**Personal Statement**

As a child, I looked up to Albert Einstein as my role model, inspired by his groundbreaking inventions that held the power to transform the world and impact countless lives. His genius and innovations seemed to rival those of a superhero, igniting within me a desire to someday invent something equally impactful.

During my secondary school years, I discovered my aptitude for logical thinking, critical reasoning, and organization. Naturally drawn to subjects like math, physics, and chemistry, I immersed myself in national and international academic competitions, earning several awards along the way. Recognizing my capabilities for technical skills, I was reminded of my childhood aspirations and felt compelled to continue pursuing that dream. Throughout high school, I found immense joy in mentoring STEM subjects to fellow students, organizing free workshops and astronomy courses, and leading outdoor field trips for star gazing and physics experiments. The experiences and knowledge I had from my educational background left me wondering: how could I best combine my scientific abilities with my passion for spreading knowledge to as many young people as possible? The answer became clear: computer science.

As I delved deeper into computer science, I realized the potential for it to align perfectly with my strengths in logical thinking and my desire to disseminate knowledge through online platforms. Now, as I embark on this journey, the VURF program presents itself as a valuable opportunity to further explore my interests and solidify my specialization. Having spent nearly a year as a research assistant in NLP, I've come to truly appreciate the field's blend of creativity and resourcefulness, particularly in my collaboration with Dr. Nwafor. Assigned with tasks that demanded ingenuity and determination, I found fulfillment in overcoming challenges and meeting the professor's expectations.

With the current research proposal, I eagerly anticipate continuing the work I've been immersed in since May 2023, contributing to a project poised to yield tangible results. Additionally, this research offers me the chance to delve deeper into the field of NLP and assess its potential as a future career path. Whether it's pursuing a PhD in NLP or other machine learning fields or entering the industry with a specialization in machine learning, particularly NLP, this opportunity promises to shape my professional trajectory in profound ways.

**Budget**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Item** | **Quantity** | **Cost** | **Justification** | **Total** |
| Google Colab Pro+ | 3 | $49.99/month | Used for fine-tuning NLP models in a faster, more efficient way compared to performing the task in local laptop | $150 |
| GPT-3.5-turbo (fine-tuning) | Training:  ~ 40,000,000 tokens  Input:  ~ 400,000 tokens  Output (estimated):  ~ 410,000 tokens | Training: $0.0080/1k tokens (words)  Input usage:  $0.0030/1k tokens (words)  Output usage:  $0.0060/1k token (words) | One of the state-of-the-art models that costs money to be fine-tuned and tested for its performance on various NLP tasks. | $324 |

**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.

1. **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.

1. **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.

1. **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

1. **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.

1. **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 higher education, such as a PhD in machine learning/NLP, and help me navigate my career path, allowing me to explore whether NLP is the field where I truly belong in the industry.

**Bibliography**

Khurana, D., Koli, A., Khatter, K. et al. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 82, 3713–3744 (2023).

<https://doi.org/10.1007/s11042-022-13428-4>

Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). ChatGPT MT: Competitive for High- (but not Low-) Resource Languages. Retrieved from arXiv preprint repository (2309.07423). <https://doi.org/10.48550/arXiv.2309.07423>