UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY

DEPARTMENT OF COMPUTING AND TECHNOLOGY

EASTER 2025 SEMESTER

PROGRAM: BSCS, BSDS 3:1

COURSE: BIG DATA MINING AND ANALYTICS

Project Title: A review article titled “Building natural language processing tools for Runyakitara”

submitted by

Name	Registration Number	Access Number
Obba Mark Calvin	S23B23/047	B24277

Date: April 18, 2025

1 An essay describing the insights derived from review article

The review article ”Building Natural Language Processing Tools for Runyakitara” by Fridah Katushemererwe, Andrew Caines and Paula Buttery, published in 2021, offers a comprehensive exploration of applying computational linguistics to preserve and revitalize low-resource Bantu languages in Uganda. Runyakitara, encompassing Runyankore, Rukiga, Runyoroand Rutooro, is spoken by approximately eight million people, representing a significant portion of Uganda's population. Despite these numbers, the article reveals critical vulnerabilities: while Ethnologue classifies the languages as ”educational” or ”developing” on the Expanded Graded Intergenerational Disruption Scale (EGIDS), the authors argue this assessment overlooks declining intergenerational transmission and restricted domains of use, labeling them as ”threatened” or ”endangered” in key dimensions like education and urban migration.

A central insight is the historical and sociolinguistic context shaping Runyakitara's challenges. Originating from Runyoro, the languages diverged due to missionary orthographic variations in the 19th century, with colonial policies promoting English and Kiswahili while marginalizing indigenous tongues. Post-independence efforts in the 1990s unified them under ”Runyakitara,” yet implementation of Uganda's ”mother tongue” policy remains inconsistent, with private schools favoring English and limited materials hindering literacy. The article highlights common spelling errors among even educated users, attributing this to recent orthographic formalization and morphological complexity—verbs like ”shoma” can generate over 40 million forms through affixes, extensionsand classifiers. This underscores the need for tailored NLP tools to address agglutinative structures unique to Bantu languages.

tailored NLP tools to address agglutinative structures unique to Bantu languages.
The project's core contributions—RunyaCorpus, RunyaMorph, RunyaSpellerand LearnRunya—demonstrate a cyclical approach to NLP development. RunyaCorpus compiles diverse data from speeches, essays, newspapersand The Bible, using optical character recognition (OCR) like Tesseract for digitization. RunyaMorph, a finite-state morphological analyzer, achieves 70% recognition accuracy, parsing intricate word structures to support spell-checking and CALL systems. Insights into CALL reveal pilots like RU-CALL and iCALL, which enhance grammar and syntax learning but highlight needs for bilingual interfaces to reach non-proficient heritage speakers.

Broader implications emphasize NLP's role in language revitalization amid globalization. The authors position Runyakitara as a case study for low-resource languages, where technology counters prestige loss by expanding domains into digital media, journalismand education. Since the article's publication, similar efforts have advanced; for instance, AI-driven projects using large language models (LLMs) for Indigenous languages have shown success in automated transcription and translation, aligning with RunyaMorph's goals. GPT-4 applications have generated learning materials for endangered tongues, suggesting potential extensions for Runyakitara. An ACL study on NLP for revitalization advocates machine-in-the-loop processing and community collaboration, echoing the article's call for open resources.

In conclusion, the article illuminates how computational tools not only improve literacy and accuracy but also elevate cultural prestige, fostering autonomous learning in multilingual contexts. By bridging gaps in resources, it models sustainable preservation, inspiring global initiatives for endangered languages facing similar threats from dominant ones like English.

Recommendations to Ethnic Leaders of the Runyakitara Speakers on Using NLP Models for Language Preservation

- **Integrate NLP into Educational and Policy Frameworks:** Advocate for the mandatory incorporation of tools like RunyaSpeller and LearnRunya into national curricula, aligning with Uganda's mother tongue policy. Partner with institutions such as Makerere University to train teachers on these systems and push for government funding to develop mobile apps accessible offline, targeting both rural and urban youth. This approach mirrors successful AI initiatives for Indigenous languages, where automated tools have significantly boosted classroom engagement.

- **Expand Community-Driven Data Collection:** Encourage widespread community participation in enriching RunyaCorpus through crowd-sourced platforms for collecting oral stories, songs, and modern dialogues. Utilize NLP for automated transcription and analysis to preserve dialects, ensuring ethical data handling practices. Inspired by projects leveraging GPT-4 for language documentation, involve elders in validating outputs to maintain authenticity and foster intergenerational bonds.

- **Enhance Digital Presence and Prestige:** Collaborate with media outlets such as Radio West and newspapers (e.g., Orumuri) to embed NLP tools for real-time error correction in broadcasts and articles. Develop social media campaigns showcasing Runyakitara content generated by large language models (LLMs), associating the language with innovation to attract younger speakers. This strategy draws from revitalization efforts, such as machine translation for endangered languages, which have elevated their status in digital domains.

- **Adopt Advanced AI for Revitalization:** Explore fine-tuning open-source large language models (e.g., those available on Hugging Face) for Runyakitara-specific tasks such as translation and chatbots, building on tools like RunyaMorph. Draw inspiration from ACL research emphasizing machine-in-the-loop collaboration, where NLP assists linguists in creating immersive learning applications. Pilot projects similar to those for Lakhota NLP, even if initial attempts face challenges, to iterate toward robust systems.

- **Form International Alliances and Monitor Progress:** Partner with global NLP research groups, such as those at the University of Michigan or Sheffield, for knowledge exchange on low-resource languages. Establish metrics for tracking usage, literacy gains, and speaker retention through analytics dashboards. Regularly update tools based on community feedback to ensure adaptability to emerging challenges like urbanization.

- **Address Ethical and Inclusivity Concerns:** Prioritize inclusive development by involving diverse community voices, including women and youth, to avoid biases in NLP models. Advocate for open-source releases to enable global contributions, as seen in student-led projects for low-resource languages. This ensures equitable access and sustains momentum beyond initial efforts.

github repo : https://github.com/KingHash23/BIG-DATA-and-ANALYSIS-Project-test-1.