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

Nuwan I. Senaratna
nuwans@alumni.stanford.edu

October 5, 2025

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.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 thedevelopment of natural language processing (NLP)and computational social science. Large corpora suchas Common Crawl[?], WikipediaDumps[?], and OpenWeb-Text[?]have powered models thatgeneralize across domains. However, these resourcesare dominated by data from high-resource languagesand global institutions.

Regional initiatives have sought to address this imbalance by creating domain-specific collections. Examples include the Indian Kanoon legal corpus[?], the OpenSubtitles multilingual dataset[?], and the African News Corpus[?]. Such datasets 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[?][?].

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

3.1 Hansard

A Hansard is the official verbatim record of parliamentary debates, preserving lawmakers' words and decisions for history, law, and public accountability.

- Time Updated: 2025-10-04 18:36:13
- Number of Documents: 1,665
- Date Range: 2006-02-01 to 2025-09-24
- Dataset Size: 17.9 GB

3.2 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.

- Time Updated: 2025-10-04 18:44:41
- Number of Documents: 10,146
- Date Range: 2012-04-23 to 2025-10-03
- Dataset Size: 10.3 GB

3.3 Supreme Court Judgements

A Supreme Court judgment is a binding legal decision that interprets the Constitution and laws, shaping justice, governance, and citizens' rights.

- Time Updated: 2025-10-04 18:44:12
- Number of Documents: 1,575
- Date Range: 2016-07-22 to 2025-10-03
- Dataset Size: 1.3 GB

3.4 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.

- Time Updated: 2025-10-04 18:28:49
- Number of Documents: 736
- Date Range: 2025-05-01 to 2025-10-03
- Dataset Size: 250.5 MB

3.5 Acts

A legal act is a law passed by Parliament that governs rights, duties, economy, and society, shaping daily life and national policy.

- Time Updated: 2025-10-04 18:42:02
- Number of Documents: 3,928
- Date Range: 1981-01-22 to 2025-09-22
- Dataset Size: 6.8 GB

3.6 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.

- Time Updated: 2025-10-04 18:37:26
- Number of Documents: 4,077
- Date Range: 2010-05-10 to 2025-10-26
- Dataset Size: 1.8 GB

3.7 Extraordinary Gazettes

An Extraordinary Gazette is an official government publication used to announce urgent laws, regulations, or public notices with immediate effect.

- Time Updated: 2025-10-04 18:39:51
- Number of Documents: 101,532
- Date Range: 2010-01-01 to 2025-10-03
- Dataset Size: 19.4 GB

3.8 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.

- Time Updated: 2025-10-04 18:08:09
- Number of Documents: 10,369
- Date Range: 2010-09-27 to 2025-09-22
- Dataset Size: 125.3 MB

3.9 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.

- Time Updated: 2025-10-04 18:18:09

- Number of Documents: 133

- Date Range: 2015-09-08 to 2025-07-30

- Dataset Size: 142.9 MB

3.10 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.

- Time Updated: 2025-09-26 08:23:47

- Number of Documents: 2,182

- Date Range: 2024-09-23 to 2025-09-24

- Dataset Size: 55.9 MB

3.11 News

A collection of news documents.

- Time Updated: 2025-10-04 18:41:30

- Number of Documents: 79,049

- Date Range: 2021-09-12 to 2025-10-04

- Dataset Size: 1.2 GB

3.12 Tourism Weekly Reports

Report on Weekly Tourist Arrivals to Sri Lanka.

- Time Updated: 2025-10-04 18:36:40

- Number of Documents: 33

- Date Range: 2023-01-01 to 2025-09-01

- Dataset Size: 91.5 MB

3.13 Tourism Monthly Reports

Report on Monthly Tourist Arrivals to Sri Lanka.

- Time Updated: 2025-10-04 18:36:59

- Number of Documents: 127

- Date Range: 2015-01-01 to 2025-08-01

- Dataset Size: 308.9 MB