Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages

Farhana Shahid[1], Mona Elswah[2], Aditya Vashistha[1]

[1]Cornell University, [2]Center for Democracy and Technology

Full paper

Provocation

- Content moderation failures in the Global South are cast as a ***"data problem"*** of low-resource languages.
- Would moderation really improve if these languages had lots of data?
- Why there are not enough data in these languages despite being spoken by millions in the Global South?
- Why current language-agnostic technologies perform poorly in these languages?

Research Question

- **RQ1.** What systemic barriers impact automated moderation pipelines for low-resource languages?
- **RQ2.** How might we improve automated moderation for low-resource languages?

Methodology

Interviews with 22 AI researchers and practitioners

- Tamil (South Asia)
- Swahili (East Africa)
- Quechua (South America)
- Maghrebi Arabic (North Africa)

Contributions

- **Empirical evidence** of systemic issues across moderation pipeline
- **Theoretical contribution** surfacing coloniality behind these systemic inequity

Key Findings (🟢 socio-political and 🔵 technical issues)

💰 **Lack of financial interest** to invest in moderation pipelines for low-resource languages

Data sources

Manifestations of digital colonialism

🔒 **Data restriction** by tech companies **to build proprietary LLMs** hinder grassroots moderation efforts

💣 News articles that **portray Muslims as terrorists** are used as **Arabic** data sources

✝️ Google uses **Bible translations** as data for **Indigenous languages** like Quechua

💸 Companies spend a lot for moderation in Western contexts but expect **voluntary labor** from Global South communities

Monolithic assumptions

🤬 Companies often use **fixed list of slurs as a patching solution** for low-resource languages ignoring the regional diversity

🗣️ **Machine translated data** for low-resource languages often rely on **outdated corpora** (Sheng vs. Shembeteng) and **overlook dialectical variations** (Tanzanian vs Kenyan Swahili)

Data annotation

Corporate profit vs safety

💰 **Lack of financial interest to recruit annotators** for diverse Global South languages

🧑🏿‍💻 Global South data workers **mostly annotate harmful content in English**

💰 Historic **lack of resources** in Global South institutions hinder sustainable annotation practices

Western centrism

💀 **Sentiment and toxicity analysis models** misclassify non-Western contexts based on **Western notions of harm**

🙈 Language detection technologies **overlook code-mixing** in the Global South, which complicates the annotation of harmful content

Data preprocessing

Normative assumptions in technology design

🔚 Colonial suppression of native languages and limited support for non-Latin scripts led to **code-switching, romanization,** and **code-mixing** among Global South users

👇 Preprocessing pipelines treat **code-mixed, romanized data** which are absent in English **as *"low quality"***

👁️ Colonial linguists perceived morphologically complex **agglutinative languages** (e.g., Tamil, Swahili, Quechua) **as *"less evolved"*** than Western languages

⬇️ Preprocessing techniques optimized for data-rich languages like English **underperform in complex agglutinative languages** that have distinct word formations **than English**

Model training

Normalizing data-intense and language-agnostic approaches

💰 Current design of **data** and **resource-intensive** multilingual models are **ill-suited** to detect harmful content in the Global South

🙈 Tech companies **overlook language-aware approaches** due to corporate arms race to build language agnostic models

Language naïve models

❌ Large multilingual models **fail to infer correct linguistic properties from different language families**

🔨 AI models **flatten the diversity in annotation** by allowing a singular label, especially for content **with rich dialectical variations**

Contact: Farhana Shahid | fs468@cornell.edu | https://farhana-shahid.github.io/