Natural Language Processing of a Low-Resource Language (Igbo, one of African Languages)
NLP Of Low Resource Language

Principal Investigator
Dr. Stanley Chinedum Nwoji
Harrisburg University of Science and Technology

Mentor
Dr. Iheb Abdellatif
Harrisburg University of Science and Technology

Student Investigator
Atajan Abdyyev
Harrisburg University of Science and Technology
Though there are only 20 languages that fall into the high-resource category, most natural language processing (NLP) advancements have been accomplished in these 20 languages, excluding thousands of the low-resource languages spoken by millions of people in the world. Our goal is to narrow down the NLP gap for Igbo a low-resource category language.
Problem

Spoken Languages in West Africa

- Low Resource: 99.7%
- High Resource: 0.3%

West Africa Population

- Population: 400,000,000

Igbo Speaking Population

- Population: 10,000,000
The Igbo language is one of the major languages spoken in Nigeria, but there is a lack of comprehensive language corpora available for this language. This presents a significant challenge for natural language processing (NLP) applications, including machine translation, and text-to-speech synthesis, which all rely on large amounts of high-quality language data.
Importance of the project: This project aims to address the lack of Igbo language corpora by creating a comprehensive corpus of text data in the Igbo language. This corpus will be an invaluable resource for researchers, language learners, and developers of NLP applications, enabling them to build more accurate and effective language models that can handle Igbo text.
Solution provided by the project: By creating a high-quality corpus of Igbo language data, this project will help to preserve and promote the use of the Igbo language, while also enabling the development of new and innovative NLP applications that can support Igbo speakers and learners. The project will also contribute to the broader goal of promoting linguistic diversity and cultural preservation through the use of technology.
This project is divided into six milestones.

1. Milestone #1: Development of the Igbo Corpora consisting of News Content from BBC Igbo, Nigerian Television Authority, etc.
2. Milestone #2: Cleaning of the Igbo Corpora
4. Milestone #4: Text Categorization Using the Igbo Corpora
5. Milestone #5: Information Extraction Using the Igbo Corpora
6. Milestone #6: Machine Translation Using the Igbo Corpora, Give a Wrap-up presentation and finalize project
• What I hope to learn
  – NLP Techniques for low resource languages
  – Machine Learning applications
  – Data Management and Storage
  – Understand languages better
• **Goals for Next Month**
  – Create Scripts to collect text and convert to raw Corpora for IGBO Language
  – Clean the Corpora
  – Attempt initial ML-NLP analyses
• Help needed
  – To be discovered as we dive deeper in the project
Thank you
Any Question?