UCLA
Recent Work

Title

Underrepresentation and Power: Data Colonialism in IndonesianLanguage-based AI Development

Permalink

https://escholarship.org/uc/item/3hr4h2qr

Author

Lambey, Katrina

Publication Date

2026-03-10

eScholarship.org Powered by the California Digital Library
University of California

Underrepresentation and Power: Data Colonialism in Indonesian Language-based AI Development

Katrina Aurelia Lambey[1]

[1]Bioengineering Department, University of California, Los Angeles

klambey@ucla.edu

Abstract: This paper examines how data colonialism shapes language-based AI and its development in Indonesia, arguing that extraction, benchmarking, and optimization practices structurally underrepresent Bahasa Indonesia and regional languages, limiting linguistic equity and national digital sovereignty.

INTRODUCTION

Digital infrastructure has now shaped people's ability to access information and participate in society, and language-based artificial intelligence (AI) systems have become increasingly prominent in that shift. English has been treated as the default language for language-based AI, creating uneven performance across other linguistic and cultural contexts. Countries such as Indonesia, despite their linguistic diversity and growing digital population, remain underrepresented in the datasets used to train large language models, limiting system performance for large segments of the population. Indonesia is one of the world's most linguistically varied countries, with over 700 living languages and dialects (Eberhard et al., 2023). Because of this, language-based AI systems may fail to adequately support large portions of its population.

Several scholars suggest that this disparity is not simply accidental. Relating technology to capitalism, Couldry and Mejias (2019) characterize digital infrastructures as a form of data colonialism. Thatcher, O'Sullivan, and Mahmoudi (2016: 990), following David Harvey, describe this process as "capitalist accumulation by dispossession," in which human communication is collected and reused as raw material for commercial gain. Noble (2018) contends that algorithmic bias emerging from institutional design choices and priorities is systemic, with discrimination embedded in code and increasingly in AI systems. Benjamin (2019) coined the phrase "default user," describing how technology is designed to benefit dominant populations while keeping others in structural inequality.

With these ideas in mind, this paper will examine the extent to which data colonialism shapes the development of language-based AI systems in Indonesia, particularly through the underrepresentation of Indonesian and regional languages in AI training data.

METHODS

This paper takes a qualitative critical research approach grounded in standpoint theory and data colonialism, examining how language-based AI systems produce linguistic inequality in Indonesia. AI is not treated as a neutral tool but as a sociotechnical infrastructure shaped by institutional priorities and global power hierarchies. Therefore, unequal model performance is treated because of design priorities rather than a technological limitation.

Standpoint theory (Harding, 1991) argues that marginalized groups possess a more "objective and privileged" understanding of social reality because they experience structural inequalities directly. This study looks at language-based AI from the perspective of linguistically marginalized communities rather than assuming English as the neutral standard. Couldry and Mejias' data colonialism framework is used to evaluate how Indonesian linguistic data is extracted, filtered, and monetized within global AI systems.

The theoretical foundation will be grounded on qualitative critical literature in AI ethics, while the empirical evaluation of language representation relies on multilingual NLP benchmarks, dataset documentation, and corporate transparency reports.

RESULTS AND INTERPRETATION

Data Extraction Mechanism

Large language models (LLMs) are trained using publicly available datasets and web-scraped corpora. A major dataset used by LLMs is Common Crawl, a non-profit organization that provides web archive data used in training models such as GPT-3, LLaMA, and T5. According to Common Crawl (n.d.), the repository contains petabytes of data collected from billions of web pages. However, Common Crawl does not collect data around the goal of linguistic equity. It aggregates publicly accessible web text at scale, which is later repurposed into AI training datasets by downstream developers (Common Crawl, n.d.; Dodge et al., 2021).

An example is the Colossal Clean Crawled Corpus (C4), derived from Common Crawl to train T5 (Raffel et al., 2020). In building C4, Raffel et al. (2020) removed “noisy, low-quality” text, including short documents, duplicated content, and material filtered through language detection systems. Dodge et al. (2021) later showed that C4 documentation was limited and that filtering decisions shaped representation. These filtering standards are not linguistically neutral. If Bahasa Indonesia or regional dialect content is informal or includes code-switching, it is more likely to be classified as “noise” under heuristics shaped by English-dominant norms. C4 does not merely reflect web imbalances; it institutionalizes them. Indonesian language data is evidently present in these datasets. It appears in CC100 (Wenzek et al., 2020) and mC4 (Xue et al., 2021), both derived from Common Crawl and used to pretrain multilingual models globally. However, control over extraction, filtering, and classification remains with the institutions that constructed these corpora (Raffel et al., 2020; Dodge et al., 2021), not with Indonesian linguistic communities or national institutions. From a data colonialism perspective, this shows an asymmetrical infrastructural control: communication is extracted locally, but classification standards set externally. Standpoint theory further shows how dominant institutional preferences are treated as neutral standards.

Indonesian language data enters global AI systems, but the authority to define its value does not. These datasets are integrated into commercial AI systems (Wenzek et al., 2020; Xue et al., 2021; Stanford AI Index, 2024). Yet Indonesian institutions do not exercise governance authority over how their linguistic data is structured, evaluated, or monetized. This asymmetry captures a key defining feature of data colonialism: extraction without reciprocal sovereignty (Couldry & Mejias, 2019).

Evaluation Benchmarks and Optimization Priorities

Data colonialism also operates through evaluation systems. XTREME is a benchmark used to evaluate cross-lingual generalization in multilingual models (Hu et al., 2020). It includes Bahasa Indonesia but does not include major regional languages such as Javanese or Sundanese (Hu et al., 2020). Realistically, benchmarks cannot include all 700+ Indonesian languages, but it's the point that inclusion criteria are largely based on data availability and existing resources (Hu et al., 2020). Data availability is shaped by historical digitization patterns, research funding, and institutional investment (Joshi et al., 2020), and NLP research disproportionately focuses on a small subset of high-resource languages despite the existence of over 7,000 living languages worldwide.

Around 60% of filtered Common Crawl used to train GPT-3 was in English (Brown et al., 2020), demonstrating how large-scale scraping reflects linguistic concentration. The scraping and benchmark selection reproduce pre-existing linguistic hierarchies rather than correcting them. Benchmark inclusion shapes research incentives: models are optimized for the languages that are measured (Hu et al., 2020).

With English frequently being the anchor language for cross-lingual transfer and as the baseline against

which multilingual performance is evaluated, languages other than English that are excluded from benchmarks are less likely prioritized for future research cycles (Joshi et al., 2020). The Stanford AI Index (2024) shows that frontier model development is aligned with benchmark-driven performance gains, reinforcing optimization toward already dominant languages. This is similar to what Benjamin (2019) describes as the "default user," in which systems are implicitly optimized around dominant linguistic populations. From a standpoint theory lens, what appears as neutral selection based on "data availability" reflects institutional perspectives about which languages matter computationally.
Thus, the issue is not total exclusion but hierarchical inclusion. Indonesian data participates in global AI development under evaluation systems that structurally favor English-dominant environments.

Real World Governance Consequences in Indonesia
Uneven language representation in the AI systems can even be seen in Indonesia's digital governance environment. In 2022, an international organization that promotes the freedom of expression called ARTICLE 19 partnered with UNESCO to write an article about the content moderation and local stakeholders in Indonesia. It reports that Facebook's Indonesian-language misinformation guidelines once displayed a warning that some content was "not yet available in Indonesian language," whereas the English version of the Community Standards served as the authoritative master document. This indicates that Indonesian translations are occasionally incomplete or insufficient. Governance is more than just making rules, but also about who is able to have full access to the authoritative version of those rules. While English remains the dominant reference point, Indonesian obviously took a secondary position (ARTICLE 19, 2022).

This issue is much more complex as Indonesia is so diverse linguistically, and many of these languages lack sufficient annotated datasets, benchmarks, and resources (Aji et al., 2022). Even if Bahasa Indonesia (the main language of Indonesia) appears in benchmarks such as XTREME, regional ones remain underrepresented in NLP research and system development. The result is a layered inequality. English remains dominant, Bahasa Indonesia is secondary in position with partial inclusion, and regional languages are marginalized.

Indonesia's government and its own policy frame this imbalance as a sovereignty issue. Compared to countries such as the United States and China, which have dominance over AI standards and benchmarks through concentrated research output, private investment, and computational capacity, Indonesia operates within a global AI ecosystem shaped by those centers of infrastructural power. Although global models use Indonesian language data, Indonesia does not have the same agency as countries like the U.S. or China in determining how that data is filtered, evaluated, or optimized (Stanford Global Vibrancy Tool). The National Strategy on Artificial Intelligence has recognized infrastructure and data as key strategic priorities for Indonesia's near future. The government has emphasized the importance of ensuring that Indonesian data is not controlled by foreign parties (Ministry of Research and Technology/National Research and Innovation Agency, 2020). Recently, in 2022, the Indonesian government has released a Personal Data Protection Act that establishes regulatory authority over data processing that affects the country (Library of Congress, 2022). These policy efforts indicate Indonesia's recognition of the importance of data control and digital sovereignty. It isn't simply just if the Indonesian language data is present, but also who controls how the data is being collected, evaluated, and commercialized. This is the extent to which data colonialism organizes AI development in Indonesia.

CONCLUSIONS

The paper has shown how data colonialism influences AI development in Indonesia at the extraction, evaluation, and governance levels. Rather than being merely excluded, Indonesia holds a structurally secondary role within global AI ecosystems. Indonesian data penetrates global infrastructures, yet quality, benchmarking, and optimisation criteria are determined elsewhere. This hierarchical inclusion determines which languages are investable for future research direction and which stay peripheral, thus limiting

Indonesia's capacity to influence the development of its own language-based AI systems.

To address this would need not just participation, but also real control, via community consent, locally set criteria, equitable benefit-sharing, and increased institutional power over linguistic data. Without such structural changes, participation will not result in technical sovereignty.

REFERENCES

Aji, A. F., Winata, G. I., Koto, F., Cahyawijaya, S., Romadhony, A., Mahendra, R., Kurniawan, K., Moeljadi, D., Prasojo, R. E., Baldwin, T., Lau, J. H., & Ruder, S. (2022). *One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia*. arXiv. https://arxiv.org/abs/2203.13357

Benjamin, R. (2019). *Race after technology: Abolitionist tools for the new Jim code*. Polity Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*, 1877–1901.

Couldry, N., & Mejias, U. A. (2019). *The costs of connection: How data is colonizing human life and appropriating it for capitalism*. Stanford University Press.

Dodge, J., Gururangan, S., Card, D., Schwartz, R., & Smith, N. A. (2021). Documenting large webtext corpora: A case study on the Colossal Clean Crawled Corpus. *Proceedings of EMNLP 2021*, 1286–1305.

Eberhard, D. M., Simons, G. F., & Fennig, C. D. (Eds.). (2023). *Ethnologue: Languages of the world* (26th ed.). SIL International. https://www.ethnologue.com

Harding, S. (1991). *Whose science? Whose knowledge? Thinking from women's lives*. Cornell University Press.

Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., & Johnson, M. (2020). XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *Proceedings of ICML 2020*, 4411–4421.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. *Proceedings of ACL 2020*, 6282–6293.

Library of Congress. (2022). *Indonesia: Personal data protection law enacted*. https://www.loc.gov/item/global-legal-monitor/2022-10-07/indonesia-personal-data-protection-law-enacted/

Ministry of Research and Technology/National Research and Innovation Agency. (2020). *National strategy for artificial intelligence 2020–2045*. Government of Indonesia.

Noble, S. U. (2018). *Algorithms of oppression: How search engines reinforce racism*. NYU Press.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., … Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research, 21*(140), 1–67.

Stanford Institute for Human-Centered Artificial Intelligence. (2024). *AI index report 2024*. Stanford University.

Thatcher, J., O'Sullivan, D., & Mahmoudi, D. (2016). Data colonialism through accumulation by dispossession: New metaphors for daily data. *Environment and Planning D: Society and Space, 34*(6), 990–1006.

Wenzek, G., Lachaux, M. A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., & Grave, E. (2020). CCNet: Extracting high quality monolingual datasets from web crawl data. *Proceedings of LREC 2020*, 4003–4012.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou', R., Siddhant, A., … Raffel, C. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. *Proceedings of NAACL 2021*, 483–498.

ARTICLE 19. (2022). *Content moderation and local stakeholders in Indonesia*. ARTICLE 19 & UNESCO.