Improving the Mkulima Repository: Utilizing Theses, Dissertations, and LLMs for Agricultural Knowledge Dissemination in Kiswahili

Abstract

Motivation and Problem

The Sokoine National Agricultural Library (SNAL) at Sokoine University of Agriculture (SUA) has two major mandates: offering library and information services to the SUA community and disseminating agricultural information at a national level. Given that over 95% of Tanzania's population speaks Kiswahili, there is a high demand for agricultural and food information in this language. However, SNAL has had difficulty finding sufficient bibliographic materials in Kiswahili, owing to the fact that agricultural and food research outputs, such as scientific publications, theses, and dissertations, tend to be written in English. Thus, the Swahili-speaking public has limited access to key agricultural and food knowledge due to the language barrier.

To address part of this issue, SUA established the Mkulima Repository, which is largely focused on curating agricultural and food-related content in Kiswahili. The collection includes Swahili agricultural and food books, booklets, brochures, and leaflets. These resources are the outcome of numerous sponsored programs at SUA, with researchers and student volunteers translating the findings into Kiswahili. The Mkulima collection's major objective is to help share or disseminate agricultural and food information and knowledge in Kiswahili to agricultural practitioners and the wider Swahili-speaking community, as well as to promote good agricultural practices (GAPs).

Objective

Despite this initiative, the Mkulima repository struggles to find appropriate Kiswahili content. One potential approach for boosting the Mkulima collection content is to leverage the extensive amount of knowledge contained in theses and dissertations written by SUA postgraduate students. SUA's

postgraduate requirements require students to submit electronic copies of their final copy-edited and error-free PDF theses or dissertations to the SUA Institutional Repository (SUAIR). To raise awareness of the findings of these studies, the institution is pushing for postgraduate students to write a Swahili version of their thesis or dissertation title, which will be read to the public during the graduation ceremony. However, this ends with the ceremony, and thus, does not benefit the community. This needs the implementation of additional steps to aid in the extraction of crucial information from these studies in Kiswahili, which should be made available. As a result, this study investigates the use of machine translation of abstracts from theses and dissertations with large language models (LLMs) to improve the Mkulima repository.

Methods

To translate abstracts from English to Kiswahili, we are exploring different open-source LLMs, including MarianMT and Meta models from the Hugging Face Transformers library, particularly the 'Helsinki-NLP/opus-mt-en-sw' and 'facebook/nllb-200' models. We shall first choose a representative sample of electronic theses and dissertations from the SUAIR. The abstracts will then be extracted and pre-processed into a format suitable for translation, treating each abstract as an independent document. Following translation, multilingual specialists, mainly Swahili-English language experts will assess the Kiswahili texts to ensure accuracy and contextual relevance. The validated documents, alongside their metadata, will then be uploaded to the Mkulima repository. In addition to using LLMs to translate the abstracts, the study will look into best practices for encouraging students to self-archive their translated abstracts in the Mkulima repository. To achieve this, the study will primarily undertake a literature review and benchmarking against other institutions around the world.

Results

By automatically translating theses and dissertation abstracts into Kiswahili, agricultural and food research will become more accessible to the Swahili-speaking population, expanding the reach and impact of SUA's research outputs via the Mkulima repository. Implementing a streamlined self-archiving system will encourage postgraduate students to submit work in both languages (the full thesis or dissertation in English to the SUAIR and the abstract in Kiswahili to the Mkulima repository), lowering administrative overheads and improving repository content. Making research output synopses available in Kiswahili is expected to increase Mkulima repository content engagement from local people, raising the institution's research profile.

Furthermore, the study's findings will provide actionable recommendations for improving the Mkulima repository through policy changes at SUA concerning thesis and dissertation, improved self-archiving processes, and the use of advanced NLP technologies for content translation in open-access institutional repositories. These enhancements will assist to close the gap between academic research and practical application, helping Tanzanian farmers and other players in the agriculture-food value chain.

Conclusion

Completion of this work is expected to set precedent for other academic institutions in Tanzania, as well as other bi- and multilingual countries, that face comparable issues in providing knowledge in one language while the bulk of consumers use another. This would also promote inclusive and accessible dissemination of knowledge, increasing the reach and impact of academic research in Tanzania and globally.

Keywords: Kiswahili, Agricultural Information, Thesis, Dissertation, Machine Translation, NLP, LLMs.