Fine-Tuning NLP tools in two low-resourced languages: Nigerian Pidgin & Igbo

1. Abstract

This research proposal aims to address the lack of robust Natural Language Processing (NLP) tools tailored for low-resourced languages, specifically Igbo and Nigerian Pidgin. Despite the growing interest in NLP, many languages are often overlooked due to their low resource availability. This proposal outlines a method to fine-tune existing NLP models for text classification, named entity recognition (NER), and translation tasks in Igbo and Nigerian Pidgin. The methodology involves collecting and curating datasets, fine-tuning popular NLP models, and evaluating the performance improvements compared to baseline models. By improving the efficiency of NLP tools in these languages, the research aims to contribute to the development of more inclusive and equitable NLP technologies.

2. Introduction

Natural Language Processing (NLP) is a subfield of Artificial Intelligence. It deals with communication between humans and computers through natural human languages, enabling computers to understand and respond to words or statements in a desired manner. In recent years, NLP has witnessed remarkable advancements, revolutionizing various fields such as machine translation, spam detection, information extraction, and most recently, AI chatbots. However, the developments benefited languages with abundant linguistic resources - high-resourced languages (HRLs) - leaving low-resourced languages (LRLs) at a disadvantage. Trends reveal that large language models such as GPT, when compared with traditional Machine Translation (MT) models, perform equally well in HRLs but are "especially disadvantaged" when communicating in LRLs. The disparity in NLP development among languages stems from the unequal distribution of linguistic resources and research effort. For example, while languages like English and French receive abundant attention and research, languages such as Igbo and Nigerian Pidgin often lack adequate datasets, pre-trained models, and specialized tools. This difference not only hinders technological progress but also exacerbates linguistic inequalities on a global scale.

To address this gap, this research focuses on fine-tuning existing NLP models to improve their performance in text translation, text classification, and NER in Igbo and Nigerian Pidgin languages. Self-collected datasets and those collected from credible sources will be used for fine-tuning. Finally, evaluations will be performed to measure the improvements of fine-tuned models compared to their baseline models.

3. Methods

Initially, data acquisition will involve making use of the formerly curated datasets from our previous research and having them translated from Igbo/Nigerian Pidgin to English by native speakers, as well as utilizing datasets from well-known and credible platforms such as GitHub and Hugging Face. The collected data sets will be preprocessed and tailored based on the intended NLP task of the model to ensure quality and consistency. Alongside data preprocessing, we will research state-of-the-art NLP models suitable for fine-tuning, with particular emphasis

on models built for text translation, text classification, and/or NER tasks. Subsequently, the fine-tuning process will take place for several weeks, with constant modification of training algorithms to optimize model performance. Finally, comprehensive testing and evaluation are conducted to gauge the effectiveness of the fine-tuned models against that of baseline models.

4. Timeline

Week 1: Data collecting/ Data translation from HITT workers

Week 2: Data collecting/ Data translation from HITT workers

Week 3: Data preprocessing. Research NLP models

Week 4: Fine-tuning translation models

Week 5: Fine-tuning translation models

Week 6: Fine-tuning models for text classification and NER

Week 7: Fine-tuning models for text classification and NER

Week 8: Test and compare models

Week 9: Test and compare models; Start writing final report

Week 10: Draw conclusions; Finished final report

5. Broader Impacts

With the constant development and implementation of AI, it is essential for developers and researchers to have robust NLP tools. This ensures that new technologies are not biased towards widely used languages, thus avoiding the neglect of other languages and providing a significant disadvantage in developing similar technologies for different languages.

By focusing on low-resourced languages such as Igbo and Nigerian Pidgin, this research not only provides more efficient NLP tools to support future AI research in these languages but also contributes to linguistic equality within the digital landscape. Equalities in technological advancements can help avoid greater inequalities among countries in terms of economic, environmental, and social development. The methodologies and insights gained from this research will pave the way for advancements in AI, particularly in the domain of language processing for underrepresented languages.

In summary, this research has far-reaching implications across various sectors, underscoring the significance of linguistic diversity and fair technological advancement in an interconnected world.

6. Trajectory/Future Goals

With an interest in NLP and data science, I hope to apply the skills and knowledge learnt to continue future research in the field of NLP. I envision myself working on projects like developing AI chatbots, like ChatGPT, but tailored specifically for low-resourced languages. Furthermore, this research journey will both lay a strong foundation for me to potentially purse