**African NLP projects are largely undocumented, unlicensed, and legally uncertain, and the communities whose languages are involved have almost no control over the resources being built from them.**

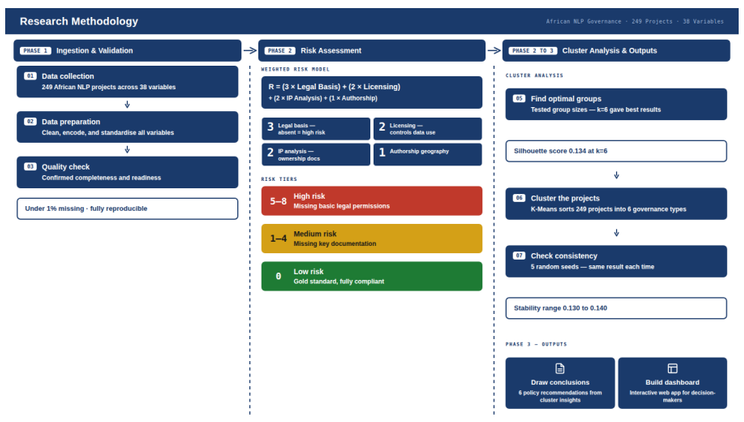
# Copyright Compliance and Licensing Patterns in African NLP Datasets: A Veridical Audit of African NLP

Merhawi Hailu, Malwandla Ngobeni

## 1 Introduction

- Artificial intelligence (AI) relies on big data to enhance performance.
- We analysed 249 African Natural Language Processing (NLP) projects, focusing on data sourcing, usage permissions, and governance documentation to promote responsible, ethical innovation.
- The PSC framework was applied to the data to help ensure decisions are made based on factual data, with all the legal checks in place.

## 2 Research Methodology




## 3 Results and Discussion

| | Full abbreviated reference | Project funding source | Authorship geography | NLP Resource Type | Licensing clarity | Explicit IP analysis | Total_Risk_Score | Risk_Profile |
|---|---|---|---|---|---|---|---|---|
| 1 | Meyer 2024a | none stated | Mixed | Dataset and two pretrained langua | absent | No | 5 | No Legal Basis, Ambiguous License |
| 2 | Andersland, M. 2024 | corporate | Global North only | Benchmark | explicit | No | 3 | No Legal Basis |
| 3 | Ardila et al., LREC 2020 | mixed | Mixed | Dataset | explicit | Yes | 0 | Legally Robust |
| 4 | Aremu et al. 2024 | none stated | Mixed | Dataset | ambiguous | No | 5 | No Legal Basis, Ambiguous License |
| 5 | Artetxe and Schwenk 2019 | none stated | Global North only | Model | ambiguous | No | 5 | No Legal Basis, Ambiguous License |
| 6 | Aryabumi et al., 2024 | corporate | Global North only | Model | explicit | Yes | 0 | Legally Robust |
| 7 | Meyer 2022a | none stated | Mixed | Dataset | absent | No | 5 | No Legal Basis, Ambiguous License |
| 8 | Aryal et al. 2023 | none stated | Global North only | Benchmark | explicit | No | 0 | Legally Robust |
| 9 | Mirbostani 2023a | nsf | Mixed | Deep active learning algorithm | ambiguous | No | 5 | No Legal Basis, Ambiguous License |
| 10 | Asai et al. 2024. | mixed | Global North only | Benchmark | explicit | No | 0 | Legally Robust |
| 11 | Atuhurra et al. 2024. | mixed | Global North only | Tool | ambiguous | No | 5 | No Legal Basis, Ambiguous License |
| 12 | Awobade et al. 2024 | none stated | Mixed | Model | ambiguous | No | 5 | No Legal Basis, Ambiguous License |
| 13 | Moghe 2023a | mixed | Global North only | Dataset | absent | No | 5 | No Legal Basis, Ambiguous License |
| 14 | Azime and Mohammed 2021 | none stated | Global South only | Dataset | ambiguous | No | 5 | No Legal Basis, Ambiguous License |
| 15 | Azime et al. 2024 | none stated | Mixed | Dataset, Model and Pipeline | ambiguous | No | 5 | No Legal Basis, Ambiguous License |
| 16 | Azunre et al. 2021 | none stated | Mixed | Dataset, Model, Tool, and Pipeline | ambiguous | No | 5 | No Legal Basis, Ambiguous License |
| 17 | Lawson et al. 2021b | none stated | Mixed | Dataset | ambiguous | No | 5 | No Legal Basis, Ambiguous License |
| 18 | Babu et al. 2022 | corporate | Global North only | Model | ambiguous | No | 5 | No Legal Basis, Ambiguous License |
| 19 | Mohamed and Mohamed 2022 | none stated | Mixed | Lexicon and Rule-based lemmatizc | not stated | No | 5 | No Legal Basis, Ambiguous License |
| 20 | Mohammadshahi, A. et al, 20 | global north public | Global North only | Model | not stated | No | 5 | No Legal Basis, Ambiguous License |

Figure 1: Risk audited projects

- Projects that were found to be 'High Risk' (Score 5-8) were missing the necessary legal basis and explicit IP documentation. Most projects have a medium risk score.
- A significant portion of research was conducted by researchers located in the Global North.
- The projects with the highest risk resulted in the creation of a dataset, which shows a gap in the datasets being used for African NLP projects.

- Figure 4 shows the risk of distribution using a density curve which identified where most projects sit on the 0-8 risk scale. Higher peaks indicate the most common risk profile for that specific geography.
- Figure 5 visualizes the transparency of usage permissions.

- Figure 6 shows the projects that are predicted to be 'High Risk' (Score 5-8) based on their funding source.
- Figure 7 shows which project outcomes (e.g., Datasets, Models) carry the most legal risk.

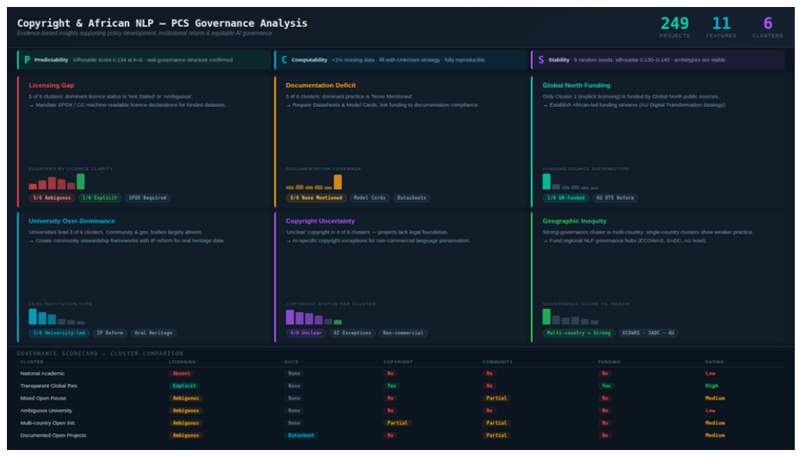



Figure 2: Governance gaps in African NLP projects reveal licensing, documentation, and equity challenges.

- Licensing Gap: Dominant status remains "NotStated " or "Ambigous" across 5 of 6 cluster.
- Documentation deficit: Standard Practise default to "None Mentioned" in 5 of 6 clusters.
- Global North funding: only the cluster featuring explicit licensing is funded by global public source.
- University over-dominance: Universities lead 3 of 6 cluster, while community and government bodies are largely absent.
- Copyright Uncertainty: project lack legal foundation, with "Unclear " copyright spanning 4 of cluster.

- Figure 8 shows K-Means (k = 6) tested with five random seeds, producing silhouette scores between 0.1233 and 0.1577, all above the 0.10 threshold for meaningful governance data. Despite label switching, the spread of 249 projects across six groups remains consistent, confirming the six governance archetypes as stable, reproducible structures that policymakers can rely on.
- Figure 9 shows inertia dropping until k = 3–4 before flattening, while silhouette scores stay above the 0.10 threshold. Together, they confirm stable governance patterns and meet the PCS Framework's Predictability requirement, validating six archetypes for policy use.

### Policy Insight

| # | Finding | Evidence | Recommendation |
|---|---|---|---|
| 1 | **Licensing Gap** | 'Not stated' dominates 5/6 clusters | Mandate SPDX licences for all public-funded NLP data |
| 2 | **Documentation Deficit** | 'None mentioned' in 5/6 clusters | Require Datasheets & Model Cards for all published datasets |
| 3 | **Global North Funding Skew** | Only Cluster 1 has transparent Global North funding | Create African-led funding streams with governance requirements |
| 4 | **University Over-Dominance** | 3/6 clusters university-led | Introduce community stewardship frameworks |
| 5 | **Copyright Uncertainty** | 'Unclear' in 4/6 clusters | AI-specific copyright exceptions aligned to African legal context |
| 6 | **Capacity Inequality** | Single-country projects show weaker governance | Fund regional NLP governance hubs at SADC/AU level |

Figure 3: Key-finding and recommendation

## Extra figures

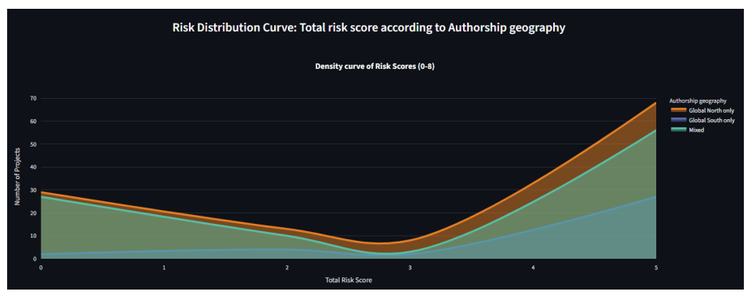



Figure 4: Risk Distribution Curve

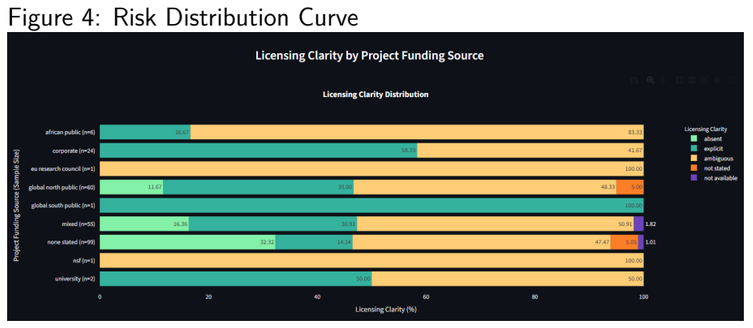



Figure 5: Licensing Clarity by Project Source

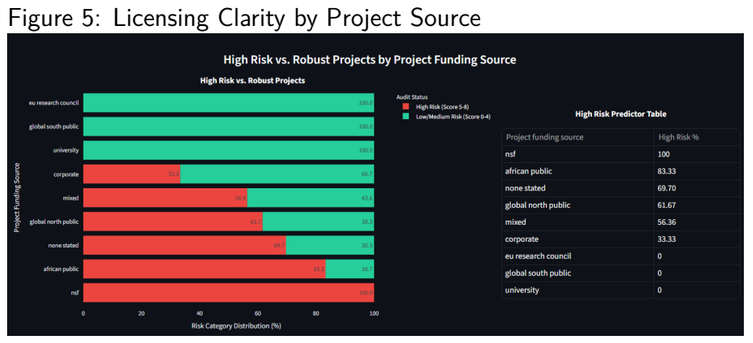



Figure 6: Predictive Analysis

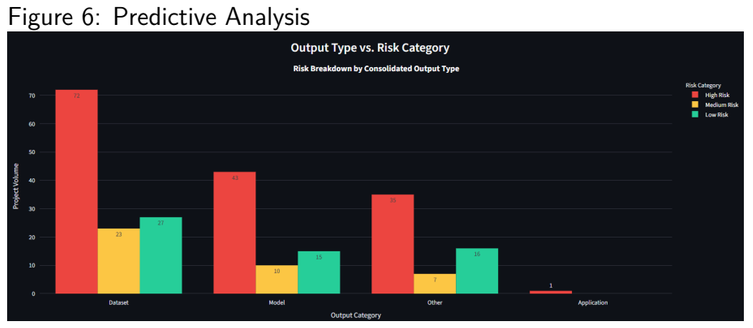



Figure 7: Categorical Risk Mapping

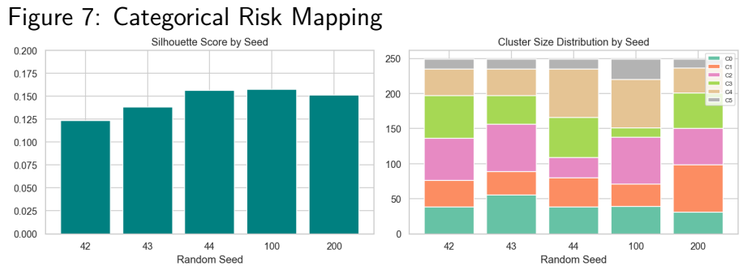



Figure 8: Stability check

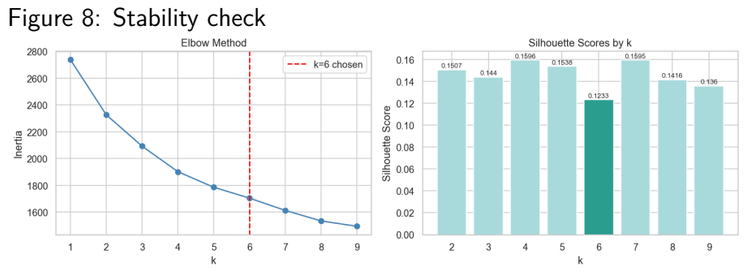



Figure 9: Predictability using Elbow Silhouette

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