Data Sourcing and Preprocessing Report

1. Problem Statement

Nigerian Pidgin is one of the most widely spoken languages in West Africa, used daily by millions. However, it remains underrepresented in modern NLP applications due to a lack of clean, structured, and large-scale datasets.

This project addresses that gap by preparing a high-quality dataset for Pidgin-to-English translation, enabling downstream tasks such as:

- Machine Translation
- Text-to-Speech (TTS)

2. The Problem

We are solving the problem of data scarcity and inconsistency in low-resource languages like Nigerian Pidgin. Without proper preprocessing, noisy or duplicated data can significantly degrade the performance of NLP or TTS models.

3. Dataset Overview

For this project, the dataset used was:

Tommy0201/pidgin-to-English (Hugging Face)

It includes 116,331 Pidgin-English sentence pairs.

Key Stats:

- Total entries: 116,331
- Unique Pidgin entries: 116,315

- Unique English entries: 115,745

- Most frequent English sentence: "Where did this photo come from?" (86 times)

Observations:

- Minimal duplication in Pidgin

- Some repetition in English

- These patterns highlight the importance of cleaning and normalization.

4. Objective

To build a clean, consistent, and well-structured Pidgin-English dataset for use in NLP tasks by:

- Removing empty, invalid, or duplicate records

- Normalizing text (e.g., punctuation, casing)

- Ensuring quality and consistency for model training

5. Data Sourcing Steps

The dataset was sourced directly from Hugging Face using the Python datasets library:

```
from datasets import load_dataset
dataset = load_dataset("Tommy0201/pidgin-to-english")
```

Initial inspection included:

- Reviewing dataset metadata

- Previewing the top 5 rows

- Checking the dataset schema and number of entries

6. Data Preprocessing

6.1 Initial Exploration

- Confirmed shape and structure

- No null values found

- Counted words and sentences

- Viewed common sentence patterns

6.2 Data Cleaning

Removed samples that:

- Lacked a translation field

- Had missing or empty text

- Contained non-string values

6.3 Normalization

- Lowercased all text

- Removed special characters (except apostrophes)

- Trimmed whitespace

7. Challenges and Observations

7.1 Data Sourcing Difficulties

- Access and integration into Colab required careful setup

- Dataset structure took time to interpret

7.2 Cleaning Challenges

- Over-filtering initially removed all data

- Regex for punctuation required fine-tuning

- Edge cases with non-string or missing fields

7.3 Observations

- High diversity in Pidgin (low duplication)

- Common phrases repeated in English

- Dataset had no missing values