

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

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

October 4, 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 $\sim 215,552$ 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
- 2 Related Work
- 3 Datasets
 - 3.1 Hansard
 - 3.2 Appeal Court Judgements
 - 3.3 Supreme Court Judgements
 - 3.4 Police Press Releases
 - 3.5 Acts
 - 3.6 Bills
 - 3.7 Extraordinary Gazettes
 - 3.8 Cabinet Decisions
 - 3.9 Treasury Press Releases
 - 3.10 Pmd Press Releases
 - 3.11 News
 - 3.12 Tourism Weekly Reports
 - 3.13 Tourism Monthly Reports
- 4 Data Collection Pipeline
- 5 Licensing and Access
- 6 Conclusion and Future Work