

Harmonizing 4.1 Million Multilingual Tax Records — Rwanda Revenue Authority's NLP Clustering Pipeline

AI in the Public Sector · Good practice · Rwanda

Quality score: 60/100

Evidence strength: Descriptive / self-reported

Summary

A study built with Rwanda Revenue Authority data used MiniLM, UMAP and HDBSCAN to harmonize 4.1 million multilingual product names from EBM tax records into 425,103 standardized names, exposing pricing anomalies.

Key dimensions

Dimension	Score
Evidence of impact / measured public value	2/3
Transparency, fairness & accountability	2/3
Transferability / demonstrated replication	3/3
Scalability beyond pilot	1/3
Governance, capability & sustainability	1/3

Evaluated by C-NAPSE

Data sources

C-NAPSE web research — accessed 2026-08-15

Rwanda Revenue Authority (Strategy and Risk Analysis Department), research by Adventist University of Central Africa — accessed 2026-08-15

IJISRT: Harmonizing Multilingual Product Data Using Machine Learning — A Case Study of the Rwanda Revenue Authority — accessed 2026-08-15

ICTD: Technology Evolution and Tax Compliance — Evidence from Rwanda — accessed 2026-08-15

Cite this — Evidoria. *Harmonizing 4.1 Million Multilingual Tax Records — Rwanda Revenue Authority's NLP Clustering Pipeline*. Persistent ID: f5cf26f1-dca5-48c7-a0a4-31b49d1a24ff.