

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

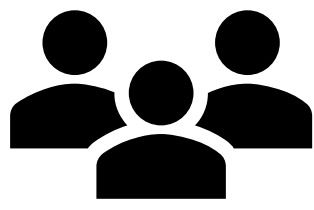


Farhana Shahid<sup>1</sup>, Mona Elswah<sup>2</sup>, Aditya Vashistha<sup>1</sup>

<sup>1</sup>Cornell University, <sup>2</sup>Center for Democracy and Technology



Full paper

Provocation	Research Question	Methodology	Contributions
<ul style="list-style-type: none"><li>Content moderation failures in the Global South are cast as a “<b>data problem</b>” of low-resource languages.</li><li>Would moderation really improve if these languages had lots of data?</li><li>Why there are not enough data in these languages despite being spoken by millions in the Global South?</li><li>Why current language-agnostic technologies perform poorly in these languages?</li></ul>	<ul style="list-style-type: none"><li><b>RQ1.</b> What systemic barriers impact automated moderation pipelines for low-resource languages?</li><li><b>RQ2.</b> How might we improve automated moderation for low-resource languages?</li></ul>	 Interviews with 22 AI researchers and practitioners	<ul style="list-style-type: none"><li><b>Empirical evidence</b> of systemic issues across moderation pipeline</li><li><b>Theoretical contribution</b> surfacing coloniality behind these systemic inequity</li></ul>

Key Findings ( <span style="color: green;">●</span> socio-political and <span style="color: blue;">●</span> technical issues)			
Data sources	Data annotation	Data preprocessing	Model training
<span style="color: orange;">💰</span> <b>Lack of financial interest</b> to invest in moderation pipelines for low-resource languages			
<div><p><u>Manifestations of digital colonialism</u></p><p><span style="color: orange;">🔒</span> <b>Data restriction</b> by tech companies to <b>build proprietary LLMs</b> hinder grassroots moderation efforts</p><p><span style="color: purple;">🗞️</span> News articles that <b>portray Muslims as terrorists</b> are used as <b>Arabic</b> data sources</p><p><span style="color: purple;">✝️</span> Google uses <b>Bible translations</b> as data for <b>Indigenous languages</b> like Quechua</p><p><span style="color: orange;">💵</span> Companies spend a lot for moderation in Western contexts but expect <b>voluntary labor</b> from Global South communities</p></div> <div><p><u>Monolithic assumptions</u></p><p><span style="color: red;">🗨️</span> Companies often use <b>fixed list of slurs</b> as a <b>patching solution</b> for low-resource languages ignoring the regional diversity</p><p><span style="color: blue;">🗣️</span> <b>Machine translated data</b> for low-resource languages often rely on <b>outdated corpora</b> (Sheng vs. Shembeteng) and <b>overlook dialectical variations</b> (Tanzanian vs Kenyan Swahili)</p></div>	<div><p><u>Corporate profit vs safety</u></p><p><span style="color: orange;">💰</span> <b>Lack of financial interest</b> to recruit <b>annotators</b> for diverse Global South languages</p><p><span style="color: blue;">👤</span> Global South data workers <b>mostly annotate harmful content</b> in English</p><p><span style="color: orange;">💰</span> Historic <b>lack of resources</b> in Global South institutions hinder sustainable annotation practices</p></div> <div><p><u>Western centrism</u></p><p><span style="color: red;">💀</span> <b>Sentiment and toxicity analysis models</b> misclassify non-Western contexts based on <b>Western notions of harm</b></p><p><span style="color: red;">🗣️</span> Language detection technologies <b>overlook code-mixing</b> in the Global South, which complicates the annotation of harmful content</p></div>	<div><p><u>Normative assumptions in technology design</u></p><p><span style="color: orange;">📺</span> Colonial suppression of native languages and limited support for non-Latin scripts led to <b>code-switching, romanization, and code-mixing</b> among Global South users</p><p><span style="color: yellow;">🗨️</span> Preprocessing pipelines treat <b>code-mixed, romanized data</b> which are absent in English as “<b>low quality</b>”</p><p><span style="color: brown;">👁️</span> Colonial linguists perceived morphologically complex <b>agglutinative languages</b> (e.g., Tamil, Swahili, Quechua) as “<b>less evolved</b>” than Western languages</p><p><span style="color: blue;">⬇️</span> Preprocessing techniques optimized for data-rich languages like English <b>underperform in complex agglutinative languages</b> that have distinct word formations <b>than English</b></p></div>	<div><p><u>Normalizing data-intense and language-agnostic approaches</u></p><p><span style="color: orange;">💰</span> Current design of <b>data</b> and <b>resource-intensive</b> multilingual models are <b>ill-suited</b> to detect harmful content in the Global South</p><p><span style="color: red;">🗣️</span> Tech companies <b>overlook language-aware approaches</b> due to corporate arms race to build language agnostic models</p></div> <div><p><u>Language naïve models</u></p><p><span style="color: red;">✖️</span> Large multilingual models <b>fail to infer correct linguistic properties from different language families</b></p><p><span style="color: red;">🔨</span> AI models <b>flatten the diversity in annotation</b> by allowing a singular label, especially for content <b>with rich dialectical variations</b></p></div>