Legal ambiguity is the primary driver of **licensing risk** in African NLP datasets, while most projects fall within **Open & Transparent** governance topology.

Copyright Compliance and Licensing Risk in African NLP Datasets: A Data-Driven Governance Analysis

INTRO

African languages NLP datasets are becoming more important for AI development in African context but many datasets lack licensing and governance documentation. In this project we developed a theory guided licensing risk analysis and governance topology clustering.

METHODS

Risk Scoring → Model Training → Evaluation & Analysis → Visualization & Deployment

1. Dataset: 247 African NLP projects
2. EDA: Python (pandas, numpy, matplotlib)
3. Legal variables were selected to construct a rule based Licence **Risk Scoring Framework**; Low, Medium & High Risk.
4. **Logistic Regression, Random Forest and Decision Tree models** were trained for Licensing Risk categorization.
5. K-Means, Agglomerative Ward & Gaussian Mixture were used for topology clustering
6. Validation: **PCS** framework, cross-validation & repeated train-test split
7. Visualization: interactive Licensing Risk assessment **app with Streamlit** (figure 4).

Tankiso Kolobe, **Andries Monyebodi**

RESULTS

- Medium risk dominated with 144 projects while High and low Risk only had 60 and 45 respectively (figure 1).
- Logistic Regression achieved the best balance of predictability, stability and interpretability.
- Dominant governance topology is Open & Transparent (Table 1 below).

Typology	% of Projects
Open & Transparent	39%
Commercial/Deployable	32%
Community-Driven	15%
Implicit/Ambiguous	10%
Rights-Reserved	3%

Table 1: Topology Clustering Results

DISCUSSION

- Medium risk classifications were primarily associated with ambiguity in licensing documentation. High Risk drivers include, copyright protected material, absent licensing & unstated legal basis for use (figure 2).
- Although the dataset represents a static set of African NLP, the framework provide evidence-based guidance for future governance assessment.

jupyter Streamlit

Visualizations

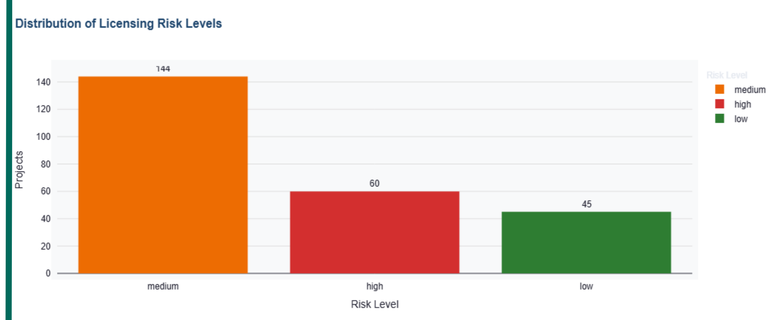

Figure 1: Licensing Risk Level Distribution

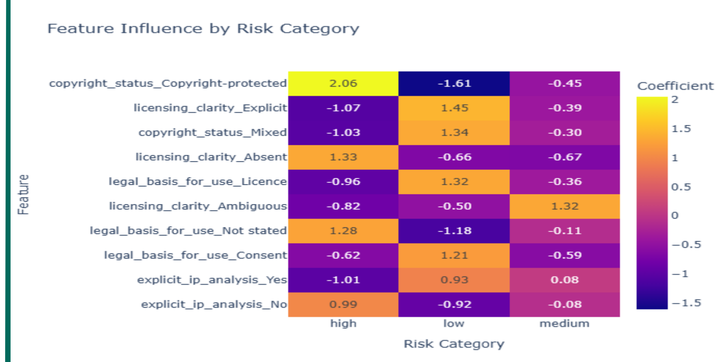

Figure 2: Feature Influence by Risk Level

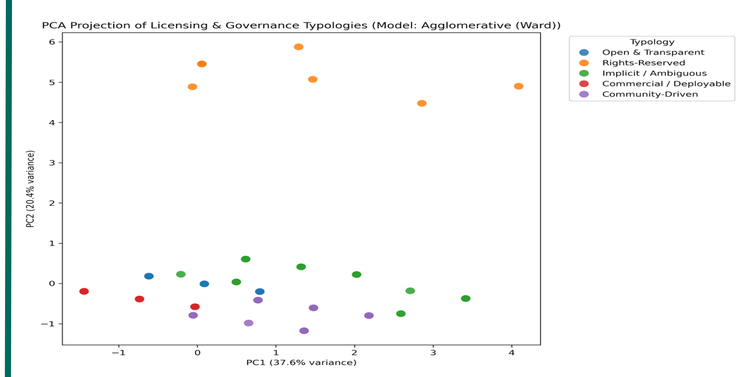

Figure 3: PCA Governance topologies

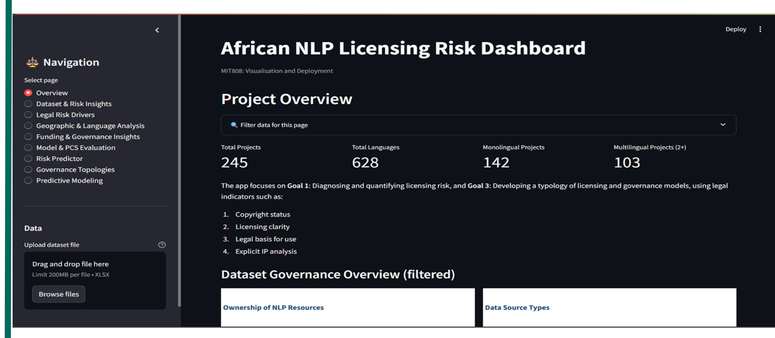

Figure 4: Deployment Overview

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)

Scan me