced **bilingual (Pidgin–English)** translation systems that better handle the **grammatical structures**, **idioms**, and **contextual meanings** unique to Pidgin English.

Sentiment analysis framework for Pidgin: A robust **sentiment analysis model** capable of accurately interpreting **emotional tones** and **opinions** expressed in Pidgin across **social media platforms**.

Annotated Pidgin-English dataset: Compilation and release of a curated, **linguistically diverse dataset** of Pidgin English text from **social media** and other digital platforms, **annotated** for both **translation** and **sentiment analysis** tasks.

Insights into online social discourse: Analysis of how Pidgin English is used in online conversations to express **cultural identity**, **political opinions**, **humor**, **resistance**, and **solidarity**, offering **sociolinguistic insights**.

Contribution to inclusive AI: Advancement in **ethical** and **inclusive AI** by expanding NLP capabilities to **underrepresented languages and dialects**, aligning with global efforts to **democratize language technologies**.

5 Timeline

July 2025 – Data Collection and Preparation

- Scrape Pidgin-English text pairs from social media platforms (Twitter, Reddit, Facebook)
- Clean and normalize text for consistent tokenization
- Begin manual annotation of translation and sentiment labels

August 2025 – Dataset Finalization and Initial Modeling

- Finalize annotated dataset for translation and sentiment tasks
- Fine-tune multilingual models (e.g., mBART, mT5) on custom parallel corpus
- Evaluate translation accuracy using BLEU and other metrics

September 2025 – Sentiment Model Development and Discourse Analysis

- Train sentiment analysis models using NaijaSenti and enriched lexicons
- Evaluate performance using F1 score and confusion matrices
- Begin qualitative analysis of Pidgin in online discourse (identity, humor, protest, etc.)

October 2025 – Tool Integration and Application Prototyping

- Build lightweight NLP tools (translation, sentiment, tagging) for browser or mobile use
- Develop simple web interface for live text input and model predictions
- Integrate Pidgin NLP tools with social media monitoring or education platforms

November 2025 – Final Evaluation and Report Writing

- Conduct end-to-end evaluation of system performance on real-world test data
- Refine models and documentation for release
- Write and finalize project report and academic paper for submission