SINHALEGAL: A Benchmark Corpus for Information Extraction and Analysis in Sinhala Legislative Texts

Minduli Lasandi[a] **and Nevidu Jayatilleke**[b]
[a]School of Computing, Informatics Institute of Technology, Sri Lanka
[b]Department of Computer Science & Engineering, University of Moratuwa, Sri Lanka
minduli.20220374@iit.ac.lk, nevidu.25@cse.mrt.ac.lk

Abstract

SINHALEGAL introduces a Sinhala legislative text corpus containing approximately 2 million words across 1,206 legal documents. The dataset includes two types of legal documents: 1,065 Acts dated from 1981 to 2014 and 141 Bills from 2010 to 2014, which were systematically collected from official sources. The texts were extracted using OCR with `Google Document AI`, followed by extensive post-processing and manual cleaning to ensure high-quality, machine-readable content, along with dedicated metadata files for each document. A comprehensive evaluation was conducted, including corpus statistics, lexical diversity, word frequency analysis, named entity recognition, and topic modelling, demonstrating the structured and domain-specific nature of the corpus. Additionally, perplexity analysis using both large and small language models was performed to assess how effectively language models respond to domain-specific texts. The SINHALEGAL corpus represents a vital resource designed to support NLP tasks such as summarisation, information extraction, and analysis, thereby bridging a critical gap in Sinhala legal research.

1 Introduction

Legal documentation forms the backbone of modern legal systems. These documents provide an authoritative textual basis for legislation, interpretation and judicial decision making (Pietrosanti and Graziadio, 1999). As a result, legal texts require a high level of precision, consistency, and well-defined structure. They commonly contain complex sentence constructions and specialised legal vocabulary that differ from those found in general-purpose texts (Jayatilleke and Weerasinghe, 2025). These characteristics make legal documents more difficult to process automatically and highlight the need for specialised computational approaches to support various tasks.

The digitalisation of legal documents is an essential prerequisite for building reliable legal NLP systems (Boella et al., 2019). In many such contexts, legal texts are available only in scanned or image-based formats, introducing additional challenges related to *Optical Character Recognition* (OCR), layout preservation, and noise reduction. These constraints restrict broader access to legal information, reinforcing the need for systematically constructed, high-quality legal text datasets.

The Sinhala language is part of the Indo-European language family, specifically within the Indo-Aryan branch. It is the first language (L1) spoken by approximately 16 million people in Sri Lanka (De Silva, 2019). Sinhala features a unique script that descends from the Indian Brahmi script (Fernando, 1949). Although Sinhala is classified as a large institutional language by the Ethnologue categorisation system, it is considered a low-resource language (Category 02) according to the criteria outlined by Ranathunga and de Silva (2022).

In this study, we introduce SINHALEGAL [1], a dataset that includes Sinhala legal acts and bills from 1981 CE to 2014 CE. We provide a detailed discussion of the systematic steps taken in creating this dataset, which include data collection, preprocessing, filtration, and text extraction using OCR. This process is followed by manual post-processing and concludes with the creation of metadata.

2 Related Work

Researchers have prioritised the development of datasets containing legislative text in various languages and jurisdictions. These datasets support tasks such as summarisation, classification, information retrieval, and both diachronic and synchronic analysis, forming the basis for significant advancements in the field.

[1] https://bit.ly/4buVbKx

2.1 Sri Lanka Document Dataset

This repository is a comprehensive collection of official Sri Lankan governmental, legal and administrative documents spanning several decades and sources from authoritative institutions (Senaratna, 2025).

The repository contains official Sri Lankan governmental, legal, and administrative documents, including parliamentary records such as Hansard, Acts (1981–2025), Bills, and Extraordinary Gazettes (2010–2025), government communications such as police press releases and Treasury announcements, documents from the Disaster Management Centre, sector-specific reports from the Ministry of Fisheries, historic Central Bank annual reports, and educational publications from the Educational Publications Department. In total, the collection comprises 230,091 documents spanning from the 1950s to the 2020s.

SINHALEGAL focuses exclusively on the Acts and Bills contained in the Sri Lanka Document Dataset by Senaratna (2025). In the original repository, these legal documents are primarily available as PDF files within a much broader collection. Our study builds on this existing resource by extracting, cleaning, and structuring the Acts and Bills into a dedicated machine-readable corpus, which is further discussed in section 3.

2.2 Cambridge Law Corpus (CLC)

The *Cambridge Law Corpus* (CLC) is a substantial dataset for legal Artificial Intelligence research that comprises over 250,000 UK court cases (Östling et al., 2023). This corpus consists of 258,146 court cases drawn from 53 courts spanning over the 16th century to the 21st century It includes approximately 0.8 billion tokens, stored in XML format that captures both the full case body and rich metadata such as judge names, parties, and dates.

During the process of creation, the word and PDF files were cleaned, OCR processed through the Tesseract engine (Kay, 2007) and normalised into XML format and iteratively refined through a cycle query-driven methodology inspired by Voormann and Gut (2008). Due to the corpus size, only a stratified subset of 638 cases received expert-annotated outcomes.

The CLC dataset has become an important benchmark for advanced legal AI tasks. It supports applications such as case outcome prediction and long-form legal text processing. Previous studies have tested models such as RoBERTa (Liu et al., 2019) and GPT-4 (OpenAI, 2023) on this corpus, showing that long legal cases require models to handle difficult reasoning and strong semantic links across the text

2.3 Other Legal Datasets

Considering other legal datasets, BIGPATENT is one of the most influential large-scale datasets used for summarisation (Sharma et al., 2019). This consists of 1.3 million records of U.S. patent documents, sourced from Google Patents Public Datasets.[2]. Each entry pair has a full patent description and a human-written abstract (the gold-standard summary). The *Japanese Tort-case Dataset* (JTD), the first legal judgment prediction resource for the Japanese jurisdiction, consist of 3,477 real civil judgments focused on tort cases such as defamation and privacy infringement (Yamada et al., 2025).

Extending this line of multilingual legal-NLP work, the *Indian Legal Corpus* (ILC) by Trivedi et al. (2023) offers 3,000+ expert-written abstractive summaries of Indian legal judgments. Similarly, Nigam et al. (2025) introduced NyayaAnumana, a large-scale Indian legal judgment-prediction dataset with 702,945 processed cases from across the judiciary. Ma et al. (2021) introduced LeCaRD, a Chinese legal case-retrieval dataset, comprising 107 query cases and over 43,000 candidate cases drawn from Supreme People's Court criminal judgments.

Elaraby et al. (2024) created a curated research subset from the *Canadian Legal Information Institute* (CanLII[3]), an open-access repository of Canadian case law, containing 1,049 long-form judicial opinions with expert-written abstractive summaries, each annotated for argument roles including Issue, Reason, and Conclusion. Leitner et al. (2020) introduced a German legal Named Entity Recognition dataset under the EU (European) *Lynx*[4] project, with 750 court decisions, 54,000 manually annotated entities across 19 categories. And the survey done by Ariai et al. (2024) gives a review of the current landscape of NLP, focusing extensively on datasets and benchmarks in the legal domain.

Other datasets include LEGAL-UQA, the first Urdu-English legal question-answering dataset with 619 parallel question-answer pairs derived from Pakistan's constitution (Faisal and Yousaf, 2024);

[2] https://bit.ly/4rSS4BN
[3] https://www.canlii.org/en/
[4] http://www.lynx-project.eu/

the *Hindi Legal Documents Corpus* (HLDC) with 912,568 district court documents for bail prediction (Kapoor et al., 2022); the ILDC with 34,816 Supreme Court cases for judgment prediction (Malik et al., 2021); MultiLegalPile, a 689 GB multilingual corpus spanning 24 languages and 17 legal systems for LLM pretraining (Niklaus et al., 2024); and VLQA, a Vietnamese dataset with 3,129 expert-annotated questions for legal question answering and information retrieval (Nguyen et al., 2025).

Collectively, existing legal NLP datasets show substantial progress for high-resource languages such as English, Chinese, German, and Hindi, supporting tasks including judgment prediction, summarisation, and question answering. In contrast, Sinhala resources are largely limited to general text collections, with no dedicated legal-domain datasets. This gap motivates the present study, which aims to develop a foundational resource for Sinhala legal NLP.

3 Methodology

This methodology section describes the complete process followed to create the dataset. The workflow begins with collecting publicly available legal documents, followed by organising the files and extracting text from them using OCR. After extraction, several post-processing steps are implemented to correct errors and standardise the content. Finally, the structure of the dataset is established, and the inclusion of metadata information is discussed.

3.1 Data Acquisition

The initial stage involved gathering Sinhala legal documents from a publicly available repository on GitHub[5] (Senaratna, 2025), which is detailed in 2.1. These documents were available in PDF format and contain the Sinhala version of national laws.

At the time the repository was accessed *(August 2025)*, the documents were organised into four categories: Acts, Bills, Gazettes, and Extraordinary Gazettes. The collection included 1,500+ Acts, 1,300+ Bills, 6,300+ Gazettes and 35,000+ Extraordinary Gazettes. During the data acquisition process, Gazettes and Extraordinary Gazettes were excluded because many of the PDF files had multi-column layouts and dense formatting, which are known to reduce OCR accuracy (Fleischhacker et al., 2025). All the accessible Acts and Bills were downloaded to create a raw collection for further processing. In total, 2,865 PDFs were gathered. These documents covered a wide range of publication years, with Acts spanning from 1981 to 2025 and Bills from 2010 to 2025.

3.2 Data Organisation

All the downloaded legal documents were systematically organised to ensure consistency. For each metadata file processed, the corresponding Sinhala PDF (if available) was downloaded. Each file was saved using a descriptive and uniform naming convention automatically generated during the downloading process. The file name was constructed using three main components: the document type, publication date, and the cleaned description field *(doc_type_date_cleaned-description-or-id_si.pdf)*. If the generated file name exceeded the file system length limits, it was automatically truncated during the download process to ensure compatibility.

Documents published in languages other than Sinhala were also excluded. If duplicate files were available, they were automatically detected when downloading and skipped to prevent redundancy. The downloaded documents were arranged into a hierarchical directory structure based on the document type (Bills or Acts) and again into subfolders based on their publication year.

3.3 Text Extraction Using OCR

After the document organisation process, all the 2,865 documents were processed using Google Document AI [6] to perform OCR and extract the text from the PDF documents. A comparative study conducted by Jayatilleke and de Silva (2025b) evaluated the performance of various OCR engines. Among the five engines tested on a synthetically created dataset for Sinhala, Surya[7] emerged as the standout performer. However, when assessing a dataset of real scanned Sinhala documents, it became clear that Document AI achieved higher accuracy in its results (Jayatilleke and de Silva, 2025a).

Since Google Document AI has a limit of 15 pages per processing request, the documents that had more than 15 pages were divided into chunks of 15 pages each. This ensured that all documents, regardless of length, were fully processed without losing any content. During this process, information such as the OCR confidence, the number of

[5] https://github.com/nuuuwan/lk_legal_docs

[6] https://cloud.google.com/document-ai/

[7] https://github.com/VikParuchuri/surya

pages in the document, the number of chunks processed, document type, and published year was recorded.

After extraction, the text files were organised by publication years to maintain the chronological structure. The same filename convention used during the data acquisition was retained for consistency. This ensured that every extracted text file could easily be traced back into its original PDF and document category.

3.4 Data Filtration

We performed an *Exploratory Data Analysis* (EDA) on all the documents to assess the dataset's structure, distribution, and OCR quality before implementing any filtration steps. Based on the findings from the EDA, we applied several filtering steps to ensure that only high-quality, usable documents were retained for building the dataset.

3.4.1 Exploratory Data Analysis

Acts and Bills were analysed separately due to differences in length and formatting. Bills were generally longer with an average of 17.2 pages, compared to 14.3 pages for Acts. OCR performance across both categories was strong, with average confidence scores of 0.967 for Acts and 0.950 for Bills.

The dataset spans 46 years, beginning in 1981. The most legislative years were identified based on the combined number of Acts and Bills. The analysis showed that 2016 had 144 documents, 2021 had 190 documents, 2022 had 175 documents, 2023 had 168 documents, and 2024 had 161 documents. This trend highlights a substantial increase in document publication in recent years.

OCR confidence values were available for all 2,865 documents. Documents were categorised into three quality levels based on their OCR confidence scores: High quality for scores above 0.8, Medium quality for scores between 0.6 and 0.8, and Low quality for scores below 0.6. Of the 2,864 documents, 2,767 (96.6%) were classified as High Quality, 92 (3.2%) as Medium Quality, and 5 (0.2%) as Low.

Based on the page count, 1,825 documents (63.7%) were classified as Small (<10 pages), 857 (29.9%) as Medium (11–50 pages), and only 183 (6.4%) as Large (>50 pages). The page distributions for each category can be seen in Figure 1. More information on the EDA is discussed in the Appendix A.

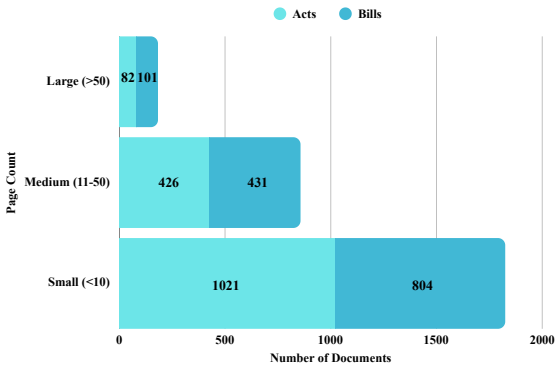


Figure 1: Distribution of document size of Acts and Bills

3.4.2 Filtration Strategy

The downloaded documents contained a total of 2,865 legal documents, of which 1,529 were Acts, and 1,336 were Bills. As a first step, the dataset was restricted based on publication year, retaining only Acts published between 1981 and 2014 and Bills published between 2010 and 2014, which resulted in 1,238 Acts and 155 Bills. Acts from the years 1992, 1996, and 1997, comprising 96 documents, were subsequently excluded due to visible double-sided printing that caused severe OCR errors. This ended with a count of 1,142 Acts. A page-count filter was then applied to the remaining documents, and those exceeding 50 pages were removed, as longer Acts and Bills often contained extensive tables and complex layouts that reduced OCR reliability; within the restricted year ranges, this step excluded 49 Acts and 13 Bills.

In addition to the page-count filtering, documents containing tables and multi-column layouts were removed, as these formats produced fragmented or unusable OCR text. Within the time ranges, this layout-based filtering excluded a further 26 Acts and only 1 Bill.

After the filtration process, 1,065 Acts and 139 Bills advanced to the next stages of research. This resulted in a total of 1,206 retained documents. A summary of the filtering stages, including the initial types of documents, page count categories, and the proportions of retained to removed documents, is presented in Figure 2. This collection of 1,206 high-quality legal documents provided a final dataset suitable for a viable post-processing procedure.

3.5 Document Post-Processing

After OCR, several post-processing steps were performed to ensure the dataset was clean and con-

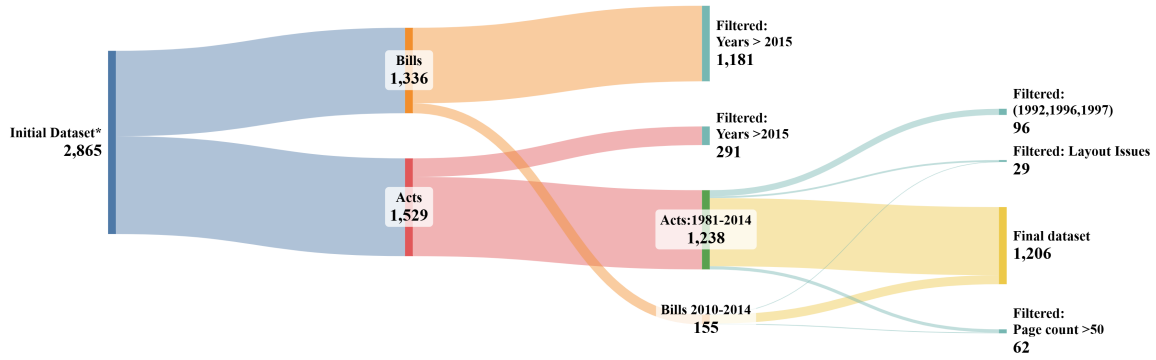


Figure 2: The flow of documents through each stage of the filtration process. *The initial dataset consists of documents that were available in the repository on the date of access (20th August 2025).

sistent. Despite having a high accuracy score in OCR for most of the documents, the extracted text still contained structural inconsistencies that required careful cleaning. These steps were carried out manually by the authors, who are native Sinhala speakers.

The post-processing included the following corrections to address identified issues based on document-level analysis and the work by Jayatilleke and de Silva (2025a):

Word-level corrections: OCR output often contained misspelt words, broken words or incorrect character substitution caused by poor quality. These were manually corrected to preserve the accuracy of the text.

Removal of footer content and page numbers: Legal documents included footers and page numbers that disrupted the flow of paragraphs. This included the removal of such footers and page numbers.

Removal of extra sentences: Most of the acts contained small sentences that could be seen outside the involved removing such sentences to maintain the flow and accuracy of the textragraphs. This step included removing such sentences to maintain flow, as well as the accuracy of the sentences.

Removal of seal content and prices: This step included removing the identified watermarks that were shown as official stamps. The prices mentioned at the start of the document were also removed.

Removal of repeated titles: Since document titles appeared multiple times per page, they were removed. The title on the title page and the first page of the document were kept.

Spacing errors: This involved the correction of the inconsistent spaces between sentences and paragraphs.

Removal of unnecessary characters: Occasionally, the extracted text contained characters such as underscores and dashes. These were removed since they interrupted the flow of the sentences and since they were not related to the content of the document.

These post-processing steps were conducted on all 1,065 Acts and 141 Bills in the SINHALEGAL corpus. Appendix B provides further discussion and examples of these steps.

3.6 Creating the Structure of SINHALEGAL

The dataset was first categorised into document type: Acts and Bills. Each document type was further organised into year-wise folders based on the year of publication. Within each year, separate directories were created for individual legal documents, with each directory named after the corresponding document. Each document directory contained the full text of the legal document and an accompanying metadata file with structured descriptive information. An example of the composed dataset structure is depicted in Figure 3. Furthermore, the creation of metadata files and their records is further described in Appendix C.

4 Evaluation

4.1 Corpus Statistics

First, we conducted a document-level evaluation of the SINHALEGAL dataset using the entire corpus. Our analysis reveals that, on average, each document contains 1,677 word tokens, with a median length of 1,213 word tokens. The distribution of

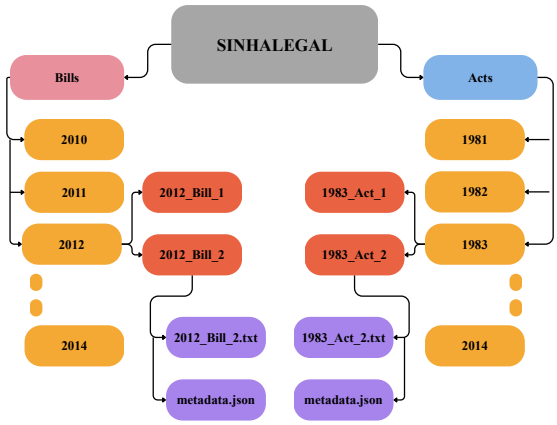


Figure 3: Structure of the SINHALEGAL dataset

word token counts demonstrates significant variability, ranging from short texts of 95 word tokens to lengthy documents that exceed 23,000 word tokens. In total, the corpus consists of 12.8 million characters, averaging 10,678 characters per document. The figures presented in Table 3 illustrate the heterogeneity of the legal documents.

Statistic	Value
Average word tokens per document	1,677
Median word tokens per document	1,213
token count range	95 - 23,430
Total characters	12,877,970
Average characters per document	10,678

Table 1: Summary statistics of the SINHALEGAL dataset

4.2 Lexical Diversity

The lexical diversity of the dataset was assessed through the *Type Token Ratio* (TTR) and the distribution of hapax legomena[8]. Tokenisation was performed using a simple rule-based whitespace tokeniser after the non-Sinhala characters were removed using Unicode range filtering. A summary of the total word tokens, vocabulary size, and TTR for Acts, Bills, and the overall dataset is provided in Table 2.

The dataset contains over two million tokens and 39,169 unique word types. TTR was length-normalised using Herdan's C (Ross and Herdan, 1960), computed as the ratio of the logarithm of vocabulary size to the logarithm of total tokens. It is observed that the Acts account for the majority of the word tokens (1,778,265) and show a TTR of 0.7315. In contrast, Bills are smaller in size (243,942 word tokens) and demonstrate a TTR of 0.7456.

Statistic	Acts	Bills	Total Corpus
Total word tokens	1,778,265	243,942	2,022,207
Vocabulary size	37,326	10,390	39,169
TTR	0.7315	0.7456	0.7284

Table 2: The number of total word tokens, vocabulary size and type token ratio for Acts, Bills and the total dataset.

The analysis of hapax legomena further highlights the distribution of rare words. Across the corpus, 18,074 word types (46.14% of the vocabulary) occur only once. Acts contain 17,632 hapax types (47.24% of their vocabulary), while Bills contain 4,026 hapax types (38.75%) as depicted in Table 3. This high proportion of single-occurrence words reflects the specialised nature of legal language, where frequent formulaic terms coexist with a long tail of rare items such as unique case names, bill titles, and technical terminology.

Documents	Hapax legomena	
	Count	Ratio
Acts	17,632	47.24%
Bills	4,026	38.75%
Total corpus	18,074	46.14%

Table 3: Distribution of hapax legomena across Acts, Bills, and the total corpus.

4.3 Word Frequency and Coverage

Word frequency analysis provides insight into the distribution of lexical items across the SINHALEGAL corpus. Coverage statistics show that a relatively small set of high-frequency words accounts for a substantial proportion of the text. In Acts, the top 20 words cover 23.00% of all word tokens, while in Bills the top 20 words cover 23.32%. Expanding to the top 50 words increases coverage to 35.04% in Acts and 35.53% in Bills, and the top 100 words account for nearly half of the corpus (45.89% in Acts and 46.39% in Bills). These figures highlight the repetitive and formulaic nature of Sinhala legal language.

It could be seen that conjunctions such as හෝ, සහ \ hɔ, sʌhʌ, particle words such as වන, ඒ, මේ \ vʌnʌ, e:, me: and other terms such as කිරීම, යුතු, සඳහා \ kɾi:mʌ, juθu, sʌnðʌhɑ: were repeated mostly. Bills show a similar distribution, with high coverage by a similar set of words. The top 10 identified frequent words are discussed in Appendix D.

[8]Word types that occur only once within a given corpus.

The coverage statistics demonstrate that Sinhala legal texts rely heavily on a small core vocabulary, while still maintaining lexical breadth through lower-frequency items. This shows that these documents are highly standardised in their functional framing, yet expansive in their incorporation of specialised terminology.

4.4 Calculating Text Accuracy and Structure

This dataset was evaluated using character-level and word-level error metrics, following a similar approach to CLC (Östling et al., 2023). For this, the corrected text was taken as the ground truth. Word Error Rate (WER) and Character Error Rate (CER) were computed with and without text normalisation. The results show WERs of 26.87% and 23.44%, and CERs of 24.07% and 24.06%, respectively. Structural differences such as line breaks were also analysed; further details are provided in Appendix F and G.

4.5 Named Entity Recognition

A rule-based *Named Entity Recognition* (NER) was implemented to identify salient entities in this dataset. Although various libraries exist for NER, they are not domain-specific and are incompatible for the legal domain (Badji, 2018), especially considering Sinhala. Therefore, a rule-based approach was implemented in `Python`, utilising regular expression matching and keyword-based rules to identify legal entities. This approach was designed to capture six major types of entities:

Date: Since dates are central to legal documents, the years and date expressions were captured using digit-based patterns (e.g., `\b\d 4}\b`) and extended rules for textual date formats.

Title: Institutional roles and titles such as ජනාධිපතිවරයා \ dʒʌnɑːðʰɪpʌθɪvʌrʌjɑː (President) and අමාත්‍යවරයා \ ʌmɑːθyvʌrʌjɑː (Minister) were identified and listed. This ensured capturing references to officials and positions within the legal system.

Organisation: These were extracted using a keyword dictionary of institutional terms frequently occurring in Sinhala legal texts such as අධිකරණය \ ʌðʰɪkʌrʌnʌjʌ (court of justice), පාර්ලිමේන්තුව \ pɑːrlɪmeːnθʊvʌ (parliament).

Law: Law names are highly formulaic, often ending with the word පනත \ pʌnʌθʌ (‘act’), and to avoid false positives, common prefixes such as දරන, එම \ ðʌrʌnʌ, ɛmʌ (bear with, that same), were removed.

Person: Personal names in legal texts are typically followed by honorifics such as මහතා, මහත්මිය \ mʌhʌθɑː, mʌhʌθmɪjʌ (Mister, Miss). This was used to capture up to two preceding words, ensuring that both single and compound names were recognised.

Amount: Monetary values are expressed with numerals followed by currency markers, such as රුපියල් \ rʊpɪjʌl or the letter රු \ rʊ (rupees), ensuring accurate identification of financial references.

The pipeline extracted a total of 28,937 entities across the corpus and the frequencies are depicted in Table 4. It could be seen that, on average, each document contains 24 entities, with a maximum of 361 and a minimum of 4 for a document which highlights the density of entities in the corpus.

Entity type	Number of Entities
Date [DATE]	13,532
Title [TITLE]	8,736
Organisations [ORG]	4,281
Law [LAW]	2,255
Person [PERSON]	126
Amount [AMOUNT]	7

Table 4: Number of entities extracted from the SINHALEGAL dataset

4.6 Topic Modelling

We performed topic modelling to explore the thematic structures within the dataset using *Latent Dirichlet Allocation* (LDA) (Blei et al., 2003), and it was implemented using the `Gensim` library in `Python`. Prior to this implementation, we performed standard preprocessing steps, including tokenisation and removal of stop words.

The corpus was preprocessed to ensure consistency and reduce noise. Texts were tokenised into word units, normalised to reduce orthographic variation (Manning, 2008). The Sinhala stop words were taken from an available public GitHub [9] repository that was created by Lakmal et al. (2020), and later modified manually with the common stop words in the SINHALEGAL dataset.

Topic coherence was computed to analyse the model’s behaviour across different values of K, with k=15 achieving the highest value. However, prior studies have shown that coherence alone is

[9] https://bit.ly/4tihUQj

insufficient for determining the optimal number of topics, as larger values of K may lead to over-clustering and unstable topic solutions (Greene et al., 2014). Therefore, the topic model was trained with ten topics to balance interpretability and coverage (Griffiths and Steyvers, 2004).

The results revealed recurring themes centred on the legislative acts (පනත \ pʌnʌθʌ), institutional references such as courts (අධිකරණය \ ʌðʰɪkʌrʌnʌjʌ), themes related to money (මුදල් \ muðʌl), pension (විශ්‍රාම \ ʋiʃˈraːmə), commissions (කොමිෂන් \ kɒmɪʃʌn) and elections (ඡන්ද \ tʃʰʌnðʌ). Representative word distribution for each topic is provided in the Appendix E.

4.7 Evaluation of Language Models

Perplexity is a standard evaluation metric in language modelling that measures how well a model predicts the next token (Meister and Cotterell, 2021). Perplexity was used to evaluate how well different language models handle legal domain text, and the scores were compared with a general Sinhala dataset named `MADLAD CulturaX` [10] (Aravinda et al., 2025).

To calculate the perplexity, we created balanced evaluation samples. The SINHALEGAL dataset was divided into sentences and clustered using the `KMeans` algorithm with k set to 10. We selected 200 sentences from each of the 10 clusters, resulting in a total of 2,000 sentences for our analysis. This clustering step ensured that sentences were grouped by similarity, allowing us to sample proportionally from each cluster. This approach preserved diversity in terms of sentence length, topics, and overall coverage.

For the general Sinhala dataset, which contained 10 million sentences, we randomly selected a sample of 100,000 sentences. We then applied the same clustering method as used for the legal dataset, ultimately extracting another set of 2,000 sentences.

To compare the differences between general Sinhala and legal Sinhala corpora, we selected two subsets of 2000 sentences each, created using `MADLAD CulturaX` and SINHALEGAL. Each sentence was tokenised using a custom Sinhala tokeniser that removes non-Sinhala characters and splits text into tokens, as detailed in subsection 4.2. We then computed normalised word distributions for both corpora and measured their divergence using the *Jensen-Shannon divergence* (JSD) (Lin, 2002), which quantifies the similarity between two probability distributions.

The computed JSD between the legal and general Sinhala corpora was 0.614, indicating a substantial difference in their word distributions. This confirms that legal Sinhala employs a distinct vocabulary and word usage compared to general language, complementing the perplexity-based evaluation and providing a quantitative measure of domain-specific linguistic characteristics.

For the perplexity-based evaluation, we considered several modern transformer architectures that support Sinhala, including `Llama-3.1-8B` [11](Kassianik et al., 2025), `Mistral-7B` [12] (Jiang et al., 2023), and `Falcon-7B` [13] (Almazrouei et al., 2023), `Deepseek-1.3B` [14](Guo et al., 2024), `DistilGPT-2` [15], a distilled variant of `GPT-2`. We also included `Gemma-2B` [16] (Team et al., 2024) a compact model released by Google that demonstrates competitive performance in resource-constrained environments. This diverse selection of models allowed us to examine how the size and architecture influence perplexity across legal and general Sinhala corpora.

Model Name	MADLAD CulturaX	SINHALEGAL
Llama-3.1	<u>3.05</u>	**2.55**
Deepseek-1.3B	3.30	2.94
Mistral-7B	3.68	3.18
Falcon-7B	**2.77**	<u>2.61</u>
DistilGPT-2	6.45	5.77
Gemma-2B	8.75	5.59

Table 5: Comparison of perplexity scores of the two datasets. **Bold:** indicates best performance and <u>Underline:</u> indicates the second best.

During the evaluation, all models exhibited lower perplexity scores on the SINHALEGAL corpus in comparison to the `MADLAD CulturaX` dataset. This suggests that domain-specific legal text is more predictable than general cultural content. Even though legal terms are more complex than general Sinhala, lower perplexity can likely occur due to repetitive structures and frequent patterns in texts (Yao et al., 2025). Frequent phrases such as "මෙම පනතට ඇතුළත් වන්නේ " meaning "this act contains" (mɛmʌ pʌnʌθʌtθʌ ʌθʊlʌθ vʌnneː), can

[10] https://huggingface.co/datasets/polyglots/MADLAD_CulturaX_cleaned

[11] https://bit.ly/49Ieoa4

[12] https://bit.ly/49nZLHD

[13] https://huggingface.co/tiiuae/falcon-7b

[14] https://bit.ly/3N76iiC

[15] https://huggingface.co/distilbert/distilgpt2

[16] https://huggingface.co/google/gemma-2b

be seen multiple times across documents. This repeated usage, along with the findings on the non-uniformity of word frequencies discussed in subsection 4.3, likely contributes to the lower perplexity scores observed when compared to the general Sinhala dataset.

`Llama 3.1` and `Falcon-7B` achieved the lowest perplexity on both datasets, followed by `Deepseek-1.3B`, indicating strong predictive performance. `Mistral-7B` also performed competitively, with slightly higher perplexity scores. As expected, the smaller models exhibited high perplexity values, in which `DistilGPT-2` produced moderate scores, while `Gemma-2B` showed the highest perplexity, particularly on the general corpus as presented in Table 5.

5 Conclusion

This study introduced SINHALEGAL, a Sinhala legal dataset designed to support research in legal NLP and information extraction tasks, specifically facilitating diachronic analysis of legal documents. The dataset includes a total of 1,206 legal documents, of which 141 are Bills ranging from 2010-2014 CE, and 1,065 are Acts ranging from 1981-2014 CE. The process of creating this dataset included performing OCR, filtering unwanted documents, and post-processing them manually to reduce noise and improve quality. The conducted evaluation included the lexical diversity, word frequency and coverage, NER and topic modelling to identify the number of entities and topics within the dataset. Finally, the perplexity scores were measured on selected language models to see how well the models respond to domain-specific data.

For future work, this dataset can be expanded with additional types of legal documents beyond Acts and Bills. Its utility can also be enhanced by applying further post-processing methods, such as segmenting documents into sections. SINHALEGAL fills a significant gap in legal NLP for Sinhala and provides a reliable foundation for future research.

Limitations

Scope restricted to Acts and Bills: This study considers only Acts and Bills. But the new repository in GitHub [17] mentioned in section 2.1 has been updated later and contains many additional categories.

[17] https://github.com/nuuuwan/lk_datasets

Temporal coverage for Acts and Bills: Although Acts in the repository span a broader range from 1981 to 2025, and Bills a range from 2010-2025, this analysis consists of Acts and Bills published between 2014.

Document structure not explicitly segmented: While some documents contain section boundaries, the documents in the dataset are provided as continuous text and are not consistently segmented into structural sections (e.g., preamble, definitions, clauses, schedules).

Language coverage limited to Sinhala: Although official English and Tamil versions of legal documents, including Acts and Bills, were available, this study focuses exclusively on the Sinhala versions of the documents.

Manual evaluation of the NER task: Due to the language-specific characteristics of named entities, automated evaluation methods were not fully applicable. Consequently, the NER task can be considered for manual evaluation.

Consideration of lengthy documents: For practical reasons for manual post-processing, documents longer than 50 pages were not considered in this study. They can be considered for future expansion of this dataset.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, and 1 others. 2023. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

HWK Aravinda, Rashad Sirajudeen, Samith Karunathilake, Nisansa de Silva, Rishemjit Kaur, and Surangika Ranathunga. 2025. Sinllama-a large language model for sinhala. In *2025 Moratuwa Engineering Research Conference (MERCon)*, pages 617–622. IEEE.

Farid Ariai, Joel Mackenzie, and Gianluca Demartini. 2024. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. *ACM Computing Surveys*.

Ines Badji. 2018. *Legal entity extraction with NER systems*. Ph.D. thesis, ETSI_Informatica.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Guido Boella, Luigi Di Caro, and Valentina Leone. 2019. Semi-automatic knowledge population in a legal document management system. *Artificial intelligence and Law*, 27(2):227–251.

Nisansa De Silva. 2019. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358*.

Mohamed Elaraby, Huihui Xu, Morgan Gray, Kevin Ashley, and Diane Litman. 2024. Adding argumentation into human evaluation of long document abstractive summarization: A case study on legal opinions. In *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 28–35, Torino, Italia. ELRA and ICCL.

Faizan Faisal and Umair Yousaf. 2024. Legal-uqa: A low-resource urdu-english dataset for legal question answering. *arXiv preprint arXiv:2410.13013*.

PEE Fernando. 1949. Palaeographical development of the brahmi script in ceylon from 3rd century bc to 7th century ad.

David Fleischhacker, Roman Kern, and Wolfgang Göderle. 2025. Enhancing ocr in historical documents with complex layouts through machine learning. *International Journal on Digital Libraries*, 26(1):3.

Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 498–513. Springer.

Thomas L Griffiths and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl_1):5228–5235.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Yu Wu, YK Li, and 1 others. 2024. Deepseek-coder: When the large language model meets programming–the rise of code intelligence. *arXiv preprint arXiv:2401.14196*.

Nevidu Jayatilleke and Nisansa de Silva. 2025a. Sidiac: Sinhala diachronic corpus. *arXiv preprint arXiv:2509.17912*.

Nevidu Jayatilleke and Nisansa de Silva. 2025b. Zero-shot OCR accuracy of low-resourced languages: A comparative analysis on Sinhala and Tamil. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing - Natural Language Processing in the Generative AI Era*, pages 471–480, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Nevidu Jayatilleke and Ruvan Weerasinghe. 2025. A hybrid architecture with efficient fine tuning for abstractive patent document summarization. In *2025 International Research Conference on Smart Computing and Systems Engineering (SCSE)*, pages 1–6. IEEE.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Arnav Kapoor, Mudit Dhawan, Anmol Goel, Arjun TH, Akshala Bhatnagar, Vibhu Agrawal, Amul Agrawal, Arnab Bhattacharya, Ponnurangam Kumaraguru, and Ashutosh Modi. 2022. Hldc: Hindi legal documents corpus. In *Findings of the association for computational linguistics: ACL 2022*, pages 3521–3536.

Paul Kassianik, Baturay Saglam, Alexander Chen, Blaine Nelson, Anu Vellore, Massimo Aufiero, Fraser Burch, Dhruv Kedia, Avi Zohary, Sajana Weerawardhena, and 1 others. 2025. Llama-3.1-foundationai-securityllm-base-8b technical report. *arXiv preprint arXiv:2504.21039*.

Anthony Kay. 2007. Tesseract: an open-source optical character recognition engine. *Linux Journal*, 2007(159):2.

Dimuthu Lakmal, Surangika Ranathunga, Saman Peramuna, and Indu Herath. 2020. Word embedding evaluation for Sinhala. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1874–1881, Marseille, France. European Language Resources Association.

Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. 2020. A dataset of German legal documents for named entity recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4478–4485, Marseille, France. European Language Resources Association.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Simon M. Lucas, Alex Panaretos, Luis Sosa, Anthony Tang, Shirley Wong, Robert Young, Kazuki Ashida, Hiroki Nagai, Masayuki Okamoto, Hiroaki Yamamoto, Hidetoshi Miyao, JunMin Zhu, WuWen Ou, Christian Wolf, Jean-Michel Jolion, Leon Todoran, Marcel Worring, and Xiaofan Lin. 2005. ICDAR 2003 robust reading competitions: entries, results, and future directions. 7(2):105–122.

Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. 2021.

Lecard: a legal case retrieval dataset for chinese law system. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2342–2348.

Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. Ildc for cjpe: Indian legal documents corpus for court judgment prediction and explanation. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 4046–4062.

Christopher D Manning. 2008. *Introduction to information retrieval*. Syngress Publishing,.

Clara Meister and Ryan Cotterell. 2021. Language model evaluation beyond perplexity. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5328–5339, Online. Association for Computational Linguistics.

Tan-Minh Nguyen, Hoang-Trung Nguyen, Trong-Khoi Dao, Xuan-Hieu Phan, Ha-Thanh Nguyen, and Thi-Hai-Yen Vuong. 2025. Vlqa: The first comprehensive, large, and high-quality vietnamese dataset for legal question answering. *arXiv preprint arXiv:2507.19995*.

Shubham Kumar Nigam, Deepak Patnaik Balaramamahanthi, Shivam Mishra, Noel Shallum, Kripabandhu Ghosh, and Arnab Bhattacharya. 2025. Nyayaanumana and inlegalllama: The largest indian legal judgment prediction dataset and specialized language model for enhanced decision analysis. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11135–11160.

Joel Niklaus, Veton Matoshi, Matthias Stürmer, Ilias Chalkidis, and Daniel Ho. 2024. Multilegalpile: A 689gb multilingual legal corpus. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15077–15094.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1.

Andreas Östling, Holli Sargeant, Huiyuan Xie, Ludwig Bull, Alexander Terenin, Leif Jonsson, Måns Magnusson, and Felix Steffek. 2023. The cambridge law corpus: A dataset for legal ai research. *Advances in Neural Information Processing Systems*, 36:41355–41385.

Ettore Pietrosanti and B Graziadio. 1999. Advanced techniques for legal document processing and retrieval. *Artificial intelligence and law*, 7(4):341–361.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the NLP world. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 823–848, Online only. Association for Computational Linguistics.

Alan Strode Campbell Ross and Gustav Herdan. 1960. Type-token mathematics : a textbook of mathematical linguistics.

Nuwan I Senaratna. 2025. Sri lanka document datasets: A large-scale, multilingual resource for law, news, and policy. *arXiv preprint arXiv:2510.04124*.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Pawan Trivedi, Digha Jain, Shilpa Gite, Ketan Kotecha, Anant Bhatt, and Nithesh Naik. 2023. Indian legal corpus (ilc): A dataset for a dataset summarizing indian legal proceeding using natural language. *Engineered Science*, 27:1022.

Holger Voormann and Ulrike Gut. 2008. Agile corpus creation. *Corpus Linguistics & Linguistic Theory*, 4(2).

Hiroaki Yamada, Takenobu Tokunaga, Ryutaro Ohara, Akira Tokutsu, Keisuke Takeshita, and Mihoko Sumida. 2025. Japanese tort-case dataset for rationale-supported legal judgment prediction. *Artificial Intelligence and Law*, 33(3):783–807.

Junchi Yao, Shu Yang, Jianhua Xu, Lijie Hu, Mengdi Li, and Di Wang. 2025. Understanding the repeat curse in large language models from a feature perspective. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7787–7815, Vienna, Austria. Association for Computational Linguistics.

A Exploratory Data Analysis

Figure 5 presents the yearly distribution of the documents, separated by Acts and Bills. This shows a clear increase in legislative document production in recent years, particularly after 2010. While Acts are consistently present throughout the entire time span, Bills become more prominent in later years.

The stacked representation further shows that Bills contribute significantly to the overall document count in peak years such as 2021, 2022 and

2023. Overall, this figure highlights how legal documentation has evolved, with an increasing volume and complexity in recent years, potentially reflecting changes in governance, policy focus, or administrative practices.

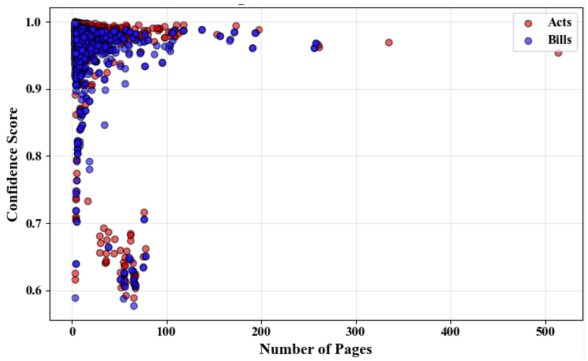


Figure 4: Relationship between the page count and OCR confidence

The scatter plot shown in Figure 4 depicts the relationship between document length (number of pages) and the OCR confidence scores for both Acts and Bills. It was seen that the shorter documents generally exhibited high OCR confidence values, often exceeding 0.9. As document length increases, greater variability in OCR confidence can be observed, mostly for documents exceeding 50 pages.

Longer documents frequently contain complex layouts such as multi-column formatting and tables. This can negatively impact the performance of OCR and lead to noisy text extraction. This pattern directly informed the page-count-based filtration criteria applied during dataset construction.

Figure 6 illustrates the distribution of page counts for Acts and Bills. Both document types generally have a low median page count, indicating that most acts and bills are relatively short. However, there is a notable presence of outliers, particularly among acts, with some extending beyond 500 pages. This suggests that while the majority of these documents are concise, acts tend to vary more widely in length and can be significantly longer than bills. The variability in page count highlights the diverse complexity and scope of legislative documents within these categories.

B Document Post-Processing

Seven errors were identified and addressed during the post-processing phase. This was carried out by the author, who is a native Sinhala speaker and is fluent in Sinhala

The acts contained small sections of sentences that could be seen next to paragraphs. These were often seen to be breaking the flow of the paragraphs and creating low accuracy of the meaning of the mentioned text. Hence, these were removed. Some examples are shown in Table 6.

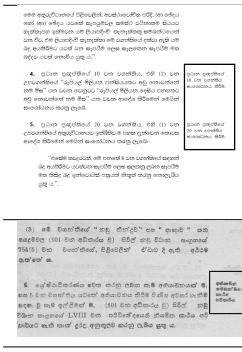
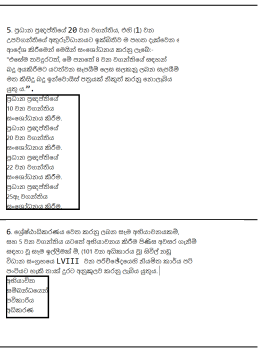
Extra Sentences	Extracted Text
	

Table 6: Examples of extra sentences in the scanned PDF and extracted text removed during post-processing.

Some of the years contained low-quality scanned PDFs, hence the output contained fragmented and misspelt words. Several paragraphs were seen to be broken, and this required manually correcting the spelling and order of the sentences. Some words could not be identified at all, and they too had to be manually typed and added into the extracted text. This took the most amount of time compared to the other errors that were fixed. Some of the above-mentioned errors and the corrected version can be seen in Table 17.

These documents also contained footers and page numbers that were not relevant to the legal content of the document. The page numbers were normally present at the end of each page, and the footers at the end of each document or section. Some of the Acts could be seen with seal content, which also does not have an impact on the content present in the legal document. Hence, these seal contents were removed from the documents. The titles of the document could be seen repeatedly on every page of the document. The title on the title page (the very first page) and the first page of the document were kept, and the others were removed to keep the flow of the document without the titles interrupting the paragraphs.

A large portion of the document exhibited inconsistent spacing. In some cases, excessive blank spaces appeared between lines, while in others, paragraphs, numbered lists and bullet points were merged with no spacing at all. These inconsis-

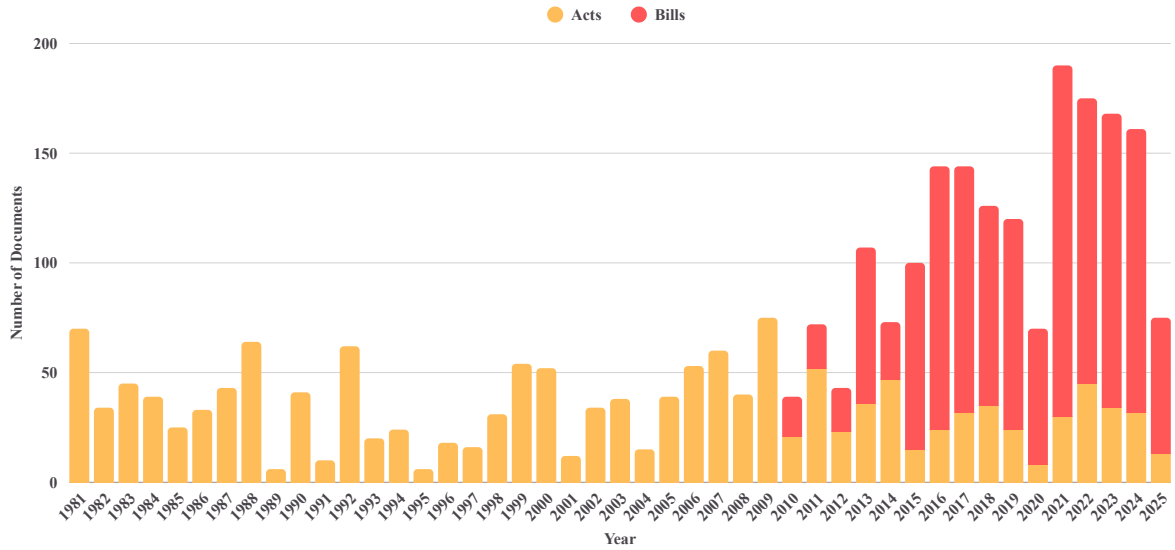


Figure 5: Distribution of Acts and Bills throughout the years from 1981-2025

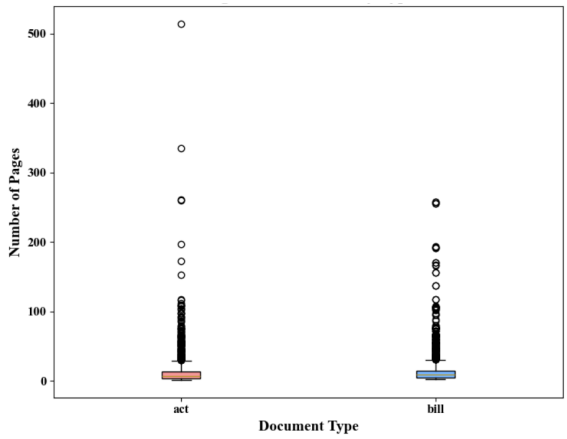


Figure 6: Boxplot showing the distribution of page counts for legal documents by type.

tencies made the text visually congested and difficult to read. During post-processing, proper line breaks and spacing were restored between paragraphs, numbered lists and bullet points to ensure a clear and well-structured document layout. An example of the mentioned spacing error and its respective corrections can be seen in Table 18.

The extracted text also contained various unnecessary characters produced by OCR misidentification, including underscores, dashes, semicolons, brackets, random English letters and other stray symbols *(e.g.,* `_ ; - { :] /)`. These artefacts appeared randomly throughout the text and did not carry any semantic meaning. All such characters were removed during post-processing to ensure clean and consistent documents.

C Creating Metadata Files

During the OCR process, document-level information was extracted and recorded for each legal document. Following post-processing, the relevant information related to each document was taken into separate metadata files and grouped accordingly. Maintaining document-level metadata also supports reproducibility and auditability, allowing OCR results and evaluation metrics to be traced back to their original source documents.


```
{
  "document_id": 148,
  "filename": "acts_1983-12-29_Rehabilitation_Levy_si",
  "type": "act",
  "year": "1983",
  "total_pages": 5,
  "chunks_processed": 1,
  "overall_document_confidence": 0.9883841406553984
}
```


Figure 7: Example of the metadata record for a document

The metadata files consisted of the document ID, the file name, the document type (Acts/Bills), the year of publication, the total number of pages, the number of chunks (1 chunk = 15 pages maximum) processed during the OCR process and the overall OCR confidence for the relevant document. An example of a metadata record is shown in Figure 7. The overall OCR confidence represents the

aggregate confidence score provided by the OCR engine, reflecting the estimated recognition reliability across all pages of the document.

D Word Frequency Coverage

This appendix provides detailed word frequency statistics to complement the analysis mentioned in Section 4.3. The Table 7 summarises the proportion of the corpus accounted for by the most frequent words. Coverage is calculated as the percentage of total word tokens represented by the top 20, 50, and 100 words in Acts and Bills. These figures illustrate the dominance of a small set of high-frequency function words in Sinhala legal texts.

Number of Word Covered	Acts	Bills
Top 20 words	23.00%	23.32%
Top 50 words	35.04%	35.53%
Top 100 words	45.89%	46.39%

Table 7: Coverage of the most frequent words in Acts and Bills.

Most Frequent words in Acts	Count
හෝ \ hɔ:	47,559
වන \ vʌnʌ	32,380
සහ \ sʌhʌ	29,775
යම් \ jʌm	28,440
විසින් \ vɪsɪn	25,494
කරනු \ kʌrʌnʊ	25,105
ඒ \ e:	24,947
යුතු \ jʊθʊ	24,027
ය \ jʌ	21,079
සඳහා \ sʌnðʌhɑ:	19,933

Table 8: The top 10 frequent words seen in Acts

Most Frequent words in Bills	Count
හෝ \ hɔ:	5,489
සහ \ sʌhʌ	5,309
වන \ vʌnʌ	4,136
කිරීම \ kɪri:mʌ	3,776
යුතු \ jʊθʊ	3,622
ය \ jʌ	3,372
යම් \ jʌm	3,249
විසින් \ vɪsɪn	3,235
සඳහා \ sʌnðʌhɑ:	3,124
කරනු \ kʌrʌnʊ	2,885

Table 9: The top 10 frequent words seen in Bills

The top 10 frequent words were taken from Acts and Bills separately. They are shown in Table 8 and Table 9. Unlike general Sinhala text, legal text contains complex wordings, but also contains a vast number of repetitive words, which are mostly conjunctions. As shown in the two tables, Acts and Bills mostly contained the same set of frequent words, just in different amounts. Bills mostly contained lower amounts than Acts since the number of Acts in the dataset is higher than that of Bills.

The top 10 most frequently seen bigrams[18] were also taken. This count was taken as an addition of both Acts and Bills. Words such as යුතු ය \ jʊθʊ jʌ, ශ්‍රී ලංකා \ ʒri: lʌŋkɑ:, මේ පනත \ mæ: pʌnʌθʌ could be seen to be used frequently across the document. The top 10 bigrams can be seen in Table 10.

As mentioned in the Section 4.7, these may be the reason legal documents had lower perplexity scores than that of general Sinhala text. These repetitive words and frequent structures are well known to reduce perplexity, as they make it easier for the model to guess the next word. (Yao et al., 2025).

Most Frequent Bigrams	Count
යුතු ය \ jʊθʊ jʌ	17,912
කරනු ලබන \ kʌrʌnʊ lʌbʌnʌ	9,004
අංක දරන \ ʌŋkʌ ðʌrʌnʌ	7,814
ලැබිය යුතු \ læbɪjʌ jʊθʊjʌ	7,453
කරනු ලැබිය \ kʌrʌnʊ læbɪjʌ	7,268
කරන ලද \ kʌrʌnʌ lʌðʌ	6,860
ශ්‍රී ලංකා \ ʒri: lʌŋkɑ:	6,463
විය යුතු \ vɪjʌ jʊθʊ	5,566
මේ පනත \ mæ: pʌnʌθʌ	5,202
පහත දැක්වෙන \ pʌhʌθʌ ðækvɛnʌ	5,089

Table 10: The top 10 frequent bigrams seen in SINHALEGAL

E Topic Modelling

Topic modelling was done in order to detect the main themes distributed within the dataset. Before this, the corpus text was tokenised into units. (Manning, 2008). A list of Sinhala stop words taken from a publicly available repository (mentioned in Section E), including conjunctions and other function words, was removed to reduce noise. The list was modified with some common stop words that were also seen in the SINHALEGAL dataset.

LDA (Blei et al., 2003) was selected for topic modelling because of its ability to uncover latent thematic structures in large corpora and its interpretability in corpus linguistics. The model was implemented using the Gensim library in Python. After exploratory runs, the number of topics was set to ten, balancing interpretability with coverage.

[18] A pair of consecutive written units such as letters or words

The revealing topics revealed recurring themes in Sinhala legal discourse. The word distribution for each topic is listed in Table 16. Across multiple topics පනත \ pʌnʌθʌ meaning “Act/Law”, emerged as a dominant term, reflecting the centrality of legislative acts in the corpus.

Other topics highlighted institutional references such as "Council" (සභාව \ sʌbɑːvʌ), "Court" (අධිකරණය \ ʌðʰɪkʌrʌnʌjʌ), "Commission (කොමිෂන් \ kɒmɪʃʌn)" and "Election (ඡන්ද \ tʃʰʌnðʌ)". Themes related to "Money" (මුදල් \ muðʌl), "Pension" (විශ්‍රාම \ viʃˈraːmə), "Towns" (නගර \ nʌgʌrʌ) were also to be seen among the listed words. These topics highlight the dominance of legislative references across the corpus, alongside institutional and procedural vocabulary.

F Calculating WER and CER

To calculate the WER and CER, a subset that consisted of 100 legal documents was taken. It contained over 184,000 words and 911,000 characters. Documents vary substantially in length, reflecting the natural diversity of legal texts, with word counts ranging from short to long legal documents. More information regarding the sample subset can be found in Table 11.

Statistic	Value
Documents analysed	100
Total words	184,011
Average words per document	1,840
Median words per document	1,295
Word count range	112 - 9,028
Total characters	911,524
Average characters per document	9,115

Table 11: Summary statistics of the evaluated document subset

Two types of assessments were conducted to provide a fair assessment. This included the original CER and WER, which include formatting differences and a normalised evaluation focusing solely on content errors.

First, an original evaluation was conducted where the raw OCR texts were compared directly against the post-processed texts. The randomly selected 100 documents were matched with the same files to the raw OCR files to calculate the WER and CER. The results showed a CER value of 24.07% and a WER value of 26.87%.

To assess the content-level fidelity independently of formatting, the post-processed texts were normalised (Lucas et al., 2005). The normalisation included collapsing multiple consecutive spaces into a single space, reducing multiple consecutive line breaks to a single line break, removing leading and trailing white spaces from each line, and discarding empty lines. After the normalisation, the CER and WER were recalculated using the same document subset. As shown in Table 12, normalisation reduced WER while leaving CER largely unchanged.

Metric	Original Evaluation	Normalized Evaluation
CER	24.07%	24.06%
WER	26.87%	23.44%

Table 12: Comparison of OCR evaluation metrics before and after text normalisation.

The minor change in CER suggests that the removal of formatting doesn’t affect the result of the calculation and that the earlier results of the CER were mostly accurate, while the WER dropped by nearly 3%.

G Structural Analysis

Other than calculating the WER and CER, a structural comparison was done. This was illustrated using a single document from the 100 documents already sampled to provide a concrete example (Document name: acts_1988-04 21_Evidence_Amendment_si.txt). It was seen that the post-processing effectively reduced redundant line breaks and spaces, which improves consistency and readability for future tasks.

Feature	Raw OCR	Post-processed	Difference
Line breaks	39	22	17
Spaces	173	94	79

Table 13: Structural comparison of raw OCR and post-processed text, showing the number of line breaks and spaces removed during post-processing.

To qualitatively illustrate OCR errors and post-processing corrections, a character-level comparison of the same representative document was done. The Table 14 highlights the removal of spurious digits, punctuation, redundant line breaks, and OCR-induced noise in Sinhala characters.

In addition to the document-level structural illustration, a corpus-level structural analysis was conducted over the full subset of 100 documents to quantify the overall impact of post-processing on document length characteristics. At the document

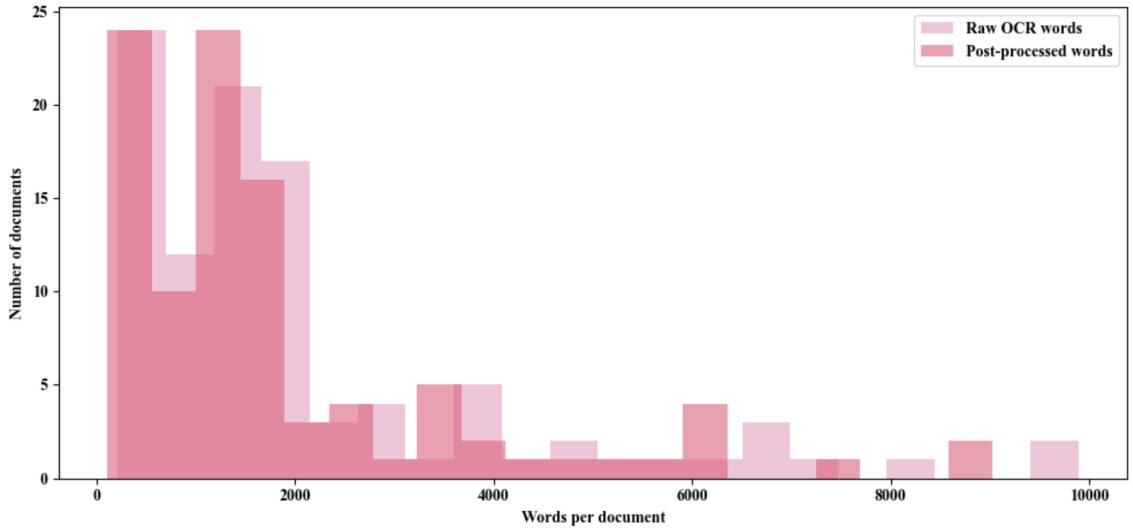


Figure 8: Distribution of words per document before and after OCR

Operation	Characters Added or Removed
Removed	2
Removed	8
Removed	1
Removed	˻˝ \ ʊ
Removed	◌ːↃ
Removed	ൈ \ nʌ
Added	\n
Added	ൽ \ jʌ

Table 14: The character-level differences between a raw OCR and post-processed text in a document

level, the average number of words per document decreased from 2,113 in the raw OCR to 1,840 after post-processing. Similarly, the average character count decreased from 12,506 to 11,007 characters per document. Removing unnecessary characters, stamp content, and footers results in a decrease in the number of words. Table 15 includes more information regarding the change of words and characters during the process. Figure 8 also demonstrates the distribution of the words per document before and after OCR.

Statistic	Raw OCR	Post-processed
Total words	211,317	184,011
Total characters	1,250,623	1,100,695
Average words per document	2,113	1,840
Average characters per document	1,2506	11007
Word count range	213–9891	112–9028

Table 15: Values comparing raw OCR and post-processed documents on the word and character count.

Taken together, these results demonstrate that the post-processing approach improves structural consistency throughout the corpus. While Table 13 and the qualitative examples highlight localised corrections within a single document, the aggregate statistics confirm that similar structural improvements are achieved across the entire dataset.

Topic	Word Distributions	Probabilities
Topic 1	පනත \ pʌnʌθʌ	0.034
	ප්‍රධාන \ prʌðʰaːnʌ	0.018
	ප්‍රඥාපනත \ prʌ/ɲaː/pʌnʌθʌ	0.018
	සංශෝධනය \ sʌnʃɔːðʰʌnʌjʌ	0.018
	උපවගන්නිය \ ʊpʌvʌgʌnθɪjʌ	0.017
Topic 2	සංස්ථාව \ sʌŋsθaːvʌ	0.020
	සංගමය \ sʌŋsθaːvʌ	0.015
	මහතා, මහත්මිය \ mʌhʌθaː, mʌhʌθmɪjʌ	0.012
	පනත \ pʌnʌθʌ	0.012
	ගමන \ gʌmʌnʌ	0.011
Topic 3	මණ්ඩලය \ mʌndʌlʌjʌ	0.021
	ආයතනය \ aːjʌθʌnʌjʌ	0.016
	කාර්ය \ kaːrjʌ	0.014
	පනත \ pʌnʌθʌ	0.012
	ප්‍රධාන \ prʌðʰaːnʌ	0.011
Topic 4	සභාව \ sʌbaːvʌ	0.039
	කොමිෂන් \ kɒmɪʃʌn	0.032
	පනත \ pʌnʌθʌ	0.014
	ව්‍යවස්ථාව \ vyʌvʌsθaːvʌ	0.009
	දක්වන \ ðʌkvʌnʌ	0.008
Topic 5	සංස්ථාව \ sʌŋsθaːvʌ	0.049
	පනත \ pʌnʌθʌ	0.019
	පරමාර්ථ \ pʌrʌmaːrθʰʌ	0.012
	ගමන \ gʌmʌnʌ	0.012
	සියලු \ sɪjʌlʊ	0.011
Topic 6	ඡන්ද \ tʃʰʌnðʌ	0.0019
	පනත \ pʌnʌθʌ	0.010
	සඳහන් \ sʌnðʌhʌn	0.009
	විශ්‍රාම \ ʋiʃˈraːmə	0.009
	උපවගන්නිය \ ʊpʌvʌgʌnθɪjʌ	0.008
Topic 7	පනත \ pʌnʌθʌ	0.0017
	අධිකරණය \ ʌðʰɪkʌrʌnʌjʌ	0.013
	මණ්ඩලය \ mʌndʌlʌjʌ	0.011
	කාර්ය \ kaːrjʌ	0.008
	සම්බන්ධයෙන් \ sʌmbʌnðʌjɛn	0.007
Topic 8	පනත \ pʌnʌθʌ	0.021
	කොටස \ kɒtʌsʌ	0.010
	උපවගන්නිය \ ʊpʌvʌgʌnθɪjʌ	0.010
	භාණ්ඩ \ bʰaːndʌ	0.010
	දක්වන \ ðʌkvʌnʌ	0.010
Topic 9	සභාව \ sʌbaːvʌ	0.037
	ප්‍රදේශය \ prʌðeːʃʌjʌ	0.026
	මණ්ඩලය \ mʌndʌlʌjʌ	0.015
	පනත \ pʌnʌθʌ	0.012
	නගර \ nʌgʌrʌ	0.011
Topic 10	පනත \ pʌnʌθʌ	0.016
	මුදල් \ muðʌl	0.011
	උපවගන්නිය \ ʊpʌvʌgʌnθɪjʌ	0.009
	ඇතළත \ æθʊlʌθʌ	0.009
	සම්බන්ධයෙන් \ sʌmbʌnðʌjɛn	0.008

Table 16: The word distributions and their word probabilities for the identified topics from topic modelling.

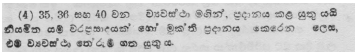
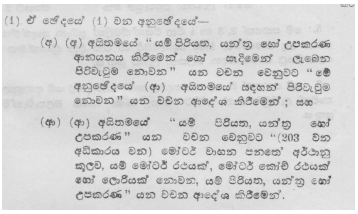
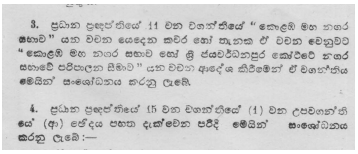
Scanned Document View	OCR Extracted Text	Corrected Text
	(4) 35, 36 සහ 40 වන නියමිත යම් වරප්‍රසාදයක් ව්‍යවස්ථා මගින්, ප්‍රදානය කළ යුතු යයි හෝ මුක්ති ප්‍රදානය කෙරෙන එම ව්‍යවස්ථා තේරුම් ගත යුතුය. ලෙස, පිළිගනු	(4) 35, 36 සහ 40 වන ව්‍යවස්ථා මගින්, ප්‍රදානය කළ යුතු යයි නියමිත යම් වරප්‍රසාදයක් හෝ මුක්ති ප්‍රදානය කෙරෙන ලෙස, එම ව්‍යවස්ථා තේරුම් ගත යුතුය.
	(1) ඒ ඡේදයෙ (1) වන අනුඡේදයේ - 99 හෝ (අ) (අ) අයිතමයේ '' යම් පිරියත, යන්ත්‍ර හෝ උපකරණ ආනයනය කිරීමෙන් හෝ සැදීමෙන් ලැබෙන පිරිවැටුම නොවන " යන වචන වෙනුවට “ මේ අනුඡේදයේ (ආ) අයිතමයේ සඳහන් පිරිවැටුම නොවන '' යන වචන ආදේශ කිරීමෙන්; සහ 66 හෝ (ආ) (ආ) අයිතමයේ යම් පිරියත, යන්ත්‍ර උපකරණ '' යන වචන වෙනුවට " (203 වන අධිකාරය වන) මෝටර් වාහන පනතේ අර්ථානු කූලව, යම් මෝටර් රථයක්, මෝටර් කෝච් රථයක් හෝ ලොරියක් නොවන, යම් පිරියත, යන්ත්‍ර හෝ උපකරණ '' යන වචන ආදේශ කිරීමෙන්.	(1) ඒ ඡේදයෙ (1) වන අනුඡේදයේ - (අ) (අ) අයිතමයේ '' යම් පිරියත, යන්ත්‍ර හෝ උපකරණ ආනයනය කිරීමෙන් හෝ සැදීමෙන් ලැබෙන පිරිවැටුම නොවන " යන වචන වෙනුවට “ මේ අනුඡේදයේ (ආ) අයිතමයේ සඳහන් පිරිවැටුම නොවන '' යන වචන ආදේශ කිරීමෙන්; සහ (ආ) (ආ) අයිතමයේ යම් පිරියත, යන්ත්‍ර උපකරණ '' යන වචන වෙනුවට " (203 වන අධිකාරය වන) මෝටර් වාහන පනතේ අර්ථානු කූලව, යම් මෝටර් රථයක්, මෝටර් කෝච් රථයක් හෝ ලොරියක් නොවන, යම් පිරියත, යන්ත්‍ර හෝ උපකරණ '' යන වචන ආදේශ කිරීමෙන්.
	3. ප්‍රධාන ප්‍රඥප්තියේ 11 වන වගන්තියේ " සභාව " යන වචන යෙදෙන කවර කොළඹ මහ නගර හෝ තැනක ඒ 46 වචන වෙනුවට කෝට්ටේ නගර 14 වන වගන්තිය ප්‍රධාන ප්‍රඥප තීයේ සංශෝධනය කිරීම. කොළඹ මහ නගර සභාව හෝ ශ්‍රී ජයවර්ධනපුර සභාවේ පරිපාලන සීමාව " යන වචන ආදේශ කිරීමෙන් ඒ වගන්තිය මෙයින් සංශෝධනය කරනු ලැබේ. 4. ප්‍රධාන ප්‍රඥප්තියේ 15 වන වගන්තියේ (1) වන උපවගන්තියෙ (ආ) ඡේදය පහත දැක්වෙන පරිදි මෙයින් සංශෝධනය කරනු ලැබේ :-	3. ප්‍රධාන ප්‍රඥප්තියේ 11 වන වගන්තියේ " කොළඹ මහ සභාව " යන වචන යෙදෙන කවර හෝ තැනක ඒ වචන වෙනුවට කොළඹ මහ නගර සභාව හෝ ශ්‍රී ජයවර්ධනපුර කෝට්ටේ නගර සභාවේ පරිපාලන සීමාව " යන වචන ආදේශ කිරීමෙන් ඒ වගන්තිය මෙයින් සංශෝධනය කරනු ලැබේ. 4. ප්‍රධාන ප්‍රඥප්තියේ 15 වන වගන්තියේ (1) වන උපවගන්තියෙ (ආ) ඡේදය පහත දැක්වෙන පරිදි මෙයින් සංශෝධනය කරනු ලැබේ :-

Table 17: Example of scanned PDF, OCR text, and corrected text for word-level corrections.

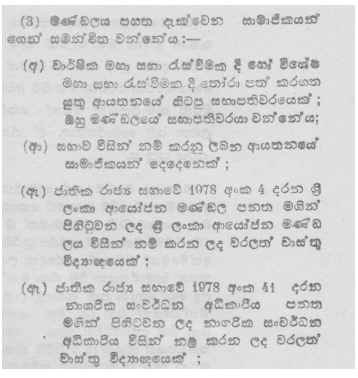
Scanned Document View	OCR Extracted Text	Corrected Text
	(3) මණ්ඩලය පහත දැක්වෙන සාමාජිකයන් ගෙන් සමන්විත වන්නේය:- (අ) වාර්ෂික මහා සභා රැස් වීමක දී හෝ විශේෂ මහා සභා රැස් වීමක දී තෝරා පත කරගත යුතු ආයතනයේ හිටපු සභාපතිවරයෙක්; ඔහු මණ්ඩලයේ සභාපතිවරයා වන්නේය; (ආ) සභාව විසින් නම් කරනු ලබන ආයතනයේ සාමාජිකයන් දෙදෙනෙක්; (ඇ) ජාතික රාජ්‍ය සභාවේ 1978 අංක 4 දරන ශ්‍රී ලංකා ආයෝජන මණ්ඩල පනත මගින් පිහිටුවන ලද ශ්‍රී ලංකා ආයෝජන මණ්ඩලය විසින් නම් කරන ලද වරලත වාසතු විද්‍යාඥයෙක්; (ඈ) ජාතික රාජ්‍ය සභාවේ 1978 අංක 44 දරන නාගරික සංවර්ධන අධිකාරිය පනත මගින් පිහිටුවන ලද නාගරික සංවර්ධන අධිකාරිය විසින් නම් කරන ලද වරලත වාස්තු විද්‍යාඥයෙක්;	(3) මණ්ඩලය පහත දැක්වෙන සාමාජිකයන් ගෙන් සමන්විත වන්නේය:- (අ) වාර්ෂික මහා සභා රැස් වීමක දී හෝ විශේෂ මහා සභා රැස් වීමක දී තෝරා පත කරගත යුතු ආයතනයේ හිටපු සභාපතිවරයෙක්; ඔහු මණ්ඩලයේ සභාපතිවරයා වන්නේය; (ආ) සභාව විසින් නම් කරනු ලබන ආයතනයේ සාමාජිකයන් දෙදෙනෙක්; (ඇ) ජාතික රාජ්‍ය සභාවේ 1978 අංක 4 දරන ශ්‍රී ලංකා ආයෝජන මණ්ඩල පනත මගින් පිහිටුවන ලද ශ්‍රී ලංකා ආයෝජන මණ්ඩලය විසින් නම් කරන ලද වරලත වාසතු විද්‍යාඥයෙක්; (ඈ) ජාතික රාජ්‍ය සභාවේ 1978 අංක 44 දරන නාගරික සංවර්ධන අධිකාරිය පනත මගින් පිහිටුවන ලද නාගරික සංවර්ධන අධිකාරිය විසින් නම් කරන ලද වරලත වාස්තු විද්‍යාඥයෙක්;

Table 18: Example of scanned PDF, extracted text, and corrected text for spacing errors.