



Doctracer: Intelligent Document Information Extraction and Processing Toolkit for Sri Lankan Extraordinary Gazettes

*A.U. Amarasinghe¹, R.N.C. De Zoysa¹, P.K.H.D. De Silva¹, K.D. Ranaweera¹,
K.L.K. Sudheera^{1*}, V. Abeykoon²*

¹ Department of Electrical and Information Engineering, Faculty of Engineering,
University of Ruhuna, Sri Lanka

² Lanka Data Foundation

**Corresponding author: kushan@eie.ruh.ac.lk*

Sri Lankan Extraordinary Gazettes serve as the primary legal mechanism for publishing government structures, including ministries, their allocated departments, and the Acts and functions assigned to each ministry. Following national elections and subsequent administrative revisions, these structures are frequently amended through complex legal documents involving transfers of departments, reassignment of laws, and structural reorganizations. Due to the unstructured nature, dense layouts, and heavy numerical formatting of gazette documents, manual analysis and automated extraction remain highly challenging. This research introduces Doctracer, an intelligent document information extraction and processing toolkit designed specifically for Sri Lankan Extraordinary Gazettes. This system contains customized layout aware parsing pipeline specially optimized for these gazette-specific multi column and tabular structures. This pipeline is closely aligned with controlled, few shot learning optimized Large Language Model (LLM) prompting for accurate information extraction. Extracted content is converted into structured format and modeled using a domain-specific graph schema consisting of key entities such as Gazette, Minister, Department, Law, and Function, along with their semantic relationships. Then this structured data is stored in Neo4j Graph Database to preserve temporal dependencies and enable interactive visualization. Doctracer also introduces an automated amendment comparison module that accurately detects insertions, updates, deletions, and renumbering operations between base and amendment gazettes. Comparative experimental evaluation against widely used LLMs demonstrates that Doctracer achieves significantly higher accuracy in both base gazette extraction and amendment change detection tasks. For base gazette processing, Doctracer achieved 95% extraction accuracy, outperforming ChatGPT (79%), DeepSeek (87%), and Gemini (56%), particularly in handling complex multi-column layouts and maintaining structural consistency. In amendment change detection, Doctracer achieved 100% accuracy in identifying all modification operations, matching the performance of Gemini and DeepSeek while ChatGPT showed no successful detections in this task.

Keywords: Information extraction, Knowledge graph, Layout-aware parsing, Legal document analysis, Natural language processing