**CHAPTER ONE**

1. **Introduction**
2. **Background**

The language technology is crucial today to ensure access to information and opportunities for economic development. With about two thousand different languages, Africa is a multilingual continent. It presents significant challenges for researchers who seek to promote and use African languages in the areas of business, development, education and humanitarian. The Natural Language Processing (NLP) allows these challenges in the creation of application that use languages.

Language identification is an important area of research in natural language processing (NLP) amongst other components such as machine translation, text-to-speech synthesis, and speech recognition. It involves automatically determining the language of a given text, which can be useful in various applications such as machine translation, text classification, and information retrieval. With the increasing amount of multilingual data available on the web, the need for accurate and efficient language identification systems has become more important than ever. However, for many African languages, including Yoruba, language identification is still a challenging task due to the lack of resources and research in the field.

The Yoruba language is a widely spoken language in Nigeria, West Africa, with approximately 25 million speakers. It is a tonal language with a rich linguistic and cultural heritage. As a result of its popularity, there is a need to develop an automated language identifier that can accurately detect and identify Yoruba text in digital documents.

Despite its importance and widespread use, there is a scarcity of computational resources and tools for Yoruba language processing. One of the first steps towards developing such resources is to design and develop an accurate language identifier for Yoruba. Yoruba language identification is an important task that can facilitate the development of NLP applications for the Yoruba-speaking community.

**1.2** **Problem Statement**

Currently, only a few tools and programs can reliably identify the language that a given text is written in. The lack of language identification systems fully dedicated to a indigenous Nigerian language is another reason behind the study.

Occasionally, researchers would come upon a certain text and be forced to go through a physically taxing process to determine what language it was.

The lack of precise identifiers frequently results in inaccurate translations because translators typically need to determine what text a language is in in order to translate it.

This particular deficiency also has an impact on the technology sector, in terms of language translators.

By using the Yoruba language as a point of contact, this study seeks to offer a viable language identification method that might assist with these problems.

* 1. **Scope and Limitations of Study**

The study will use a corpus of Yoruba texts collected from various online sources, which may not be representative of all genres and styles of texts in these languages. The study may also be limited by the availability of computational resources and the quality of the machine learning models used for language identification. The performance of the Yoruba language identifier will be evaluated using standard evaluation metrics and compared with existing language identification tools.

**1.4 Aim and Objectives of the Study**

The aim of this project is to propose the design of a system that can be used to identify Yoruba language through text.

The specific objectives of the study are:-

* To review existing literature on language identification and related fields.
* To collect and preprocess a corpus of texts in Yoruba and other languages.
* To design and implement a language identifier using machine learning techniques.
* To evaluate the performance of the language identifier using appropriate metrics and compare it with existing language identification systems.
* To deploy to a web framework.

**1.5 Significance of the Study:**

The significance of this study lies in the fact that it addresses an important problem in NLP, particularly in the context of Yoruba language processing. The design and identification of a Yoruba language identifier will enable researchers, developers, and other stakeholders to accurately identify the language of a given text or phrase, which can be useful in various applications such as machine translation, text classification, and information retrieval. Furthermore, the study will contribute to the development of computational resources and tools for Yoruba language processing, which can have a positive impact on language preservation, education, and communication.

CHAPTER TWO

**LITERATURE REVIEW**

**2.1 Introduction**

This chapter reviews literature on language identification, natural language processing techniques, and the Yoruba language. The literature review aims to provide a background to the study and to identify the gaps in the current state of research.

* 1. **Language Identification**

Language identification is the process of determining the language in which a given text is written. It is a crucial component in natural language processing (NLP) applications such as machine translation, text-to-speech synthesis, and speech recognition. Several approaches have been proposed for language identification, including statistical methods, rule-based methods, and machine learning methods. Statistical methods use language-specific features such as character frequency and n-gram frequency to identify the language. Rule-based methods use a set of rules based on linguistic knowledge to identify the language. Machine learning methods use a training set of labeled data to learn a model that can accurately identify the language.

* 1. **Natural Language Processing Techniques**

Natural language processing (NLP) is a subfield of computer science and artificial intelligence that deals with the interaction between computers and human language. NLP techniques are used in several applications such as text classification, sentiment analysis, machine translation, and speech recognition. Some of the commonly used NLP techniques include tokenization, part-of-speech tagging, named entity recognition, and sentiment analysis.

* 1. **The Yoruba Language**

Yoruba is one of the twelve (12) languages of the Edekiri sub-branch from the great family of Niger-Congo. It is natively spoken in southwestern part of Nigeria (the second largest ethnic group in number), in Benin and Togo by over 30 million people. Since 2011, Yoruba has a unique alphabet containing thirty (30) letters composed of 12 vowels: i, e, ẹ, a, o, ọ, u, in, n, un, n, an, and 18 consonants: b, d, f, g, gb, h, j, k, l, m, n, p, r, s, ṣ, t, w, y. Yoruba is a tonal language. It has three (03) tones: high tone (represented by the acute accent), the low tone (represented by the grave accent) and the average tone (represented by the absence of accent).

The Yoruba language is a language spoken in Nigeria and other West African countries. It is a tonal language with several dialects, which means that the meaning of the word can change depending on the tone used; and it has a rich literary tradition. The Yoruba language is written using the Latin alphabet, and it has a complex orthography with several diacritics.

Additionally, Yoruba has a large number of loanwords from other languages, such as English, Portuguese, and French, which can affect the performance of language identification systems

Methods for Yoruba Language Identification: Several methods have been proposed for Yoruba language identification, including traditional machine learning approaches and deep learning approaches.

1. Traditional Machine Learning Approaches: Traditional machine learning approaches for Yoruba language identification include Naive Bayes, Support Vector Machines (SVM), and K-Nearest Neighbors (KNN). These methods rely on hand-crafted features, such as n-grams and character n-grams, to represent the language of a given text.
2. Deep Learning Approaches: Deep learning approaches for Yoruba language identification include Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). These methods automatically extract features from the input text, which can be more effective than hand-crafted features.
   1. **Related Work**

Several studies have been conducted on language identification, natural language processing, and the Yoruba language. Ojo et al. (2015) developed a Yoruba part-of-speech tagger using a rule-based approach. Ojo et al. (2018) developed a Yoruba speech recognition system using a Hidden Markov Model. Adegbola et al. (2020) developed a Yoruba named entity recognition system using a machine learning approach.

* 1. **Literature Review**

Although several studies have been conducted on language identification, natural language processing, and the Yoruba language, there is a need to develop a Yoruba language identifier that can accurately detect and identify Yoruba text in digital documents. The existing language identification tools do not accurately detect Yoruba text, leading to incorrect processing and analysis of the text. Therefore, there is a need to design and develop a Yoruba language identifier using natural language processing techniques.

**Chapter Three: Methodology**

3.1 Introduction

This chapter describes the methodology used to design and develop the Yoruba language identifier. The methodology involves four main stages: data collection and preprocessing, feature extraction, model selection, and model evaluation.

3.2 Data Collection and Analysis

The first phase of the methodology involves collecting a dataset of Yoruba text. The dataset is used for training and testing the language identification model. Dataset of Yoruba and English text was collected from various sources, including news articles, dictionary, and online forums and GitHub repositories. The dataset contains a total of 2,306 documents, with Yoruba dataset containing an average of 200 words.

3.3 Feature Extraction

The second phase of the methodology involves extracting features from the dataset. The features are used to represent the text in a machine-readable format that can be used for training the language identification model. We used a combination of statistical and linguistic features for feature extraction. The statistical features include character frequency, n-gram frequency, and sentence length. The linguistic features include part-of-speech tags, named entities, and syntactic dependencies.

3.4 Model Development

The third phase of the methodology involves developing the language identification model. We used a supervised machine learning approach for model development. We split the dataset into training and testing sets in an 80:20 ratio. We trained several machine learning algorithms on the training set, including logistic regression, support vector machines, and decision trees. We evaluated the performance of the models on the testing set using metrics such as accuracy, precision, recall, and F1-score. We used the scikit-learn library in Python for model development and evaluation.

3.5 Model Optimization

After evaluating the performance of the models, we selected the best performing model for further optimization. We used grid search cross-validation to tune the hyperparameters of the selected model. We also experimented with different feature sets to improve the performance of the model.

3.6 Model Evaluation

The final phase of the methodology involves evaluating the performance of the optimized model on a new dataset of Yoruba text. We collected a new dataset of 2,000 Yoruba documents from different sources and used it to evaluate the performance of the model. We used the same metrics as in the testing phase to evaluate the performance of the model.

3.7 Summary

This chapter has described the methodology used in developing the Yoruba language identifier. The methodology involves collecting a dataset of Yoruba text, extracting features from the dataset, developing a language identification model using supervised machine learning, optimizing the model, and evaluating its performance on a new dataset. The next chapter will present the results of the study.

**REFERENCES**

1. *Wikipedia*, Language as an effective tool for communication.
2. Research Journal of Computer and Information Technology Sciences - **Vol. 1(1), 2-5, February (2013) -** Part-of-Speech tagging of Yoruba Standard, Language of Niger-Congo family