Sri Lanka Document Datasets: A Large-Scale, Multilingual Resource for Law, News, and Policy (v20251005)

Nuwan I. Senaratna
Independent Researcher
nuwans@alumni.stanford.edu

Abstract

We present a collection of open, machine-readable document datasets covering parliamentary proceedings, legal judgments, government publications, news, and tourism statistics from Sri Lanka. As of v20251005, the collection currently comprises 215,552 documents (59.6 GB) across 13 datasets in Sinhala, Tamil, and English, updated daily and mirrored on GitHub and Hugging Face. These datasets aim to support research in computational linguistics, legal analytics, socio-political studies, and multilingual natural language processing. We describe the sources, collection pipeline, formats, and potential use cases, while discussing licensing and ethical considerations.

1 Introduction

Sri Lanka's digital record of law, policy, and media is fragmented across numerous government and private sources. Much of this information exists as PDFs or web pages, often lacking machine-readable structure or public archival consistency. This fragmentation limits access for citizens, journalists, and researchers interested in the island's governance, history, and socio-economic trends.

The Sri Lanka Document Datasets initiative aims to bridge this gap by collecting, cleaning, and organizing key public documents into standardized, machine-readable formats. It unifies diverse materials—from Hansards and court judgements to news articles and tourism reports—under a common data framework. All datasets are openly licensed and continuously updated to ensure reproducibility and public transparency.

This effort is particularly significant for data-driven research in low-resource contexts. By providing structured data in Sinhala, Tamil, and English, the project supports the development of natural language processing models, cross-lingual studies, and digital humanities research. The datasets also enable policy analysis, legal precedent tracking, and media monitoring in a transparent, open science environment.

In this paper, we describe the scope and structure of these datasets, outline the scraping and curation processes, and highlight their potential applications in AI, governance, and public knowledge. Our goal is to create a living data archive that strengthens civic engagement and academic research through open, verifiable information.

2 Related Work

The study of open datasets has been central to the development of natural language processing (NLP) and computational social science. Large corpora such as Common Crawl[1], Wikipedia Dumps[2], and OpenWebText(Gokaslan and Cohen, 2019)have powered models that generalize across domains. However, these resources are dominated by data from high-resource languages and global institutions.

Regional initiatives have sought to address this imbalance by creating domain-specific collections. Examples include the Indian Kanoon legal corpus[3], the OpenSubtitles multilingual dataset[4], and the African News Corpus[5]. Such datasxets have improved representation for under-resourced languages and enabled comparative linguistic research.

In South Asia, efforts remain scattered and often focus on individual media outlets or institutions. Sri Lanka, in particular, lacks consolidated and machine-readable documentation of its public records. Prior datasets were either limited in size, language coverage, or temporal continuity[6][7].

The Sri Lanka Document Datasets aim to fill this gap by aggregating diverse sources—parliamentary

[1] https://commoncrawl.org/
[2] https://dumps.wikimedia.org/
[3] https://indiankanoon.org/
[4] https://opus.nlpl.eu/OpenSubtitles.php
[5] https://data.africanlp.org
[6] https://github.com/sltalk
[7] https://github.com/SriLankaNLP

debates, court judgements, gazettes, press releases, and news—into a unified, open, and multilingual repository. This complements global initiatives by providing a structured view of a unique national information ecosystem.

3 Datasets

As of v20251005, Sri Lanka Document Datasets consists of 13 datasets.

1. **Hansard**: A Hansard is the official verbatim record of parliamentary debates, preserving lawmakers' words and decisions for history, law, and public accountability. *1,665 documents, 17.9 GB, from 2006-02-01 to 2025-09-24.* Source: https://www.parliament.lk).

2. **Appeal Court Judgements**: A Court of Appeal judgment is a higher court ruling that reviews decisions of lower courts, shaping legal precedent and protecting citizens' rights. *10,146 documents, 10.3 GB, from 2012-04-23 to 2025-10-03.* Source: https://courtofappeal.lk).

3. **Supreme Court Judgements**: A Supreme Court judgment is a binding legal decision that interprets the Constitution and laws, shaping justice, governance, and citizens' rights. *1,575 documents, 1.3 GB, from 2016-07-22 to 2025-10-03.* Source: https://supremecourt.lk).

4. **Police Press Releases**: A police press release is an official update from law enforcement on crimes, arrests, safety alerts, or public notices, ensuring transparency and public awareness. *736 documents, 250.5 MB, from 2025-05-01 to 2025-10-03.* Source: https://www.police.lk).

5. **Acts**: A legal act is a law passed by Parliament that governs rights, duties, economy, and society, shaping daily life and national policy. *3,928 documents, 6.8 GB, from 1981-01-22 to 2025-09-22.* Source: https://documents.gov.lk).

6. **Bills**: A Bill is a draft law proposed in Parliament. It becomes binding once passed and enacted, shaping governance, rights, and daily life in the country. *4,077 documents, 1.8 GB, from 2010-05-10 to 2025-10-26.* Source: https://documents.gov.lk).

7. **Extraordinary Gazettes**: An Extraordinary Gazette is an official government publication used to announce urgent laws, regulations, or public notices with immediate effect. *101,532 documents, 19.4 GB, from 2010-01-01 to 2025-10-03.* Source: https://documents.gov.lk).

8. **Cabinet Decisions**: A Sri Lanka Cabinet Decision is an official policy or action agreed by the Cabinet of Ministers, shaping governance, law, and national development in the country. *10,369 documents, 125.3 MB, from 2010-09-27 to 2025-09-22.* Source: https://www.cabinetoffice.gov.lk).

9. **Treasury Press Releases**: A Sri Lanka Treasury press release shares key govt financial updates—on budgets, debt, or policy—vital for transparency, guiding investors, citizens, and markets on the nation's economic direction. *133 documents, 142.9 MB, from 2015-09-08 to 2025-07-30.* Source: https://www.treasury.gov.lk).

10. **Pmd Press Releases**: A Sri Lanka Presidential Media Division press release shares official updates on national decisions, policies, or events. It's vital as the authoritative source ensuring transparency and public awareness. *2,182 documents, 55.9 MB, from 2024-09-23 to 2025-09-24.* Source: multiple sources.

11. **News**: A collection of news documents. *79,049 documents, 1.2 GB, from 2021-09-12 to 2025-10-04.* Source: multiple sources.

12. **Tourism Weekly Reports**: Report on Weekly Tourist Arrivals to Sri Lanka. *33 documents, 91.5 MB, from 2023-01-01 to 2025-09-01.* Source: https://www.sltda.gov.lk).

13. **Tourism Monthly Reports**: Report on Monthly Tourist Arrivals to Sri Lanka. *127 documents, 308.9 MB, from 2015-01-01 to 2025-08-01.* Source: multiple sources.

4 Data Collection Pipeline

Data Collection Pipeline Our pipeline is automated, reproducible, and resilient. It continuously discovers, ingests, parses, validates, and versions documents from official Sri Lankan sources.(Ashraf et al., 2022)

GitHub Actions orchestrates the workflow. Cron jobs run several times per day, per dataset. A matrix strategy isolates each source, allowing independent retries without blocking others. Secrets manage tokens; caches speed I/O.[8]

Each run is idempotent and incremental. Before crawling, we load a manifest of known items. New or changed items are detected by stable keys (URL + date) and content hashes. Only deltas are committed to the repository.(Stodden et al., 2017)

Crawling is implemented in Python with Selenium in headless Chromium. We wait for dynamic content via explicit conditions (e.g., presence of

[8] https://docs.github.com/actions

selectors), handle pagination, and capture canonical URLs.[9]

Politeness is enforced. We respect robots.txt, throttle requests, randomize delays, and apply exponential backoff on transient failures. Errors are logged and surfaced in the Actions summary for rapid triage.(Pant and Srinivasan, 2021)

Raw artifacts are preserved alongside parsed representations. For each document we store the fetched HTML or PDF, plus normalized JSON with metadata (title, date, source, language, hashes) to enable downstream reproducibility.(Blaustein et al., 2020)

PDF parsing uses PyMuPDF (fitz). For each PDF, we extract text, metadata, and layout blocks, retain page boundaries, and normalize Unicode. When images contain embedded text, PyMuPDF's text extraction captures vector text regions.[10]

The parser records coordinates for blocks, allowing approximate structure recovery (sections, headings, tables) where present. Heuristics join hyphenated lines and preserve numbering and legal citations.(Li et al., 2021)

Quality gates run in CI. Schemas are validated, required fields are enforced, and checksums guard against corruption. Unit tests cover fetching, parsing, and serialization, and fail the job on regressions.(Pipino et al., 2022)

Historical coverage was built via a back-population pipeline. We iterate over archival indexes and date ranges, enqueue jobs in batches, checkpoint progress, and resume safely after interruptions.(Ben-David et al., 2019)

Transparency is prioritized. Run metadata, document counts, and last-updated timestamps are published to README badges. Commit messages summarize deltas, enabling clear, auditable provenance across releases.(Janssen et al., 2020)

5 Licensing and Access

All datasets and code are openly available to the public. The repositories are hosted on GitHub under the MIT License, which permits reuse, modification, and redistribution with attribution to the original author.[11]

This permissive model encourages transparency and collaboration. Researchers, developers, and institutions can integrate the datasets into their pipelines without restrictive terms or commercial barriers.(Janssen et al., 2020)

Each dataset repository includes structured metadata, versioned releases, and README files with descriptive statistics and provenance. All assets are mirrored on Hugging Face to ensure redundancy and faster global access.(Lhoest et al., 2021)

Public accessibility is a design principle. Automated GitHub Actions update metadata badges and commit summaries whenever new data are ingested. Users can clone, fork, or download any subset directly without authentication.[12]

Licensing notices are embedded in every dataset directory, clarifying reuse rights and responsibilities. The datasets intentionally avoid any personally identifiable information or restricted content to uphold ethical data-sharing standards.(Floridi and Taddeo, 2019)

The open license facilitates reproducible science and supports downstream applications in natural language processing, digital governance, and policy research. By ensuring public access, the project aligns with FAIR principles—Findable, Accessible, Interoperable, and Reusable.(Wilkinson et al., 2016)

6 Conclusion and Future Work

This project establishes an open, reproducible, and scalable foundation for Sri Lankan document datasets, spanning legal, governmental, and media sources. The pipeline integrates crawling, parsing, and versioning into a unified ecosystem for data-driven research.(Janssen et al., 2020)

The datasets already serve as a valuable resource for natural language processing, computational law, and policy analysis. They enable quantitative and qualitative insights into governance, lawmaking, and civic communication over time.(Wilkinson et al., 2016)

Future development focuses on three priorities:

First, expanding coverage by adding new datasets from additional government agencies, media sources, and historical archives.(Mishra et al., 2023)

Second, improving the linguistic accuracy of Sinhala and Tamil parsing, particularly for complex sentence structures and transliterated terms. Enhancements in tokenization, font handling, and multilingual embeddings are planned.(Conneau et al., 2022)

[9] https://www.selenium.dev/documentation/
[10] https://pymupdf.readthedocs.io
[11] https://opensource.org/licenses/MIT
[12] https://docs.github.com/actions

Third, integrating OCR parsing for PDFs with unstructured or scanned content. We are experimenting with deep-learning-based OCR pipelines that combine layout recognition and language modeling to recover high-quality text from low-quality sources.(Davis et al., 2021)

Together, these directions will further improve coverage, data quality, and usability, ensuring that the Sri Lanka Datasets initiative remains a sustainable open infrastructure for the region's digital and academic ecosystem.(Zuiderwijk and Janssen, 2020)

References

Omar Ashraf et al. 2022. A survey on mlops: Building reliable machine learning systems. *Journal of Systems and Software*.

Eytan Ben-David et al. 2019. Archival web data collection and preservation at scale. In *Proceedings of the Web Archiving Conference*.

Nathan Blaustein et al. 2020. Data versioning in practice: Challenges and solutions. *Data Science Journal*.

Alexis Conneau et al. 2022. Unsupervised cross-lingual representation learning at scale. *Transactions of the Association for Computational Linguistics*.

Benjamin Davis et al. 2021. Deep ocr: Neural architectures for document text recognition and layout parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Luciano Floridi and Mariarosaria Taddeo. 2019. What is data ethics? *Philosophical Transactions of the Royal Society A*.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. `https://skylion007.github.io/OpenWebTextCorpus`.

Marijn Janssen et al. 2020. Open data practices and governance: A comprehensive framework. *Government Information Quarterly*.

Quentin Lhoest et al. 2021. Datasets: A community library for natural language processing. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Chen Li et al. 2021. Document layout analysis via geometric and semantic segmentation. *Pattern Recognition Letters*.

Arjun Mishra et al. 2023. Scaling open data: Strategies for continuous expansion and maintenance. *Data Intelligence*.

Gautam Pant and Padmini Srinivasan. 2021. Web crawling best practices: Ethics, efficiency, and effectiveness. In *Proceedings of the Web Conference*.

Leo Pipino et al. 2022. Data quality assessment frameworks in data engineering. *Information Systems Frontiers*.

Victoria Stodden, Jennifer Seiler, and Zhaokun Ma. 2017. An empirical analysis of journal policy for reproducibility. *Proceedings of the National Academy of Sciences*.

Mark D. Wilkinson et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific Data*.

Anneke Zuiderwijk and Marijn Janssen. 2020. Sustainability of open data and data infrastructures: A research agenda. *Government Information Quarterly*.