SRI LANKAN LEGAL INFORMATION RETRIEVAL SYSTEM

Chanuka Asel Jayasooriya (IT20250942)

B.Sc. (Hons) Degree in Information Technology Specializing in Data Science

Department of Information Technology Sri Lanka Institute of Information Technology

Sri Lanka

SRI LANKAN LEGAL INFORMATION RETRIEVAL SYSTEM

Chanuka Asel Jayasooriya

(IT20250942)

Dissertation submitted in partial fulfillment of the requirements for the Bachelor of Science (Hons) in Information Technology Specializing in

Data Science

Department of Information Technology

Sri Lanka Institute of Information Technology Sri Lanka

# DECLARATION

I declare that this is my own work, and this dissertation does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any other university or Institute of higher learning, and to the best of my knowledge and belief, it does not contain any material previously published or written by another person except where the acknowledgment is made in the text. Also, I hereby grant to Sri Lanka Institute of Information Technology the non-exclusive right to reproduce and distribute my dissertation in whole or part in print, electronic, or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

|  |  |  |
| --- | --- | --- |
| **Name** | **Student ID** | **Signature** |
| Jayasooriya C. A. | IT20250942 |  |

The supervisor/s should certify the proposal report with the following declaration. The above candidates are carrying out research for the undergraduate Dissertation under my supervision.

Signature of the supervisor: Date:

# ABSTRACT

Law is a domain heavily reliant on comprehensive research, and this holds true for Sri Lanka as well. In an era of advancing technology, the accessibility of legal information in Sri Lanka has increased significantly. However, this has given rise to the challenge of accurately retrieving information for well-informed legal decision-making.

This research presents a developed Legal Decision Support System tailored to the Sri Lankan legal landscape. Leveraging natural language processing techniques and deep learning algorithms, this system efficiently handles legal queries, providing prompt and precise responses. Its primary objective is to streamline the retrieval of relevant legal data and documents, fostering informed decision-making. Moreover, the system is designed for universal accessibility, aiming to democratize legal information access in Sri Lanka.

The implemented Legal Decision Support System yields substantial benefits for the Sri Lankan legal community. It enhances access to legal information, streamlines legal research, and alleviates the workload of legal practitioners. This system has the potential to revolutionize legal research in Sri Lanka and serve as a blueprint for similar regional legal systems.

**Key Words: Legal Information Retrieval, Legal Decision Support, Document Extraction, Natural Language Processing, Transformer Based Models**

# ACKNOWLEDGEMENT

I want to take this moment to express my deep appreciation to all those who have played a pivotal role in the successful completion of the research project, Sri Lankan Legal Information Retrieval System.

First and foremost, I extend my heartfelt thanks to my supervisor, Dr. Lakmini Abeywardhana, and my co-supervisor, Dr. Dharshana Kasthurirathna. Their unwavering guidance and support throughout every phase of this project have been invaluable. Their expertise in Information Retrieval, Natural Language Processing, and Machine Learning has been instrumental in shaping this research into a comprehensive and well-structured system.

I also want to express my gratitude to our external supervisor, Miss Ruvani Maldeniya, for her invaluable legal domain expertise and guidance. Her insights and suggestions have played a critical role in ensuring that the system aligns with the needs of legal professionals.

Additionally, I'd like to thank Mr. Bilaal Marikar and the students at Sri Lanka Law College for their active participation and valuable feedback during the data gathering and requirements gathering phase. Their contributions have been essential in refining the research project into an efficient and effective system.

A special thanks goes out to my dedicated team, whose hard work and unwavering support were crucial during the system's development. Their commitment and contributions, including innovative ideas and constructive feedback, were integral to the successful completion of the project.

Lastly, I'd like to acknowledge the Faculty of Computing and the Comprehensive Design and Analysis Project (CDAP) team at the Sri Lanka Institute of Information Technology (SLIIT) for granting me the opportunity to conduct this research project. Their consistent guidance and support have been indispensable in bringing this project to fruition.

# TABLE OF CONTENTS

[DECLARATION i](#_bookmark0)

[ABSTRACT ii](#_bookmark1)

[ACKNOWLEDGEMENT iii](#_bookmark2)

[LIST OF FIGURES vii](#_bookmark3)

[LIST OF TABLES viii](#_bookmark4)

[LIST OF ABBREVIATIONS ix](#_bookmark5)

1. [INTRODUCTION 1](#_bookmark6)
   1. [Background and Literature Survey 1](#_bookmark7)
   2. [Research Gap 4](#_bookmark9)
      1. [The problems of geographically specific systems 5](#_bookmark10)
      2. [Problems in low accuracy similarity calculations 6](#_bookmark11)
      3. [The problem of low performance in pre-trained language models 7](#_bookmark12)
   3. [Existing Similar Systems and Feature Comparison 8](#_bookmark14)
2. [RESEARCH PROBLEM 10](#_bookmark16)
3. [OBJECTIVES 13](#_bookmark20)
   1. [Main Objective 13](#_bookmark21)
   2. [Sub Objectives 14](#_bookmark22)
      1. [Data extraction, pre-processing, and structuring 14](#_bookmark23)
      2. [Question answering model building 14](#_bookmark24)
      3. [Pipeline building and integration 15](#_bookmark25)
4. [METHODOLOGY 16](#_bookmark26)
   1. [Requirement Gathering and Analysis 16](#_bookmark27)
      1. [Functional requirements 18](#_bookmark28)
      2. [Non-functional requirements 18](#_bookmark29)
   2. [Feasibility Study 19](#_bookmark30)
   3. [System Designs 21](#_bookmark31)
      1. [Overall system architecture diagram 21](#_bookmark32)
      2. [Component-specific system diagram 25](#_bookmark34)
   4. [Methodology 30](#_bookmark36)
      1. [Data acquisition and preparation 30](#_bookmark37)
      2. [Model building and training 33](#_bookmark41)
   5. [Implementation of the system 39](#_bookmark46)
      1. [Building the question-answering pipeline 39](#_bookmark47)
      2. [Frontend implementation and integration 40](#_bookmark50)
   6. [System Testing 42](#_bookmark53)
      1. [Unit testing 42](#_bookmark54)
      2. [Integration testing 43](#_bookmark58)
      3. [User Acceptance Testing (UAT) 44](#_bookmark61)
   7. [Commercialization of the System 45](#_bookmark62)
      1. [Market opportunity 45](#_bookmark64)
      2. [Market competition 46](#_bookmark66)
      3. [Commercialization strategy 47](#_bookmark67)
5. [RESULTS AND DISCUSSION 49](#_bookmark69)
   1. [Results 49](#_bookmark70)
      1. [Reader model (Doc2Vec) 49](#_bookmark72)
      2. [Retriever model (Finetuned RoBERTa) 50](#_bookmark75)
   2. [Research Findings 52](#_bookmark79)
   3. [Discussion 54](#_bookmark80)
      1. [Limitations and challenges faced 54](#_bookmark81)
      2. [Future Work 55](#_bookmark82)
6. [CONCLUSION 57](#_bookmark83)

[REFERENCES 59](#_bookmark84)

[APPENDIX A: SURVEY RESULTS 63](#_bookmark85)

[APPENDIX B: CLIENT APPROVAL LETTER 66](#_bookmark86)

[APPENDIX C: PERSONNEL AND FACILITIES 67](#_bookmark87)

[APPENDIX D: DATASET APPROVAL 68](#_bookmark88)

[APPENDIX E: TURNITIN REPORT 69](#_bookmark89)

# LIST OF FIGURES

[Figure 1 Hierarchy of courts: Sri Lanka 2](#_bookmark8)

[Figure 2 Pain points in the legal industry 10](#_bookmark17)

[Figure 3 Issues faced during legal research 11](#_bookmark18)

[Figure 4 Research problem identification survey 12](#_bookmark19)

[Figure 5 Overall System Architecture 21](#_bookmark33)

[Figure 6 Component-specific system diagram 25](#_bookmark35)

[Figure 7 Target data source – Narcotic Act, No 1 of 2008 30](#_bookmark38)

[Figure 8 Manual annotation of documents 31](#_bookmark39)

[Figure 9 SQuAD format 32](#_bookmark40)

[Figure 10 Training code segment - Doc2Vec model 35](#_bookmark42)

[Figure 11 Retriever code segment - Doc2Vec model 36](#_bookmark43)

[Figure 12 Pretrained RoBERTa model import 38](#_bookmark44)

[Figure 13 Hyperparameter tuning - RoBERTa model 39](#_bookmark45)

[Figure 14 Model saving 40](#_bookmark48)

[Figure 15 QnA pipeline 40](#_bookmark49)

[Figure 16 Question answering system UI 41](#_bookmark51)

[Figure 17 Answer routing to UI 41](#_bookmark52)

[Figure 18 System logo 45](#_bookmark63)

[Figure 19 Target audience 46](#_bookmark65)

[Figure 20 Commercialization poster 48](#_bookmark68)

[Figure 21 Accuracy metric 49](#_bookmark71)

[Figure 22 F1 Score metric 49](#_bookmark73)

[Figure 23 Transformer based model performance comparison 50](#_bookmark76)

# LIST OF TABLES

[Table 1 Research gap illustration 8](#_bookmark13)

[Table 2 Existing systems and feature comparison 9](#_bookmark15)

[Table 3 Unit test case 01 - Retriever model 42](#_bookmark55)

[Table 4 Unit test case 02 - Reader model 42](#_bookmark56)

[Table 5 Unit test case 03 - Reader model 43](#_bookmark57)

[Table 6 Integration test case 01 43](#_bookmark59)

[Table 7 Integration test case 02 44](#_bookmark60)

[Table 8 Results - Retriever model 50](#_bookmark74)

[Table 9 Results - Reader model 51](#_bookmark77)

[Table 10 Training and evaluation losses - Reader model 52](#_bookmark78)

# LIST OF ABBREVIATIONS

|  |  |
| --- | --- |
| **Abbreviation** | **Description** |
| CDAP | Computing and the Comprehensive Design and Analysis Project |
| SLIIT | Sri Lanka Institute of Information Technology |
| NLR | New Law Reports |
| SLR | Sri Lankan Law Reports |
| IR | Information Retrieval |
| BERT | Bidirectional Encoder Representations from Transformers |
| NLP | Natural Language Processing |
| SRS | Software Requirements Specification |
| RoI | Return on Investment |
| CPU | Central Processing Unit |
| NER | Named Entity Recognition |
| LKG | Legal Knowledge Graph |
| RoBERTa | Robustly Optimized BERT Pretraining Approach |
| SQuAD | Stanford Question Answering Dataset |
| CSV | Comma Separated Values |
| TF-IDF | Term Frequency - Inverse Document Frequency |
| GPU | Graphics Processing Unit |
| UI | User Interface |
| UAT | User Acceptance Testing |

# INTRODUCTION

## Background and Literature Survey

The law is the backbone of modern society, humans have implemented the legal framework and all its constituents to form a modern civilized society [1]. Modern human civilization is highly dependent on its legal system for order and prevention of chaos [2]. The protection, preservation, and implementation of the law are some of the most important duties of man. The law, which is the foundational truth of modern civilizations, has been extensively documented and theoretically practiced, taught, and revised throughout history. Thus, the legal framework is made up of a large number of documents and information, which has made it way beyond the capabilities of humans to entirely consume and utilize this information [3].

With this plethora of documented legal information, it has become challenging to identify and use proper legal knowledge. Even though the increase in the availability of documented and digitally available legal knowledge seems very helpful, this has become very counterproductive as legal professionals and lay people have become flooded with information which has made the extraction of the necessary information extremely tedious and inefficient. According to an article by Paul Hemp regarding the “Information Overload” in the Harvard Business Review, due to modern digitization and technological advancements, the availability of documented information has become so abundant that it has surpassed the human brain's capacity of information absorption during a single person’s lifetime [3]. This has brought the need for comprehensive information retrieval systems into the modern world. It has become an extremely dire need in the legal sector too.

In the Sri Lankan context, a study by the author Aquinas V. Tambimuttu about the Sri Lankan legal system highlights that the judicial system of Sri Lanka comprises the Supreme Court, the Court of Appeal, the High Court, district courts, magistrate's courts, and primary courts [4] [5]. The Parliament of Sri Lanka is responsible for the creation of laws in Sri Lanka as Acts of Parliament. The judicial system is responsible for the

execution of the laws. The Sri Lankan legal system is a mixed system comprising the English common law, Roman-Dutch Law, Kandyan Law, Thesawalamai Law, and Muslim Law [4]. The Sri Lankan legal information is present in officially documented formats such as Amendments, Statutes, Case laws, New Law Reports (NLR), Sri Lankan Law Reports (SLR), Judgements, Gazettes, and Agreements [6]. The management of legal documents and their information is a very tedious task in the Sri Lankan context as the digitization of these documents was initiated in the recent past. This has given rise to a new set of problems like:

* The inconsistency of legal information.
* Unavailability of historic legal documents, especially case laws.
* Ineffective management of historic case laws.
* Availability of a large corpus of improperly managed legal documents.



*Figure 1 Hierarchy of courts: Sri Lanka*

Source: Roar Media, Report Article (June 13, 2017)

The availability of a vast number of legal documents and information has counter- intuitively made the legal literacy of the country very sparse. According to a study done by the two authors R. O. Moorthage and H. P. K. N. Hewawasam on the Impact of Legal Literacy on Crime Control and Prevention with Special Reference to Selected Communities in Rathnapura Police Division in Sri Lanka, it has been made clear that

citizens who are ignorant of the legal framework, rights, and punishments are the majority involved in illegal activities. This study has shown that due to legal illiteracy, 65% have become victims of crimes, 90% are involved in illegal activities, and due to ignorance about the criminal justice system and its procedures 85% face significant difficulties when dealing with law enforcement [2]. This could be overcome with the democratization and decentralization of legal knowledge and increasing the ability to access legal information to lay people.

Apart from the above, legal information access has become an arduous task for the personnel working and involved in the legal industry: Lawyers, judges, law students, and legal professionals. According to authors Nina Evans and James Price it has been highlighted in their study named “Managing information in law firms: changes and challenges”, that the abundance of legal information and documents due to technological advancement has made legal firms inefficient and increased their costs of data management. It is also highlighted that proper management of legal information would be beneficial for upholding the law to a higher standard. The authors have identified this as a global phenomenon and the legal industry all around the world suffers from improper information management [7]. According to another study done by the author H.M.P.P Karunarathna, about the use of electronic resources by law degree students at Anuradhapura Regional Centre of the Open University of Sri Lanka, a majority of the law students (93%) use digital resources in their studies, yet a majority of them suffer due to the unavailability of proper legal information management [8]. The issue is pertinent to the question of how effective and efficient legal education in the country is. The unavailability of proper and efficient access to accurate, timely, relevant legal documents, knowledge, and information is detrimental to the education of the students.

Legal knowledge data availability and management have a direct impact on legal professionals. Lawyers and judges must be in a constant state of knowledge renewal, and are at the biggest disadvantage due to the unavailability of proper access to accurate and consistent legal information and documents. These legal professionals have been forced

to resort to manual information searches using outdated legal archives to obtain the necessary legal information. Lawyers have the responsibility of referring and studying cases upon cases that are representative of the legal cases that they would be representing in court. The need for traversing through a plethora of homogenous legal documents would be taxing and arduous. Thus, the need for proper legal information retrieval systems that address the variety of information needs of various end users has been highlighted in the study done by the two authors A. Cammelli and E. Fameli. This study sheds light on the importance of legal information systems and legal information research [9].

## Research Gap

With the current developments in Information Technology and Data Science, people have begun the search for solutions to problems with the usage of modern technology. This has become the case for the legal industry as well. The increase in the volume of information in the legal industry has become so prominent that various individuals have come up with various systems to mitigate the issues that have arisen.

There have been developments in the legal information retrieval domain, and it is evident that a significant amount of research has been done in the field of legal information retrieval around the world.

Legal information and documents of each country around the world have their unique syntaxes and semantics. The concepts, terminology, and structure of legal information change from country to country due to the differences in cultural, traditional, and geopolitical, factors of the respective country and its legal system. Thus, the main research gap that could be identified is the absence of a legal decision support system that would cater to the Sri Lankan legal context. According to the literature survey performed it could be identified that systems built would mainly focus on the legal systems of that respective country.

## The problems of geographically specific systems

Legal decision support systems and legal advisory systems built with a focus on a particular country's legal system do not suit another country's legal system. This could be understood by the points illustrated above.

* **Research A** - This research is focused on Vietnamese law. The research has been conducted by the authors Ngo Xuan Bach, Le Thi Ngoc Cham, Tran Ha Ngoc Thien, and Tu Minh Phuong. The issue with this is, that the system has been made highly specific to Vietnamese law. The data collected represents the Vietnamese transport law and the system’s training, implementation, and testing are extremely focused on the Vietnamese context [10]. Thus, a gap was identified between this research and the problem and solution that the developed system has addressed.
* **Research B** - Like the previous research, this study is done by the authors: S. Pudaruth, R. P. Gunputh, K. M. S. Soyjaudah, and P. Domun. The research is a question-answer system for the Mauritian Judiciary. This focuses on the judgments given by the Supreme Court of the Republic of Mauritius [11]. Even though the above has similarities to the proposed system these two have drastic differences due to the geographic focus put on the research. The research methodology cannot be translated to the Sri Lankan context as it contradicts the entire Lankan legal framework. The Mauritian legal framework is a mix of the French Code Napoleon and British law, whereas the Sri Lankan system comprises the Roman-Dutch law and the British Common Law. Thus, the legal concepts, terminology, and structure greatly contradict each other.
* **Research C** - This is research done toward the building of an intelligent legal advisory system. The authors Christoph Hoppe, David Pelkmann, Nico Migenda, Daniel Hötte, and Wolfram Schenck have focused on the legal documents of the German legal domain through this research. The research has been conducted focusing on the German legal system and the entire research spanning from data gathering to solution building is solely intended for German legal knowledge [1]. Thus, a similar pattern of issues occurs in the translation of the system to the Sri

Lankan context. Germany’s Roman Law based system contradicts the Sri Lankan legal context.

This gives us a clear understanding of the current systems available, and the problems and gaps present in them. Legal decision support systems should be tailor-made to the country’s legal system as there are no universal legal systems anywhere in the world. Thus, a research gap was identified, with the need for a Legal Decision Support System unique to the Sri Lankan legal domain.

## Problems in low accuracy similarity calculations

In the information retrieval (IR) process the similarities must be calculated for the extraction of the most relevant information. Therefore, the accuracy of the similarity calculation methodology plays an integral part in the efficacy of the proposed system. With the literature survey, it was observed that similar systems to the developed system have used low accuracy and older techniques for the calculation of similarity.

* **Research D** – The authors Huu-Thanh Duong and Bao-Quoc Ho proposed a QnA system for Vietnamese legal documents. They have selected and extracted the answer from the relevant documents using a similarity calculation. They have managed to achieve a precision of nearly 70% in the experiment. In this system, however, the answer selection method is based on calculating similarity scores with tf-idf [12]. As a result, it cannot recognize the contextual relationship between words. Thus, it opened a gap in the research context where the developed system could improve the accuracy of the models by using more advanced and improved methodologies.

The accuracy of the legal decision support systems is highly critical as the entire domain is dependent on the accuracy of the information it consumes. Thus, a research gap was identified and filled through the developed system by improving the accuracy of the retrieved documents and information. The developed system addresses this with the usage of transformer-based language models which have revolutionized the field at present [13].

## The problem of low performance in pre-trained language models

The newest methodology present in the domain of Natural Language Processing is the use of pre-trained language models. These are transformer-based approaches. With further research into this domain, it was understood that these technologies could be utilized in the developed system which could improve its efficacy. These models are pre-trained upon a large general dataset and thus come with their own problems.

* **Research E** – Bidirectional Encoder Representations from Transformers (BERT) is such a research product. According to the authors Kenton, Jacob Devlin Ming- Wei Chang, and Lee Kristina Toutanova of the Google AI Language. BERT is a transformer-based language model. It is a pre-trained model trained upon a generalized set of data [14]. The model is widely used in decision support systems. Even though the system has a similar use case it cannot be utilized on a Sri Lanka legal knowledge base due to the generalized nature of the model. Thus, it opened up an opportunity gap where the improvement of these transformer-based models could be performed which would in turn increase their efficacy in the Sri Lankan legal context.

It could be concluded that the transformer-based models, even with their sophisticated nature have produced a research gap that could be solved through the developed system.

In conclusion, it could clearly be understood that several research gaps were essentially addressed by the developed system. The developed system was planned and constructed to provide solutions to these research gaps.

*Table 1 Research gap illustration*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Research A** | **Research B** | **Research C** | **Research D** | **Research E** | **Legal decision support system** |
| **Supports Sri Lankan legal framework** |  |  |  |  |  | Tick mark - Free signs icons |
| **High**  **Accuracy** | Tick mark - Free signs icons |  | Tick mark - Free signs icons |  | Tick mark - Free signs icons | Tick mark - Free signs icons |
| **High**  **performance** |  |  | Tick mark - Free signs icons |  |  | Tick mark - Free signs icons |

## Existing Similar Systems and Feature Comparison

During the planning phase of the project, market research was carried out to discover the availability of similar systems in the country. The evidence of the research proved that only two systems are present in the market that could be similar to the developed system.

* **System A** – This could be identified as the main system that is present in the market and is used by a majority of legal professionals. The system acts as a glossary of legal documents maintained by the Ministry of Justice and contains a minimal number of features. Currently, it only contains a faulty search capability to retrieve documents and it lacks the features that are introduced by the developed system. After several testing usages of the system, it was evident that this system does not have the technical nor the feature-wise capabilities to compete with the developed system.
* **System B** – This is a more technologically advanced system available in the market. Yet, the system does not have the user volume of the first system due to it being a monetized system. This system, like the first, is another glorified legal document glossary, the premium is charged due to the availability of consolidated legal acts, which is not the domain that the developed system is focused on. Thus, it could be concluded that this also lacks the ability to become a competitor of the proposed system due to the technology and availability of fewer features.

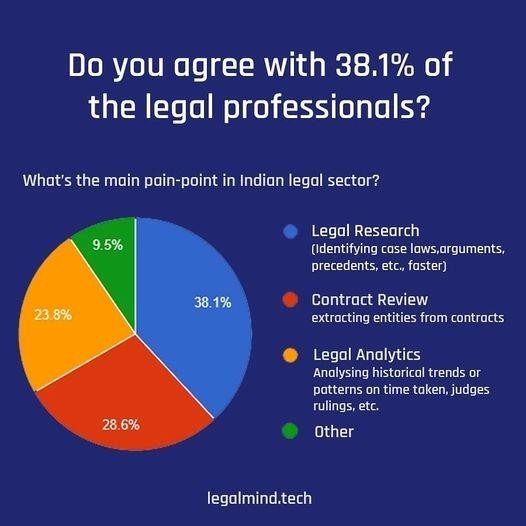
*Table 2 Existing systems and feature comparison*

|  |  |  |  |
| --- | --- | --- | --- |
|  | **System A** | **System B** | **AYCA** |
| **Relevant document retrieval** |  |  | Tick mark - Free signs icons |
| **Legal decision support** |  |  | Tick mark - Free signs icons |

# RESEARCH PROBLEM

The legal domain is a research-oriented industry where each decision is highly critical and highly valuable in time and monetary aspects [11]. The legal industry is dependent on a vast plethora of information.

This large volume of legal information has made this domain impenetrable to the general public and has disengaged the public from having a deeper understanding of the law. Apart from the public, even the personnel who are working and are associated with the legal industry face difficulties in the processing and usage of this huge wealth of information.

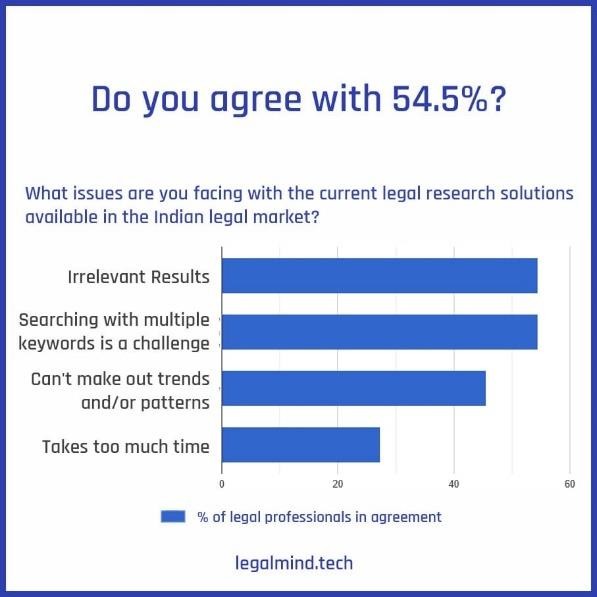


*Figure 2 Pain points in the legal industry*

Source: legalmind.tech

As a result of unmanageable volumes of legal knowledge, the legal industry has become very expert-dependent. This is mainly due to the inability to do accurate and precise legal research without having a considerable understanding of the legal concepts. This high expert dependency raises another major problem which is greatly detrimental to research, a decrease in accuracy. Due to the aforementioned factors legal research has become highly susceptible to human errors and other inaccuracies.

With inaccuracies, large document corpora, and the need for expert knowledge, legal research, and decision-making require a considerable amount of time. This excessive consumption of time has currently become one of the leading problems in the domain.



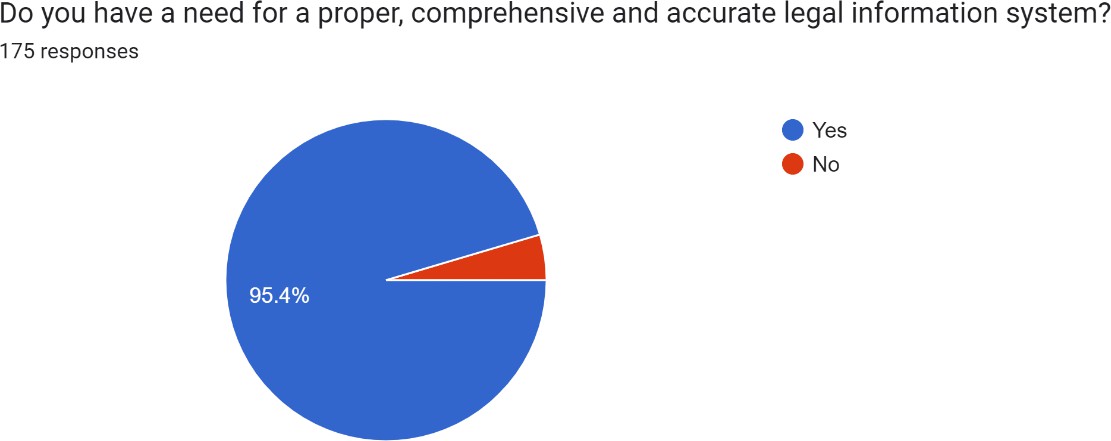
*Figure 3 Issues faced during legal research*

Source: legalmind.tech

These problems have translated to the Sri Lankan legal context too. As the Sri Lankan legal system’s digitization has been done in the recent past, the country’s legal system is in dire need of a proper system for the maintenance of its legal information and assist in legal research and knowledge gathering. This unavailability of a proper legal decision support system has even resulted in various socio-cultural issues in the country as a large population of the country does not have the proper knowledge of understanding legal concepts and does not even have the basic knowledge about their legal rights [2].

Legal professionals face various challenges due to poor access to legal information and knowledge as the country mainly depends on manual legal information retrieval systems. Wrongful conviction and improper justice could be the dangers of this issue. The unavailability of proper legal information retrieval is a hindrance to the development of future legal professionals who are the current law students of the country [8], this could bring about dire consequences for the country and the safety of its constitution.

According to the research survey (refer to [APPENDIX A](#_bookmark85)) conducted with more than 170 participants, it was unveiled how dire the problem is. A majority (95.4%) of the participants agreed that there is no such system available in the country that could enable informed and accurate legal decision-making.



*Figure 4 Research problem identification survey*

It is evident that there is a clear need for a system that could address the aforementioned problems available in the country.

Furthermore, the unavailability of a proper legal decision support system creates inequality in society as there would be oppression of people due to their ignorance of even the simplest rights. It could also give rise to problems in the rule of law as the personnel who are tasked with the protection of the Constitution are unable to make informed decisions.

In conclusion, the Sri Lankan legal industry requires a proper system that could enable and enforce proper legal decision-making through accurate and precise legal question answering. This system would bring about a new era of legal democracy where each citizen would have access to accurate legal information without any discrimination.

# OBJECTIVES

## Main Objective

The primary objective of this research component is to develop a sophisticated legal decision support system tailored specifically to Sri Lankan laws and regulations. At its core, this system revolves around a cutting-edge legal question-answer mechanism. It aims to empower users, including legal professionals, law students, and anyone seeking legal insights, by efficiently and accurately extracting answers to their legal queries from a vast corpus of documents.

The research project embraces advanced technology and adopts a transformer-based approach to fuel the question-answering engine. This approach leverages state-of-the-art natural language processing and machine learning techniques to comprehend the intricate and context-rich nature of legal language. By harnessing the power of transformers [15], the system can decipher complex legal jargon, identify pertinent details, and deliver precise answers that align with the user's queries.

This research serves a dual purpose. Firstly, it aims to bridge the gap between the intricate realm of Sri Lankan legal documents and the accessibility of this knowledge to individuals who rely on accurate legal information. Secondly, it aspires to streamline the process of legal research and decision-making, offering users a time-efficient and reliable means of obtaining answers to their legal questions.

In summary, the system strives to revolutionize the way legal information is accessed and understood in Sri Lanka. By integrating advanced technology, specifically the transformer-based question-answering engine, it seeks to empower individuals and professionals alike, making the legal landscape more accessible and comprehensible to all who seek its insights. The project's overarching objective is to empower individuals and professionals alike, ultimately facilitating a more accessible, comprehensible, and informed legal environment in Sri Lanka.

## Sub Objectives

## Data extraction, pre-processing, and structuring

This specific sub-objective of data extraction, pre-processing, and structuring is a pivotal component within the overarching goal of developing a functional legal question-answer system. This objective entails the systematic collection of legal documents from trusted sources, with a focus on data accuracy and relevance.

Data pre-processing follows to ensure consistency and uniformity within the documents. This involves text normalization and formatting corrections. Subsequently, the data is structured into the SQuAD [16] format, transforming lengthy legal texts into a structured question-and-answer framework that is machine-readable [17]. This structured dataset is then utilized for model training and evaluation.

This systematic approach ensures the accuracy and effectiveness of the legal decision support system by establishing a well-structured and standardized dataset that serves as the basis for subsequent phases in the system's development and deployment.

## Question answering model building

The sub-objective of "Question answering model building" within the development of the Legal question-answer system is a pivotal step that intricately contributes to the system's main goal. This objective involves a multi-faceted approach, carefully designed to deliver accurate and contextually relevant answers to user queries based on the extensive corpus of legal documents.

The first component of this is the implementation of a retriever model. This model serves as the initial gatekeeper, comparing user-entered queries with the comprehensive document corpus to identify the most pertinent legal documents. By leveraging advanced algorithms and natural language processing techniques, the retriever model effectively narrows down the search to retrieve documents with the highest potential to answer the user's question.

The second key element of this is the reader model, which represents the heart of the question-answering process. This model is built upon a transformer-based architecture, fine-tuned with the prepared SQuAD format dataset [17] [16]. It is equipped with the ability to analyze the retrieved legal documents and extract precise answers to user queries. This transformer-based reader model is highly adept at comprehending complex legal language, ensuring that the answers it generates are contextually accurate.

This specific objective bridges the gap between user questions and the wealth of legal knowledge contained within the corpus of documents, ultimately achieving the main objective of providing a reliable and accessible legal decision support system.

## Pipeline building and integration

The pipeline building and integration within the development of the Legal question- answer system is a critical step in realizing the main objective of creating a comprehensive and accessible legal decision support system.

The first essential element of this is the integration of the trained retriever and reader models. These models, which have been fine-tuned and optimized to efficiently retrieve and extract legal information, are integrated to form a coherent question-answering pipeline. This integrated pipeline is the backbone of the system, as it orchestrates the flow of information and processes user queries to provide accurate legal answers.

The next crucial phase of this objective involves connecting the question-answering pipeline to the front end of the web application. The UI development and integration process must prioritize user-friendliness, ensuring that individuals, regardless of their technical expertise, can easily navigate the system, input queries, and receive answers clearly and understandably.

The successful execution of this objective ensures that the Legal question-answer system is not only functionally robust but also user-centric. It transforms the system into a practical and accessible tool that can leverage the wealth of legal knowledge contained within the corpus of documents.

# METHODOLOGY

This segment provides a comprehensive insight into the systematic approach employed throughout the research journey, encompassing all the stages from initial requirements gathering to the successful implementation, and completion of the project. This section meticulously details the methodologies, strategies, and procedures followed to accomplish the specified objectives and sub-objectives.

The research methodology adopted in this study adheres to a structured and phased approach, ensuring the robustness and reliability of the processes involved. It begins with a thorough exploration of the requirements, understanding the nuanced needs and challenges associated with the development of the Legal question-answer system. Subsequently, it progresses to data acquisition, system development, integration and implementation of the system, and all the remaining steps required in the completion of the project. Furthermore, this section also addresses the critical phase of system integration with the user interface of the web application, emphasizing the significance of user-centric design and accessibility.

In essence, the Methodology section serves as the roadmap that guided the research endeavor from its inception to its successful completion. It offers a detailed account of the systematic processes, methodologies, and strategies that were meticulously executed to achieve the defined objectives and sub-objectives. Through this section, a comprehensive understanding of the structured approach that underpins the development of the Legal question-answer system and the meticulous research journey undertaken to bring this project to fruition is gained.

## Requirement Gathering and Analysis

In the initial phase of system development, requirements gathering, and analysis played a pivotal role in shaping the project's direction. This phase revolved around identifying both functional and non-functional requirements to ensure the final product aligned seamlessly

with stakeholder needs while adhering to constraints such as time, budget, and technical feasibility.

The outcome of this phase was the creation of a comprehensive Software Requirements Specification (SRS) document, meticulously detailing the identified functional and non- functional requirements. This SRS document served as an invaluable roadmap for the development team, guiding them in building the system following the precise specifications outlined by stakeholders.

The project team collaborated closely with all stakeholders (refer to the [APPENDIX C](#_bookmark87) section for further details) to collect requirements using various methodologies. These methods included:

* **Interviews and Meetings:** Multiple interviews and meetings were conducted with key user groups, including lawyers and law students, to elucidate their requirements and expectations regarding the system. The focus was on understanding the challenges they faced and devising solutions to address and rectify these issues.
* **Survey Questionnaire:** A comprehensive survey questionnaire was meticulously crafted and deployed to gather insights from target groups. This survey acted as a robust tool for conducting market research, enabling the identification of market gaps, stakeholder problems, and system requirements. (Refer to [APPENDIX A](#_bookmark85) for the Survey Results, which have been analyzed and integrated into the system.)
* **Observation:** The project team conducted on-site observations within legal settings to gain firsthand insights into the workflows and requirements of legal personnel. This firsthand experience provided a profound understanding of the practical implications of the system's implementation and underscored the system's potential impact.

This comprehensive requirements-gathering and analysis phase laid the foundation for the subsequent phases of system development. It ensured that the system was not only aligned with stakeholder needs but also helped in addressing real-world challenges effectively.

## Functional requirements

* **The system must provide concise answers to the user queries:** The system should have the ability to receive user legal queries in human-understandable language, process the user’s queries, and generate clear, concise, and accurate answers to the user questions using the legal knowledge base.
* **The system should be able to retrieve similar documents to user queries:** The system should have the ability to access a legal document database, match similar relationships containing documents with similar legal concepts, and retrieve these similar documents promptly.

## Non-functional requirements

Non-functional requirements are features that indicate how the system should behave rather than what it ought to accomplish. Some instances of non-functional requirements for the system include:

* **Performance:** Even under conditions of high load, the system should be responsive and produce results promptly. Moreover, numerous concurrent user sessions should be supported by the system without any performance deterioration.
* **Reliability:** There should be very little system downtime and interruption of user access, and the system should be very dependable and available. The system must be able to quickly recover from faults and failures.
* **User-friendliness:** The system interface should be simple and easy to use. The system should be accessible and understandable by any user without discrimination about the depth of their legal knowledge.
* **Scalability:** The system should be easily scalable. The system should possess the ability to scale up or down depending on user demand. The system should accommodate future growth and should be able to grow without significant changes to the system architecture.
* **Maintainability:** There should be little user disruption during system upgrades and maintenance.
* **Cross-browser compatibility:** Ensure that the system operates seamlessly and correctly across different web browsers (e.g., Chrome, Firefox, Safari). Users should have consistent access and functionality, regardless of their choice of browser.
* **Resource Utilization:** Efficiently utilize system resources (CPU, memory, storage) to ensure optimal performance while minimizing resource consumption. The system should strike a balance between delivering fast response times while minimizing resource consumption to ensure it runs smoothly even during periods of high demand.
* **Backup and Recovery:** Develop robust backup and recovery procedures to safeguard data integrity and system availability in case of data loss or system failures. These procedures guarantee that the system can swiftly recover and continue serving users without significant disruptions or data compromises.

## Feasibility Study

A feasibility study was conducted to determine whether the system was viable and worthwhile for development. The study examined three key aspects:

* **Scheduling feasibility**: This aspect assessed the time required to create the system. The development timetable needed to be reasonable and achievable within the allocated timeframe. The proposed system's development deadline coincided with the research project's deadline, with the requirement that the final product of the system be produced by the end of September 2023.
* **Technological feasibility**: The focus here was on determining the technological foundations necessary for the system's development. Team members were required to have extensive knowledge of Python and similar development languages, as well as expertise in various development technologies. Proficiency in machine learning algorithms, natural language processing algorithms, and data extraction

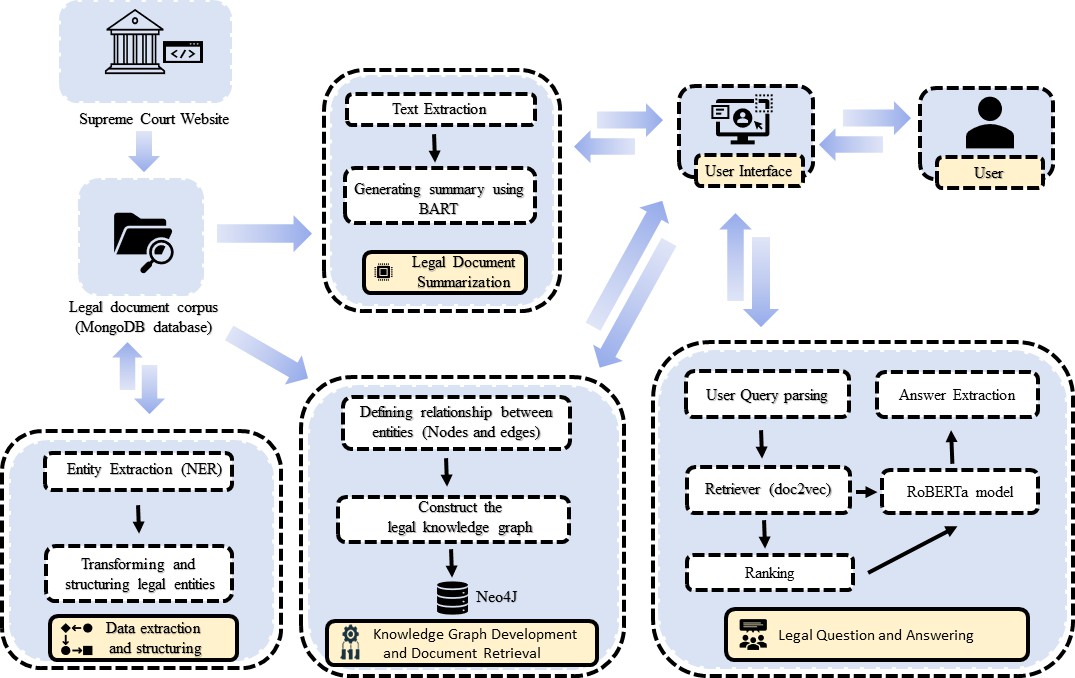
and processing were essential. Additionally, deep knowledge of software development methodologies and related tools and technologies was a prerequisite.

* **Economic feasibility**: This aspect analyzed the financial viability of the project. The development cost of the system was compared with the potential benefits it could bring. The return on investment (RoI) needed to be reasonable. The developed system includes a commercialization plan and market acceptance, as determined by prior market research. The research indicated a substantial market gap and a strong demand for such a system, suggesting a promising market attraction and return. The development of the system was carried out with minimal expenditure to maximize RoI. The initial phase of development was funded by the internal team.

In hindsight, this feasibility study provided valuable insights and guidance for the development of the system, ensuring that it proceeded on a viable and well-considered path.

## System Designs

## Overall system architecture diagram



*Figure 5 Overall System Architecture*

The Sri Lankan Legal Information Retrieval System is a sophisticated and comprehensive platform designed to streamline the process of accessing, analyzing, and retrieving legal information in Sri Lanka. This system comprises four main components:

1. Data Extraction and Structuring Module.
2. Knowledge Graph Development Module.
3. Legal Document Summarization Module.
4. Legal Question Answering Module.

Each of these components plays a vital role in enhancing the accessibility and usability of legal information for various stakeholders.

## Data Extraction and Structuring module

The Data Extraction and Structuring Module is the fundamental component of the Sri Lankan Legal Information Retrieval System. It serves as the system's backbone, initiating the process of gathering, organizing, and structuring legal case files from various sources, including the Supreme Court website and the Ministry of Justice document corpus.

This module utilizes a specially trained Named Entity Recognition (NER) model, tailored for the Sri Lankan legal context. The NER model plays a critical role in breaking down complex, unstructured legal documents, identifying and extracting key pieces of information such as the names of judges, judgments, plaintiffs, defendants, legal terms, dates, and case numbers [18].

The extracted information is then stored in a structured template within a MongoDB database. MongoDB's flexibility and scalability make it a suitable choice for this purpose. By structuring the information in this way, it ensures consistency and accessibility for the rest of the system's components.

In essence, the Data Extraction and Structuring Module serves as the foundational layer that enables the subsequent modules of the system, including the Knowledge Graph Development Module and Legal Document Summarization Module, to function effectively and provide users with valuable legal information retrieval capabilities.

## Knowledge Graph Development module

The Knowledge Graph Creation Module enhances the system's capabilities by leveraging the structured data stored in the MongoDB database. It takes this structured information and transforms it into a Legal Knowledge Graph (LKG) [19], a network of interconnected legal entities and relationships.

In this process, the module maps various legal entities, such as judges, plaintiffs, defendants, legal terms, dates, and case numbers, and establishes relationships between them to reflect the complex connections within the legal domain [20]. These relationships

represent important associations, such as which judge presided over a case or which plaintiff was a part of a particular case.

The LKG, containing this web of legal entities and relationships, is stored in a specialized database called Neo4J. Neo4J is designed to efficiently manage and query graph structures, making it an ideal choice for storing and accessing the LKG [21].

Once the LKG is in place, the module uses it for the retrieval of relevant legal documents based on user search keywords. By tapping into the interconnected entities and relationships within the LKG, users can access legal information with great precision, gaining valuable insights into the complexities of the legal domain.

## Legal Document Summarization module

The Legal Document Summarization Module capitalizes on the capabilities of the Data Extraction Module as well. This module's workflow involves taking legal case files as inputs, which are initially processed through the Data Extraction Module to identify key entities. Specifically, it focuses on extracting the extensive judgments within these documents [22]. Once this entity breakdown is complete, the component employs an abstractive summarization method to condense the case judgments into concise summaries, which are then presented to the user [23].

The summarization process is greatly aided by the utilization of a fine-tuned BART model. This model is specialized for the task of summarizing legal judgments effectively, ensuring that the extracted summaries are coherent, informative, and reflect the essential aspects of the original documents.

In summary, the Legal Document Summarization Module is an integral part of the system that streamlines the process of summarizing extensive legal case judgments [24]. It does so by first leveraging the entity extraction capabilities of the Data Extraction Module and then employing a fine-tuned BART model for the summarization task. The result is user- friendly access to the most critical information within complex legal documents, saving users valuable time and effort in their legal research.

## Legal Question Answering module

The Legal Question Answering Module is focused on responding to user queries using an extractive answer generation approach. This technique relies on two critical components within the system: a retriever model and a reader model.

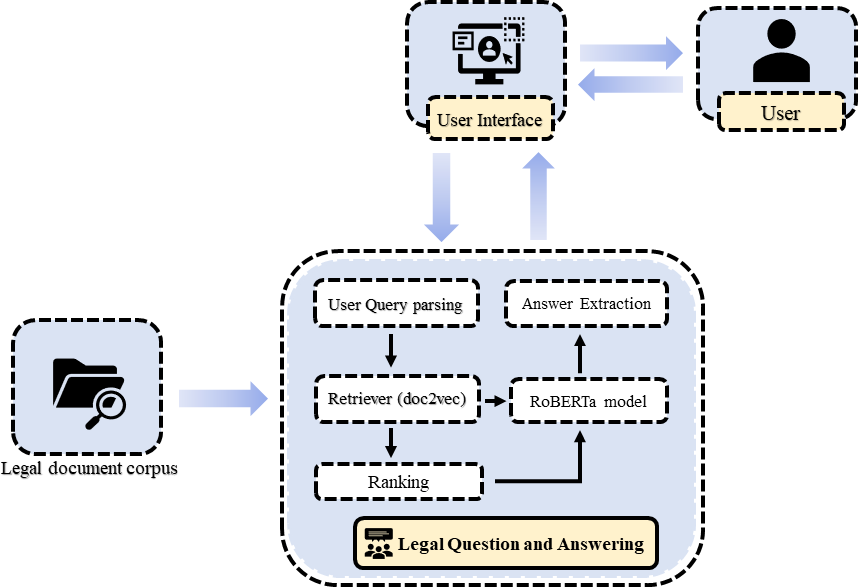
The retriever model plays the initial role of identifying and retrieving relevant legal documents from the system's database that are most suitable for addressing the user's query [25]. It uses a combination of methods such as keyword matching and semantic similarity to pinpoint these documents.

Once the retriever model selects the most relevant legal document, it becomes the context for answering the user's question. The reader model, powered by a fine-tuned transformer- based model called RoBERTa [26], takes over. The reader model's task is to extract the most appropriate answer directly from the context document, identifying the relevant section or passage that directly addresses the user's query.

In summary, the Legal Question Answering Module combines the functionalities of the retriever and reader models to provide accurate and context-aware answers to user queries. Through these advanced natural language processing techniques, the system streamlines the process of legal research, enabling users to access specific legal information quickly and effectively.

The system's components are closely interconnected and operate in a seamless flow of execution, as depicted in the figure. This integrated approach ensures that each component relies on the output of the preceding one, contributing to the overall efficiency and effectiveness of the system.

## Component-specific system diagram



*Figure 6 Component-specific system diagram*

As illustrated in the above diagram the development of the Legal Question Answering System component consists of many internal sub-components and steps. Each of the illustrated steps is critical in the Question Answering System build, and its success, accuracy, and consistency. The system focuses on information retrieval and providing accurate and consistent answers to user queries. The system receives user queries written in natural language, these queries would be parsed and processed through the system, and relevant and accurate answers would be produced to satisfy the user query. Given below are summaries of the above-illustrated steps and these would delve into the importance of each step and justify why these features should be included in the system:

## Query parsing

Query parsing is a pivotal component within the Legal question-answering system, responsible for transforming user queries into a machine-readable format. This essential

step ensures that the system can effectively understand and process the questions posed by users, facilitating accurate and relevant responses.

The process begins with the receipt of user queries as inputs. These queries, often expressed in natural language, need to be converted into a structured format that the system's algorithms can comprehend. Query parsing encompasses several key techniques:

* + - * + **Tokenization**: Tokenization involves breaking down the user's query into individual words or tokens. Each token represents a distinct unit of meaning within the query. This step enables the system to analyze the query at a granular level, making it easier to identify keywords and extract relevant information [27].
        + **Stop word removal**: Stop words are common words in a language (e.g., “and”, "the", and "in") that do not carry significant meaning and are often removed during query parsing. Eliminating stop words helps reduce noise in the query and focuses on the essential keywords that convey the user's intent.
        + **Normalization**: Normalization techniques are applied to standardize the query's format and structure. This includes converting all letters to lowercase to ensure case insensitivity in the search process. Normalization enhances consistency in query processing.

By parsing user queries using these techniques, the Legal question-answering system prepares the input data for subsequent stages, such as information retrieval and question- answering.

## Retrieval of the highest-ranking document

The retrieval of the highest-ranking documents for the Legal Question-Answer system is a crucial step that determines the relevance and precision of the information provided in response to user queries. This process involves leveraging a Doc2Vec model to effectively match user queries with the most suitable legal documents within the corpus.

## Inside a Doc2Vec model

* + - * + **Document embeddings:** The Doc2Vec model begins by generating embeddings for each document [28] in the legal corpus. These embeddings are numerical representations that capture the semantic meaning and context of the text within each document. Unlike traditional bag-of-words approaches, Doc2Vec considers the order of words and their relationships in the document, allowing for more nuanced document representations.
        + **Incorporating context:** Doc2Vec considers the context of words in a document by creating document vectors that are influenced by the words and phrases found within. This means that documents with similar content or context will have similar embeddings, making it easier to identify related documents.
        + **Query parsing:** When a user query is parsed and transformed into a machine- readable format, it is also represented as an embedding in the same vector space as the documents. This allows the query to be compared to the document embeddings to identify similarities.

## The retrieval process:

* + - * + **Matching queries to documents:** The parsed user query, represented as an embedding, is compared to the document embeddings using similarity measures such as cosine similarity. This quantifies the similarity between the query and each document in the corpus.
        + **Ranking documents:** Documents are ranked based on their similarity scores with the query [25]. Documents that are more closely related to the user's query receive higher scores and are considered the most relevant for answering the question.
        + **Retrieval of highest-ranking document:** The retriever model retrieves the highest-ranking document, which is deemed most likely to contain the information necessary to answer the user's query [26]. This document is then fed into the subsequent reader model for detailed analysis.
        + By employing a Doc2Vec model in the retrieval process, the system can effectively match user queries with the most contextually relevant legal documents within the corpus.

## Answer extraction with RoBERTa

Answer extraction is a critical phase in the Legal Question-Answer system, responsible for retrieving accurate and contextually relevant answers to user queries. In this process, a fine-tuned RoBERTa (a transformer-based model) [29] plays a central role in extracting answers from the highest-ranking legal document retrieved during the previous stage.

## The answer extraction process:

* + - * + **Input configuration:** In the answer extraction phase, the parsed user query is presented as the question, and the retrieved highest-ranking legal document is passed as the context. The model is then configured to process these inputs for question answering.
        + **Contextual understanding:** RoBERTa's bidirectional attention mechanisms [30] enable it to comprehend the context provided by the legal document. It identifies the relationships between different parts of the text, including how the query relates to the content of the document.
        + **Answer span prediction:** RoBERTa predicts the answer span within the document by identifying the start and end positions of the answer. It does this by assigning probabilities to each word in the document to determine if it marks the beginning or end of the answer span [31].
        + **N best answers:** The model generates a list of potential answer spans and their associated probabilities. To ensure accuracy, multiple candidates are considered. The N best answers are extracted, and each is assigned a likelihood score.
        + **Final answer selection:** Finally, the answer with the highest probability score is selected as the most likely response to the user query. This answer is then presented to the user as the system's response.

By incorporating a fine-tuned RoBERTa model, the Legal Question-Answer system effectively extracts precise answers from the context provided by the highest-ranking legal document.

The Legal Question-Answer system is a powerful, complex, and sophisticated tool that was carefully modeled, developed, and implemented to ensure its accuracy and efficiency. The system comprises the above steps which are highly interdependent. The completion of the system involved the precise integration of each of the above features.

The system mainly utilizes Natural Language Processing algorithms as the system handles user queries as well as legal documents that are present in human language. Natural language generation and document retrieval also play a pivotal part in the outcome generation of the system [32].

Overall, the final system can take user queries as inputs which are present in natural language. Then the input queries would be parsed and processed via the internal models of the system which have been elaborated above and the relevant answers would be extracted and presented to the user. This presented information would be the most likely result expected by the user.

## Methodology

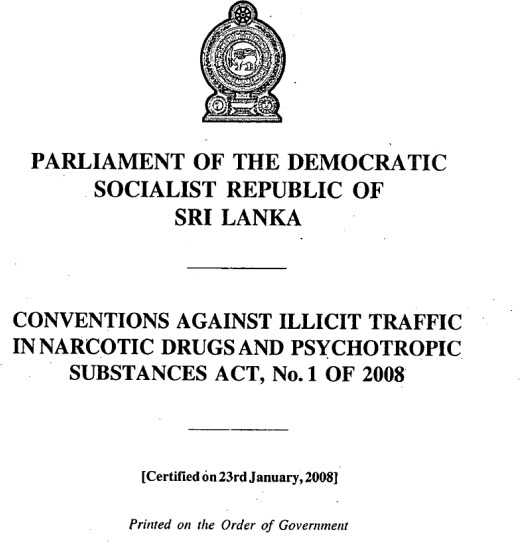
## Data acquisition and preparation

Through this section, the process of data acquisition and preparation for the Legal Question-Answer System is described in detail. This phase involves several critical steps and considerations, each contributing to the development of a robust and comprehensive dataset tailored to the specific scope of the project.

## Scope definition and data source:

The research project's scope was meticulously refined to focus exclusively on the "Conventions Against Illicit Traffic in Narcotic Drugs and Psychotropic Substances Act, No. 1 Of 2008" (Shown in Figure 7). This narrowing down of scope was a critical step in ensuring the project's precision and relevance to the legal domain. The following sources were identified for the extraction of information:

* Ministry of Justice ([www.lawnet.gov.lk](http://www.lawnet.gov.lk/))
* Court of Appeal of Sri Lanka ([courtofappeal.lk](http://courtofappeal.lk/))
* Supreme Court of Sri Lanka ([supremecourt.lk](http://www.supremecourt.lk/))
* Department of Government Printing ([documents.gov.lk](http://documents.gov.lk/))



*Figure 7 Target data source – Narcotic Act, No 1 of 2008*

One key facet that facilitated this data acquisition was the legal provisions of the Right to Information Act, No. 12 of 2016 [33]. Under this legislation, the information about the "Conventions Against Illicit Traffic in Narcotic Drugs and Psychotropic Substances Act, No. 1 Of 2008" was made available to the public as open-source information. This ensured that the project's data collection process adhered to legal and ethical standards, promoting transparency and accessibility of legal information.

## Manual annotation of data:

To ensure the quality and accuracy of the dataset, a meticulous manual annotation process was employed. An annotation tool, as illustrated in the below figure, was utilized. During this phase, domain specialists were actively involved in reviewing the data, legal queries were matched manually with corresponding answers found within the legal documents. This process aimed to capture precise legal context and insights that are vital for the success of the system.



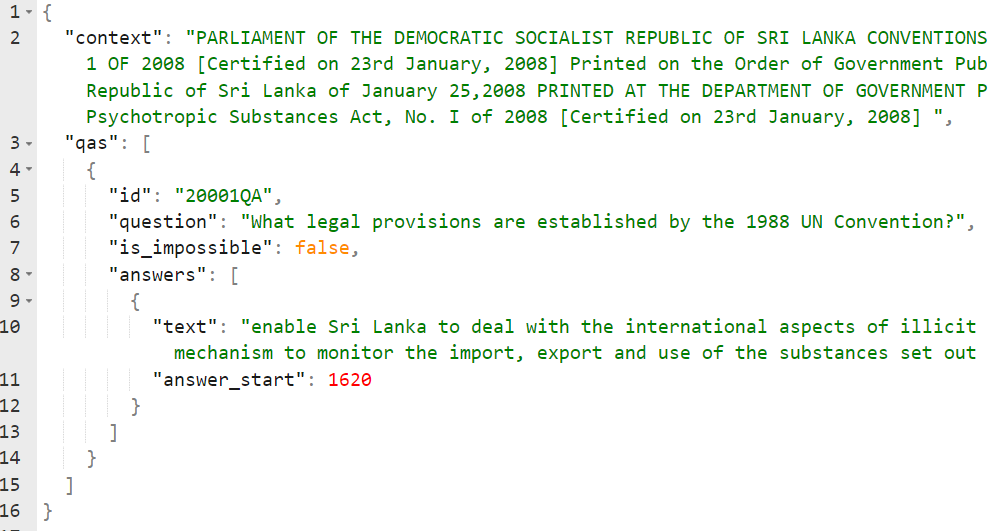
*Figure 8 Manual annotation of documents*

## Automation of SQuAD format conversion:

The outcome of the manual annotation process was a set of CSV (Comma-Separated Values) files containing essential data elements. These files included the legal context, user-generated queries, answer spans, and answer starting indices. This format needed to be converted into the SQuAD (Stanford Question Answering Dataset) format [29], which is commonly utilized for training and evaluating question-answering models.

To facilitate the transition from CSV to SQuAD format, a Python script was developed. This script automated the conversion process, ensuring that the dataset was structured

appropriately for the question-answering models. The SQuAD format, as shown in the below figure, organizes data, making it compatible with machine learning models designed for question answering, and the automation script increases the scalability, consistency, efficiency, and reproducibility of the dataset.



## Dataset review and approval:

*Figure 9 SQuAD format*

Before finalizing the dataset, a comprehensive review process was conducted. The dataset, comprising 14,361 unique data points specific to Sri Lankan law, underwent thorough scrutiny and validation. An industry professional with expertise in the legal domain reviewed and approved the dataset, confirming its accuracy, relevance, and suitability for the research project (refer to [APPENDIX D](#_bookmark88)).

In summary, the data acquisition and preparation phase of the research project involved meticulous planning and execution to ensure the quality and integrity of the dataset. This comprehensive dataset, structured in SQuAD format [17], serves as a foundational resource for training and evaluating the Legal Question-Answer System. The attention to detail and domain expertise applied throughout this process are essential for the success of the project and the delivery of accurate legal information to users.

## Model building and training

The backbone of the Legal Question Answering System lies in the development and training of two pivotal components: the retriever model and the reader model. The retriever model serves as the initial gatekeeper, efficiently sifting through vast legal document collections to retrieve contextually relevant documents in response to user queries. On the other hand, the reader model, a fine-tuned transformer-based architecture, specializes in comprehending complex legal language and extracting precise answers from the retrieved documents. Through meticulous training and optimization, they form the foundation of the system, ensuring it can provide users with accurate, context-aware responses to their legal queries.

## Retriever model development

The retriever model in the Legal Question Answering System plays a pivotal role in identifying and retrieving the most suitable document to answer a user's query. For this critical task, the Doc2Vec model has been selected as the preferred approach.

The choice of the Doc2Vec model over traditional TF-IDF (Term Frequency-Inverse Document Frequency) methods [26] for document retrieval in the Legal Question Answering System is driven by several key advantages. Unlike TF-IDF, which relies on simple term statistics, Doc2Vec leverages sophisticated neural network-based techniques, making it more adept at capturing semantic meaning and context within legal texts. Doc2Vec's ability to generate document embeddings [28] (vector representations [1]) based on context and word order enables it to understand complex relationships and nuances in legal documents, which is crucial for precise retrieval.

## Dataset and data preprocessing:

The training dataset consists of 50 legal documents.

As shown in Figure 10 below, the text data will be preprocessed utilizing the preprocessing functions of the Gensim library. Preprocessing involves mainly lowercasing and tokenization of text.

Each document of the dataset would be tagged to create TaggedDocument objects. Where each document in the dataset is associated with a unique tag (usually an integer index)

## Retriever model training:

The choice of hyperparameters for training a Doc2Vec model can significantly impact the quality of the learned document embeddings and the overall performance of the retrieval system.

As shown in Figure 10 below, the following values have been used in the tuning of the hyperparameters: (epochs=50, min\_count=10, window=15, vector\_size=300)

* **epochs** represent the number of iterations over the entire dataset during training.

For tasks involving complex documents like legal texts, training for a reasonable number of epochs (such as 50) allows the model to capture intricate patterns in the data. Legal documents can be lengthy and intricate, and more training may help the model better understand their structure and semantics.

* **min\_count** (Minimum Word Frequency) sets a threshold for word frequency. Words that appear less frequently than this value are ignored during training.

In legal documents, there may be a lot of rare legal terminology, but some of it might be specific to particular documents and not useful for retrieval. Setting “min\_count” to a moderate value like 10 filters out very rare terms while retaining relevant vocabulary.

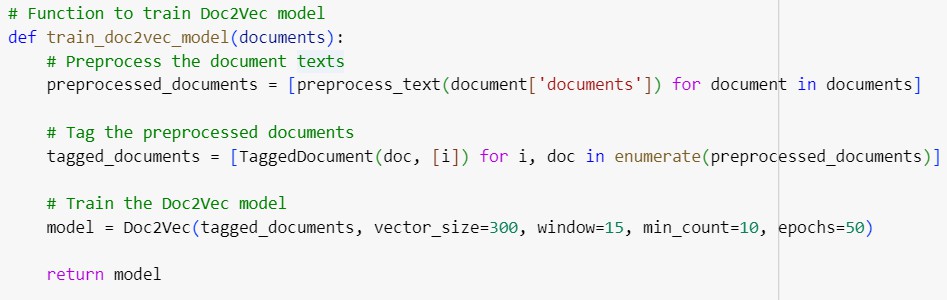
* **window** (Context Window Size) defines the maximum distance between the current word and the predicted word within a sentence. It captures the local context of words.

In legal texts, context is essential for understanding the meaning of legal terms, phrases, or sections. A larger “window”, such as 15, allows the model to consider a broader context, which can be important for understanding complex legal language.

* **vector\_size** (Embedding Dimensionality) specifies the dimensionality of the document embeddings. Larger values can capture more complex relationships in the data.

Legal documents often contain nuanced and multifaceted information. A larger embedding dimension, such as 300, allows the model to represent this complexity more effectively. It can help in distinguishing between similar but contextually different legal terms or concepts.

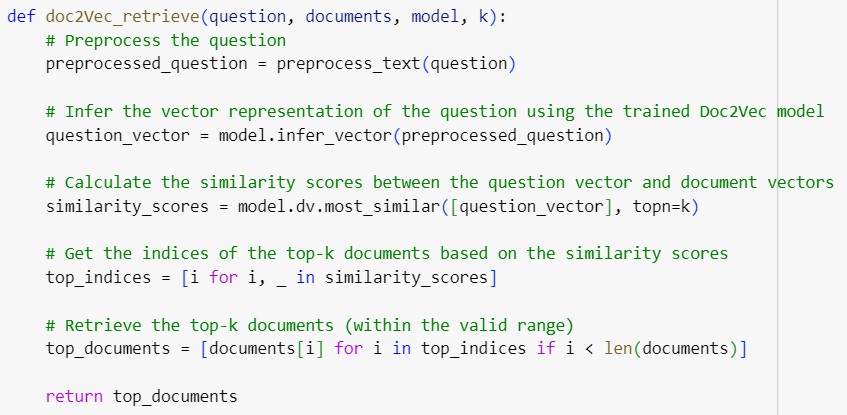
After preprocessing and tagging the documents, the Doc2Vec model has been trained using the provided hyperparameters. The Doc2Vec class from Gensim is used for this purpose.



*Figure 10 Training code segment - Doc2Vec model*

## Document retrieval with trained retriever model:

The trained Doc2Vec model was used in the retrieval of relevant legal documents in response to legal queries. The “doc2Vec\_retrieve” function (shown in the below figure) is designed for this purpose. It takes a query question, the trained model, and the number of documents to retrieve (k) as input and returns the most relevant documents based on similarity.



*Figure 11 Retriever code segment - Doc2Vec model*

## Reader model development

In the Legal Question Answering System, the reader model has the most important function of extracting answers to user queries from the retrieved legal documents. The reader model serves as the system's primary tool for interpreting and responding to user queries.

Transformers have demonstrated state-of-the-art results in various NLP applications, making them an ideal choice for complex legal language comprehension. Thus, RoBERTa, an optimized variant of BERT, Electra, and similar models [29], was selected for the reader model implementation due to its superior accuracy and robustness in understanding contextual nuances within the text.

## Dataset and data preprocessing:

The dataset is a SQuAD format dataset [17] (refer to the [Data acquisition and preparation](#_bookmark37) section) with 14361 data points.

A Python function has been written to preprocess the data, where it takes a dictionary of training examples as input, where each example includes a "question" "context" and "answers". It tokenizes questions and contexts, calculates answer span positions within

tokens, and handles cases where answers are out of context. The resulting data is essential for training the model to predict answer spans.

## Reader model training:

As shown in Figure 12, a pre-trained “RoBERTA-base” model from the “simpletransformers” library [16] has been selected to be finetuned and trained on the prepared dataset. Through meticulous research and testing, hyperparameters have been tailored for training the models while striking a balance between performance and resource consumption (refer to the below figure 13).

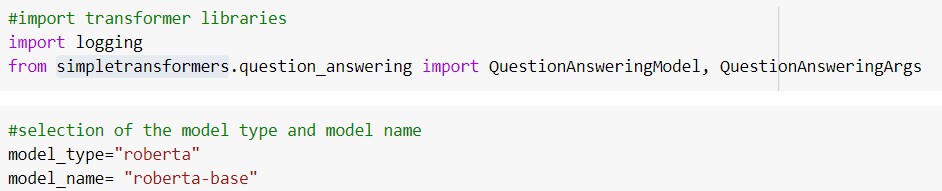
* **max\_seq\_length** (Max Sequence Length): Setting it to 128 may be suitable for legal documents, as it controls the maximum number of tokens in each input sequence. Longer documents may get truncated or split into smaller parts, which can impact the model's ability to understand context. However, increasing it could lead to memory issues.
* **num\_train\_epochs** (Number of Training Epochs): 50 epochs provide ample training time for the model to learn from the legal documents. Legal text often contains complex language and nuances, requiring extended training for the model to capture legal concepts effectively.
* **train\_batch\_size** and **eval\_batch\_size**: A batch size of 128 for training and 64 for evaluation balances memory usage and training speed. Larger batch sizes can accelerate training but may require more GPU memory.
* **evaluate\_during\_training**: Enabling evaluation during training helps monitor the model's performance as it trains. This is essential for tracking progress and early identification of issues, which is crucial when working with complex legal documents.
* **evaluate\_during\_training\_steps**: Evaluating the model every 1000 steps during training provides frequent performance feedback. This helps in detecting improvements or overfitting, especially when training on challenging legal texts.
* **n\_best\_size**: Keeping the top 3 best predictions can help in handling multiple possible answers within legal documents, increasing the chances of providing accurate responses.
* **use\_cuda**: Enabling CUDA (GPU) usage speeds up training, which is beneficial for handling the computational demands of a RoBERTa-based model.

The hyperparameter-tuned model is trained on the dataset for 50 epochs, with the training process continually evaluating its performance. During training, the model learned to predict answer spans within the legal documents given the associated questions.

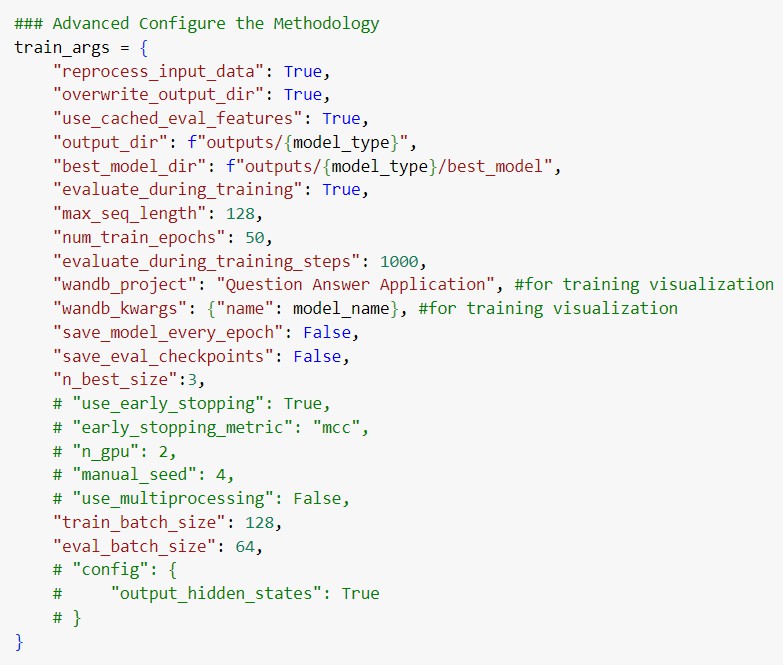
## Answer extraction with trained reader model:

The trained final model has been saved, including the checkpoints. This model was fine- tuned to extract answers from legal documents, where it can take legal questions and context (legal documents) as input and predict answer spans within those documents.

It has leveraged its understanding of legal language, context, and the hyperparameters configured during training to provide accurate responses to legal queries.



*Figure 12 Pretrained RoBERTa model import*



*Figure 13 Hyperparameter tuning - RoBERTa model*

## Implementation of the system

The implementation phase of the Question-Answer System can be divided into two key steps: the construction of the question-answering pipeline and the subsequent integration of this pipeline to facilitate seamless communication between the frontend and backend components.

## Building the question-answering pipeline

The first step in implementation involves the creation of a well-structured question- answering pipeline. This pipeline is designed to efficiently process user queries and extract accurate answers from legal documents.

## Loading of saved models:

Once the retriever and the reader models were trained and optimized, they were saved in a storage location (shown in the below figure) to ensure easy access during system

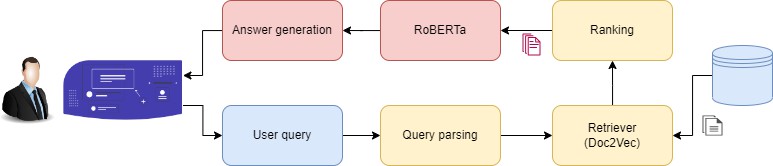
operation. This allows the model to be loaded efficiently whenever a user query is submitted.



## Integration of models:

*Figure 14 Model saving*

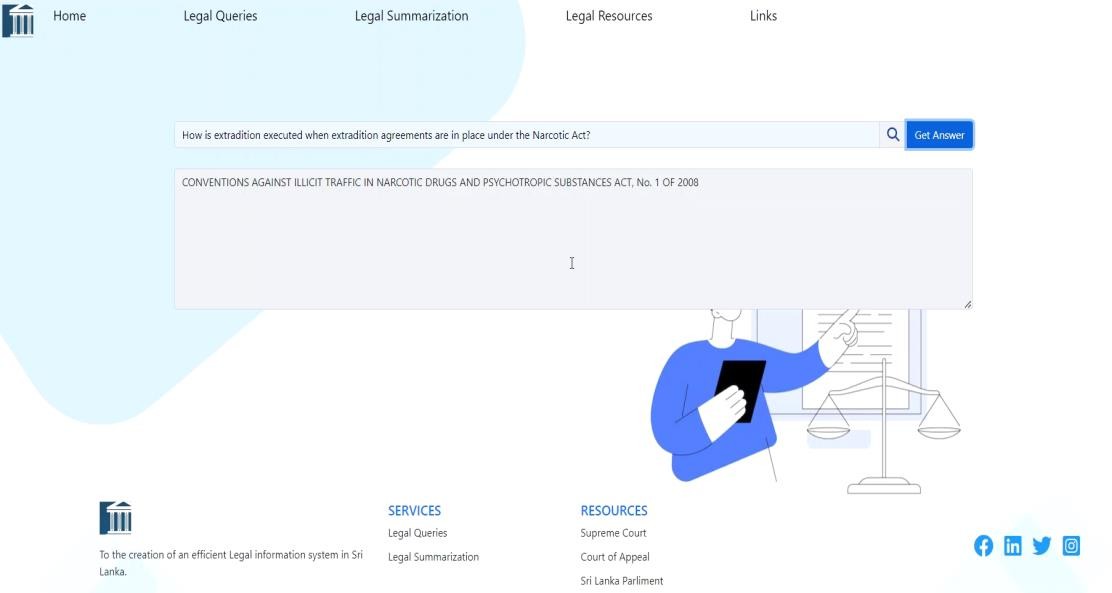
With both the retriever and reader models in place (below figure), the question-answering pipeline is built by integrating these components. The pipeline is designed to seamlessly accept user queries and follow a logical flow: the retriever model identifies the most relevant document, which is then used as context for the reader model to extract the answer.



*Figure 15 QnA pipeline*

## Frontend implementation and integration

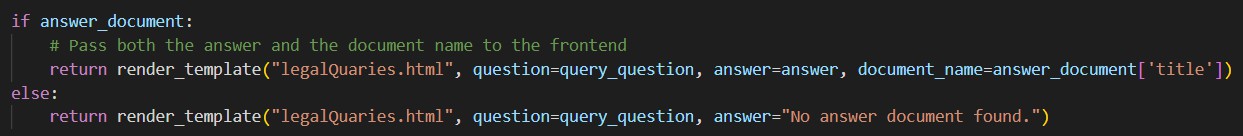
The seamless integration of the Legal Question Answering System with its frontend interface holds paramount importance, especially given the diverse user base, many of whom may not possess extensive technical proficiency. As shown in the figure below the user interface is intuitive, offering a user-friendly experience that empowers users to effortlessly input their queries and access legal information.



*Figure 16 Question answering system UI*

The system has been crafted as a web application, with Python Flask serving as the foundational framework for its development. The user interface features a text field (shown in the above figure) that allows users to enter legal queries.

This text field then channels the user's query into the question-answering pipeline. Within this pipeline, the query undergoes the process mentioned earlier to extract precise answers. Upon successfully extracting the answer, the system uses Flask routing to seamlessly relay the answer back to the user interface (shown in the below figure).



*Figure 17 Answer routing to UI*

## System Testing

The testing phase of the prepared Legal Question Answering System was a critical step in ensuring its reliability, accuracy, and overall performance. Comprehensive testing strategies were employed to evaluate various aspects of the system.

## Unit testing

Individual components of the Legal Question Answering System are tested to verify their correctness.

*Table 3 Unit test case 01 - Retriever model*

|  |  |
| --- | --- |
| Test Case No. | UT001 |
| Description | Retriever model test for relevant document retrieval |
| Input | What is the HS code of Acetone? |
| Expected Output | CONVENTIONS-AGAINST-ILLICIT-TRAFFIC-IN-NARCOTIC- DRUGS-AND-PSYCHOTROPIC-SUBSTANCES-ACT-NO-1-OF-  200.pdf |
| Actual Output |  |
| Result | PASS |

*Table 4 Unit test case 02 - Reader model*

|  |  |
| --- | --- |
| Test Case No. | UT002 |
| Description | Reader model test for answerable question |
| Input | What is the name of the narcotic act? |
| Expected Output | CONVENTIONS AGAINST ILLICIT TRAFFIC IN NARCOTIC DRUGS AND PSYCHOTROPIC SUBSTANCES ACT, No. 1 OF  2008 |
| Actual Output |  |
| Result | PASS |

*Table 5 Unit test case 03 - Reader model*

|  |  |
| --- | --- |
| Test Case No. | UT003 |
| Description | Reader model test for unanswerable question |
| Input | What are the recommendations of the Human Rights Act? |
| Expected  Output | No answer |
| Actual Output |  |
| Result | PASS |

## Integration testing

Testing was conducted to ensure proper communication between the frontend and backend components. This includes verifying that user queries are correctly routed to the pipeline and that answers are displayed accurately in the user interface.

*Table 6 Integration test case 01*

|  |  |
| --- | --- |
| Test Case No. | IT001 |
| Description | Testing long-form answer extraction for a query. |
| Input | What are the SAARC countries? |
| Expected Output | **Answer**: 1. Peples Republic of Bangladesh 2. Kingdom of Bhutan 3. Republic of India 4. Republic of Maldives 5. Kingdom of Nepal 6. Islamic Republic of Pakistan 7. Islamic Republic of Afghanistan  **Document**: Narcotic Act No. 1 of 2008 pdf Document |
| Actual Output |  |
| Result | PASS |

*Table 7 Integration test case 02*

|  |  |
| --- | --- |
| Test Case No. | IT002 |
| Description | Testing short-form answer extraction for a query. |
| Input | What is the name of the narcotic act? |
| Expected Output | **Answer**: CONVENTIONS AGAINST ILLICIT TRAFFIC IN NARCOTIC DRUGS AND PSYCHOTROPIC SUBSTANCES ACT, No. 1 OF 2008  **Document**: Narcotic Act No. 1 of 2008 pdf Document |
| Actual Output |  |
| Result | PASS |

## User Acceptance Testing (UAT)

Real users who are representatives of the target user base participated in usability testing. Feedback was provided on the system's user-friendliness, interface design, and overall user experience.

Real users, including students from Sri Lanka Law College and an external industry specialist (External Supervisor/client), participated in UAT. The system's compliance with requirements and successful delivery were confirmed through an official letter provided in [APPENDIX B.](#_bookmark86)

## Commercialization of the System

The Sri Lankan legal sector holds immense potential for profitability and marketability, yet it remains largely untapped. The solution, the "Sri Lankan Legal Information Retrieval System", is designed to address a pressing challenge within the Sri Lankan legal industry: the absence of a comprehensive, precise, and consistent legal information retrieval system.

In the initial phase of the commercialization efforts, the marketability of the system was successfully validated. The system has been named "AYCA”, and a slogan was crafted as "We Democratize Law" to encapsulate the core vision.



*Figure 18 System logo*

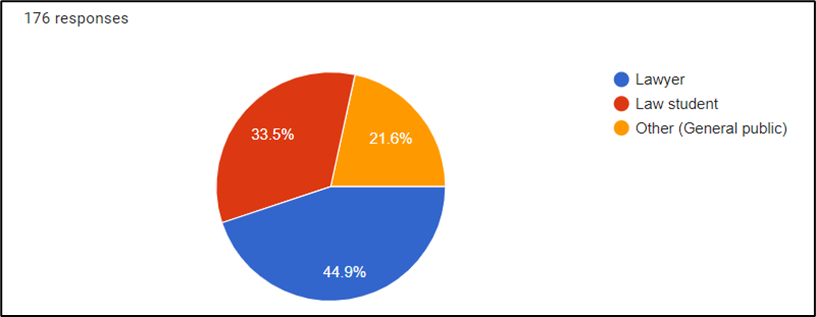
To assess the viability of commercializing the system, a market survey was conducted involving over 170 participants. The survey aimed to gain insights into crucial factors such as market opportunity, competition, and commercialization strategy.

## Market opportunity

The survey confirmed the importance of understanding the target market: Legal professionals (lawyers) emerged as the primary audience of the system, constituting 44.9% of the potential market.

Law students followed closely, making up 33.5% of the target market.

The remaining 21.6% comprises the general public.



*Figure 19 Target audience*

The survey revealed a significant market gap in Sri Lanka for a comprehensive legal information retrieval system. An overwhelming majority (95.4%) of participants recognized the need for such a system, with 87.5% indicating their willingness to adopt it once developed. This strong demand underscores the opportunity for our "Sri Lankan Legal Information Retrieval System" (AYCA) in the market (refer to [APPENDIX A](#_bookmark85) for survey results).

## Market competition

In terms of competition within the legal domain, the market analysis identified only two major systems currently available, both of which primarily serve as legal document catalogs. However, these systems fail to provide effective information retrieval capabilities and lack the comprehensive functionalities offered by the AYCA system.

The market survey further validated this assessment, with 90.3% of respondents concurring that existing systems in the market do not compete with the developed system's functionalities. Many participants expressed their willingness to transition to a new comprehensive system, as they have been using existing systems due to the absence of a suitable alternative. This highlights a clear opportunity for the "Sri Lankan Legal Information Retrieval System" (AYCA) to stand out in the market (refer to [APPENDIX A](#_bookmark85) for survey results).

## Commercialization strategy Target group coverage:

The survey showed that most of the target audience (76.7%), especially lawyers, are willing to pay a premium for a comprehensive system. Thus, a premium version would be offered to individual practitioners and firms to capitalize on the opportunity.

However, law students, the second-largest group, are less inclined to pay a premium due to their limited ROI. To address this, the system would be marketed to legal educational institutions as a package for their students.

The smallest audience, the general public, aligns with the vision of democratizing legal information. They would be offered a free limited featured system.

## Commercialization plan:

The system's commercialization plan spans three years, with a strategic pricing approach:

* **Year 1 - Market expansion:** Offer the system at a highly competitive price to aggressively gain market share. Profitability is not the primary focus during this phase; breaking even is the goal.

Prioritize rapid market growth over immediate profits.

* **Year 2 - Stabilization:** With a significant market presence, stabilize the service provision.

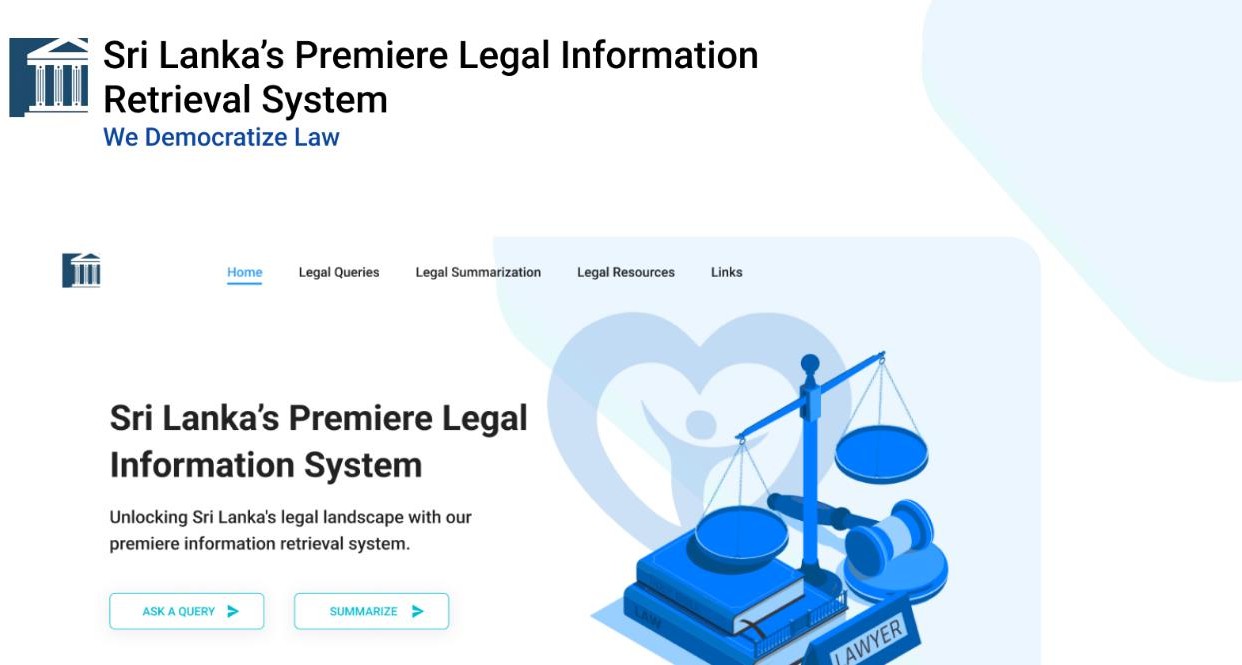
Shift to a flat-rate pricing model for consumers. Profitability remains a secondary concern as market consolidation takes precedence.

* **Year 3 and Beyond - Profitability focus:** In the third year and beyond, shift the focus to profitability. Reevaluate and restructure the pricing model, offering packaged plans and pay-as-you-go services.

Prioritize revenue generation and profitability while continuing to provide value to customers.

This phased approach allows for aggressive market capture in the initial phase, followed by stabilization and, ultimately, a focus on profitability as the system matures and gains a substantial user base.

Following this strategic commercialization plan, the system is well-positioned to capitalize on the untapped Sri Lankan legal industry. By aggressively expanding market share in the initial year, stabilizing services in the second year, and prioritizing profitability from the third year onward, the system aims to provide competitive legal information retrieval services to its target audiences. This phased approach ensures both market penetration and long-term sustainability.



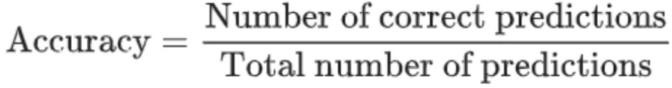
*Figure 20 Commercialization poster*

# RESULTS AND DISCUSSION

## Results

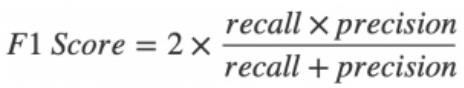
The results of the developed system have been thoroughly assessed using key evaluation metrics, including accuracy, F1 score, and precision, for both the retriever model and the reader model. These evaluations provide valuable insights into the system's performance and effectiveness.

* **Accuracy**: The accuracy of the model is quantified with this metric. The number of correct predictions is divided by the total number of predictions to calculate accuracy [25].



*Figure 21 Accuracy metric*

* **F1-Score**: To obtain the F1-score, the product of recall and precision is divided by the sum of recall and precision [10].



## Reader model (Doc2Vec)

*Figure 22 F1 Score metric*

The choice of a Doc2Vec model for the reader component was a strategic decision, driven by its superior performance compared to the TfIdf approach. The evaluation results demonstrated the advantage of the Doc2Vec model, as it achieved perfect performance scores with an accuracy and F1 score of 1.0.

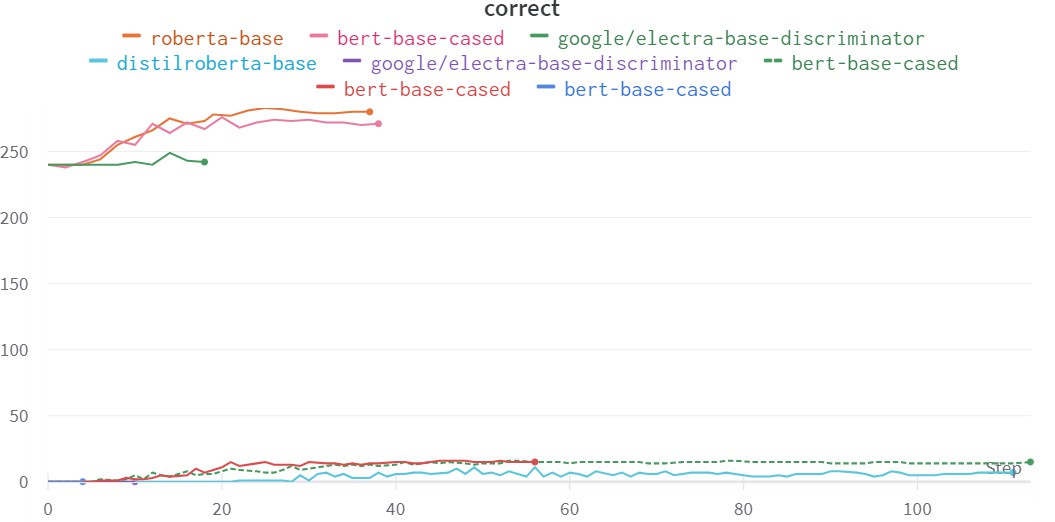
In contrast, the TfIdf approach fell short, yielding scores of 0.5, which did not align with the system's desired level of accuracy. This highlights the effectiveness and reliability of the Doc2Vec model in comprehending and extracting legal documents from the corpus, making it the preferred choice for the reader component.

*Table 8 Results - Retriever model*

|  |  |  |
| --- | --- | --- |
| **Approach** | **Accuracy** | **F1 Score** |
| TfIdf | 0.5 | 0.5 |
| Doc2Vec | 1.0 | 1.0 |

## Retriever model (Finetuned RoBERTa)

After evaluating several transformer-based approaches, including BERT, ELECTRA, and distilBERT, the decision was made to implement a RoBERTa-base based approach for the reader model [29]. This choice was driven by the exceptional performance of the RoBERTa-base [29] model in legal question answering, even before fine-tuning on the prepared dataset (refer to the below figure).



*Figure 23 Transformer based model performance comparison*

The finalized RoBERTa model, after training on the prepared dataset, delivered outstanding results, achieving an accuracy rate of 95.33% and an F1 score of 96.11%. These high-performance metrics underscore the effectiveness of the RoBERTa model in accurately and comprehensively answering legal questions, making it a valuable asset for the system's success.

*Table 9 Results - Reader model*

|  |  |  |
| --- | --- | --- |
| **Model** | **Accuracy** | **F1 Score** |
| RoBERTa (finetuned) | 95.33 | 96.11 |

Furthermore, detailed metrics regarding the training and evaluation loss during the RoBERTa model's training have been meticulously collected. These metrics provide valuable insights into the model's performance and facilitate vigilant monitoring for signs of overfitting.

During the 50-epoch training cycle, a substantial reduction in both the training loss, which decreased from 0.0944 to 2.351e-07, and the evaluation loss, which decreased from 5.321 e-06 to 3.650 e-07 was observed.

* **Training loss decrease**: The significant decrease in training loss over each epoch, from 0.0944 to 2.351e-07 after 50 epochs, is indicative of the model's ability to learn and fit the training data very well. This decreasing trend is generally expected during the training process.

It suggests that the model is effectively minimizing the error between its predictions and the actual training data labels. As the loss decreases, the model becomes more accurate in representing the training dataset.

* **Evaluation loss decrease:** The decrease in evaluation loss throughout training, from 5.321 e-06 to 3.650 e-07, signifies that the model is also performing well on data it hasn't seen during training. This indicates that the model is not only memorizing the training data (overfitting) but is also generalizing to unseen data. Lower evaluation loss is desirable, as it reflects the model's ability to make accurate predictions on new, real-world data.
* **Tapering off losses:** Towards the end of the training cycle, it could be observed, that both the training and evaluation losses start to taper off, and the decrease becomes less rapid. This behavior is typical in deep learning.

It suggests that the model has reached a point where it has learned most of the relevant patterns in the data, and further fine-tuning may yield diminishing returns. This phase is often referred to as "convergence."

*Table 10 Training and evaluation losses - Reader model*

|  |  |  |
| --- | --- | --- |
| **Epoch Number** | **Training Loss** | **Evaluation Loss** |
| 1 | 0.0944 | 5.321 e-06 |
| 48 | 3.039 e-07 | 4.895 e-07 |
| 49 | 2.533 e-07 | 4.201 e-07 |
| 50 | 2.351e-07 | 3.650 e-07 |

These results serve as compelling evidence of the Legal Question Answer system's effectiveness and precision. The impressive accuracy rate of 95.33% and the high F1 score of 96.11% demonstrate the system's capability to provide accurate and reliable answers to legal questions. Such performance undoubtedly enhances the overall legal information retrieval system, making it a valuable tool for legal professionals, students, and the general public seeking precise and efficient access to legal knowledge.

## Research Findings

One of the key findings during the research timeline was the superior performance of the Doc2Vec model in the context of legal document retrieval. This superiority stemmed from the Doc2Vec model's unique ability to comprehend the context of both questions and documents, a crucial aspect of effectively retrieving legal information.

This advantage can be attributed to the vector-based approach employed by Doc2Vec [27], as opposed to traditional scalar approaches. The vector approach allows Doc2Vec to capture the semantic relationships and nuances within legal texts [28], making it exceptionally adept at understanding and matching the complex language and context found in legal documents and queries.

Another significant finding from the research was the impact of semantic and syntactical differences in languages from various geographical regions on the understandability of transformer-based models [29]. Before fine-tuning, these pre-trained models, although trained on extensive English language corpora, exhibited lower accuracy when extracting answers to legal queries specific to Sri Lanka. However, after fine-tuning to align with the Sri Lankan context, there was a notable increase in accuracy and overall performance (refer to the [Results](#_bookmark70) section).

This discovery highlights the critical influence of regional language nuances on machine understanding of natural language. It emphasizes that transformer-based models, even with their immense capabilities, may struggle with context-specific language intricacies until adapted to the specific linguistic and cultural context. Fine-tuning plays a crucial role in bridging this gap and enhancing the effectiveness of such models in region-specific applications like legal information retrieval.

Another significant finding during the research was the presence of ambiguity in legal concepts and the unique language, document structures, and terminology used within the legal field. This presented a notable challenge for Natural Language Processing (NLP) approaches, as there is a clear distinction between legal language and everyday language.

The research emphasized that a one-size-fits-all NLP approach might not be effective for legal information processing. Instead, tailored approaches are necessary to understand the specific nuances of legal concepts and terminology. This specialized understanding is instrumental in developing an accurate and reliable legal question-answering system.

In essence, the study highlighted the importance of domain-specific knowledge and tailored approaches when working with legal texts and demonstrated the need for NLP systems to adapt to the intricacies of legal language and concepts.

## Discussion

The development of the Sri Lankan Legal Question Answering System represents a pioneering achievement in the domain of legal information retrieval. This comprehensive research effort has culminated in a system that offers a dynamic and efficient approach to addressing complex legal queries.

The foremost objective of this research was the creation of a legal decision support system, proficient in answering user queries related to Sri Lankan legal documents. This objective has been unequivocally met through the meticulous development of a question-answering pipeline. Within this framework, the retriever model, utilizing a Doc2Vec-based technique, adeptly identifies pertinent legal documents, while the reader model, anchored by the powerful RoBERTa architecture, accurately extracts context-aware answers. These models harmonize seamlessly to provide users with precise and comprehensible responses to their legal queries.

The development of the Legal Question Answering System involved overcoming multiple challenges, showcasing the adaptability of the research process. Future improvements hold the promise of extending its utility, further democratizing access to legal knowledge, and staying aligned with the evolving landscape of technology and legal requirements.

## Limitations and challenges faced

Various limitations and challenges were faced during the development of the system and during the research timeline and various assumptions and steps were taken to overcome the challenges.

* The research encountered a significant challenge in dealing with the complexity of legal language, which often contains highly specialized terminology. To bridge the gap and ensure that the system provides answers understandable to a broader audience, close collaboration with domain experts was essential.
* Post-implementation, it was observed that the question-answering pipeline required approximately 10 seconds to generate answers. This latency primarily stemmed from hardware and resource limitations in the development environment.

To optimize system performance, it is proposed that, as proof of concept has been established for the success of the legal question-answering system, significant capital investment could be directed toward hosting the models in a GPU-enabled cloud environment. This infrastructure upgrade would substantially enhance processing speed and overall system efficiency, further enhancing the system's usability and accessibility.

## Future Work

Even though the research phase has been completed, there are several promising avenues for future development and expansion of the Legal Question Answering System:

* **Widening the scope of legal knowledge:** The system initially focused on a specific subset of the legal domain to develop the question-answering system. As a future direction, the scope could be broadened to encompass a wider range of legal knowledge and practice areas. This expansion would enhance the system's utility and make it more versatile for a broader audience of legal professionals, students, and the public.
* **Multilingual support (Sinhala and Tamil):** Currently, the system primarily operates in English. A significant future enhancement could involve adapting the system to work with other languages relevant to the Sri Lankan context, such as Sinhala and Tamil. This multilingual capability would make the system more inclusive and accessible to a larger segment of the population, including those more comfortable with local languages.
* **High-resource environment implementation:** While the research has demonstrated the system's capabilities in a limited environment, future development could involve implementing the system in a high-resource environment. This investment could significantly improve system performance, including processing speed and scalability. A well-resourced environment would facilitate more extensive and efficient legal information retrieval.

In summary, while the current research provides a solid proof of concept for the Legal Question Answering System, there is ample room for future enhancements and expansion. These proposed directions would not only improve the system's performance but also increase its accessibility, usability, and overall impact on the legal community in Sri Lanka.

# CONCLUSION

In culmination, the development of the Sri Lankan Legal Question Answering System stands as a resounding success, reflecting the achievement of the research's primary objectives and its pivotal role in strengthening the broader Legal Information Retrieval System.

The success of this system lies in the adept utilization of advanced Natural Language Processing techniques and transformer-based models, notably, RoBERTa, which exhibits remarkable accuracy in comprehending complex legal language. The collaborative interplay between the retriever and reader models within the question-answering pipeline ensures that user queries, regardless of their intricacy, are met with precise and context- aware responses.

This system seamlessly integrates into the overall Legal Information Retrieval System, enhancing its efficacy. By providing users with the capability to ask specific legal questions and receive tailored answers, it improves the accessibility and usability of the entire legal database. Legal professionals, students, and laypersons alike can now navigate the intricate web of legal documents with ease, empowering them with timely, accurate, and understandable legal insights.

In the Sri Lankan context, the system bears immense significance. It not only bridges the gap between legal experts and the general populace but also streamlines the process of accessing and comprehending legal information. Its utility extends beyond the confines of academic institutions and legal chambers, reaching into the lives of ordinary citizens seeking clarity on legal matters. It aligns perfectly with the objective of making legal knowledge more accessible to all, fostering a legal environment rooted in transparency, understanding, and empowerment.

In summation, the Sri Lankan Legal Question Answering System represents a remarkable achievement in the domain of legal information retrieval. It serves as a testament to the power of innovation and dedication, effectively bridging the gap between intricate legal

language and user accessibility. This system holds immense promise for legal practitioners, law students, and anyone navigating the complex terrain of Sri Lankan legal documentation.

As the system evolves and undergoes continuous refinement, it stands poised to become an indispensable tool, democratizing access to legal knowledge. Its potential to contribute to a more informed and just society is undeniable. With the proposed optimizations and investments on the horizon, the system's expanded capabilities and its profound impact on the legal landscape are anticipated. It is through pioneering efforts like this that the future of legal information retrieval is being transformed and democratized.

# REFERENCES

|  |  |
| --- | --- |
| [1] | C. Hoppe, D. Pelkmann, N. Migenda, D. Hötte and W. Schenck, "Towards Intelligent Legal Advisors for Document Retrieval and Question-Answering in German Legal Documents," in *IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Laguna Hills, CA, USA, 2021. |
| [2] | R. O. Moorthage and H. P. K. N. Hewawasam, "A Study on the Impact of Legal Literacy on Crime Control and Prevention (With Special Reference to Selected Communities in Rathnapura Police Division in Sri Lanka)," *Vidyodaya Journal of Humanities and Social Sciences,* vol. 7, no. 02, 2022. |
| [3] | P. Hemp, "Death by Information Overload," *Harvard Business Review,* 2009. |
| [4] | A. V. Tambimuttu, "GlobaLex," 2009. [Online]. Available: https:/[/www.nyulawglob](http://www.nyulawglobal.org/globalex/Sri_Lanka.html)a[l.org/globalex/Sri\_Lanka.html.](http://www.nyulawglobal.org/globalex/Sri_Lanka.html) |
| [5] | J. S. C. Secretariate, "jsc.gov.lk," 16 June 2011. [Online]. Available: https:/[/www.jsc.gov](http://www.jsc.gov.lk/).[lk/.](http://www.jsc.gov.lk/) |
| [6] | N. L. A. D. S. BOARD, "natlib.lk," NATIONAL LIBRARY AND DOCUMENTATION SERVICES BOARD, 2021. [Online]. Available:  [http://www.natlib.lk/home/eresources.php.](http://www.natlib.lk/home/eresources.php) |
| [7] | N. Evans and J. Price, "Managing information in law firms: changes and," *Information Research: PUBLISHED QUARTERLY BY THE UNIVERSITY OF BORÅS, SWEDEN.,* vol. 22, no. 01, 2017. |
| [8] | H. Karunarathna, "Use of Electronic Resources by Law Degree Students at Anuradhapura Regional Centre of the Open University of Sri Lanka," *Journal of the University Librarians Association of Sri Lanka,* vol. 18, no. 2, p. 41–60, 2015. |
| [9] | A. Cammelli and E. Fameli, "Legal information access on the Web. A project for an information system model," in *12th International Workshop on Database and Expert Systems Applications*, Munich, Germany, 2001. |
| [10] | N. X. Bach, L. T. N. Cham, T. H. N. Thien and T. M. Phuong, "Question analysis for Vietnamese legal question answering," in *9th International Conference on Knowledge and Systems Engineering (KSE)*, Hue, Vietnam, 2017. |

|  |  |
| --- | --- |
| [11] | S. Pudaruth, R. P. Gunputh, K. M. S. Soyjaudah and P. Domun, "A Question Answer System for the Mauritian Judiciary," in *3rd International Conference on Soft Computing & Machine Intelligence (ISCMI)*, Dubai, United Arab Emirates, 2016. |
| [12] | H.-T. Duong and B.-Q. Ho, "A Vietnamese Question Answering System in Vietnam’s," in *Computer Information Systems and Industrial Management. CISIM*, Springer, Berlin, Heidelberg, 2015. |
| [13] | T. N. R. L. R. F. A. Peng Qian, "Structural Guidance for Transformer Language Models," in *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021. |
| [14] | Kenton, J. D. M.-W. Chang and L. K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for," in *naacL-HLT*, 2019. |
| [15] | D. Lamba and W. H. Hsu, "Answer-Agnostic Question Generation in Privacy Policy Domain using Sequence-to-Sequence and Transformer Models," in *International Conference on Electronics, Communications and Information Technology (CECIT)*, Sanya, 2021. |
| [16] | Simple Transformers, "simpletransformers.ai," 16 July 2020. [Online]. Available: https://simpletransformers.ai/docs/qa-data-formats/. |
| [17] | N. Abadani, J. Mozafari, A. Fatemi, M. A. Nematbakhsh and A. Kazemi, "ParSQuAD: Machine Translated SQuAD dataset for Persian Question Answering," in *7th International Conference on Web Research (ICWR)*, Tehran, 2021. |
| [18] | ,. M. M. C. L. G. V. C. A. I. Vasile P˘ais, "Named entity recognition in the Romanian legal domain," *Natural Legal Language Processing Workshop 2021,* p. 9–18, 10 November 2021. |
| [19] | F. D. ,. H. L. H. I. R. B. C. L. T. Muhamad Abdurahman, "Lex2KG: Automatic Conversion of Legal Documents to Knowledge Graph," in *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Jakarta, 2021. |
| [20] | Z. Q. T. H. B. L. Hao Lian, "Knowledge Graph Construction Based on Judicial Data with Social Media," in *14th Web Information Systems and Applications Conference*, 2017. |
| [21] | T. X. F. Z. Zhangbo Xua, "Construction of Chinese Sports Knowledge Based on Neo4j," in *IEEE 2nd International Conference on Civil Aviation Safety and*  *Information Technology (ICCASIT)*, 2020. |

|  |  |
| --- | --- |
| [22] | K. Kowsrihawat and P. Vateekul, "An Information Extraction Framework for Legal," in *12th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, 2015. |
| [23] | A. Venkataramana, K. Srividya and R. Cristin, "Abstractive Text Summarization Using BART," in *IEEE 2nd Mysore Sub Section International Conference (MysuruCon)*, Mysuru, 2022 . |
| [24] | N. Begum and A. Goyal, "Analysis of Legal Case Document Automated Summarizer," in *6th IEEE International Conference on Signal Processing, Computing and Control (ISPCC 2k21)*, Solan, 2021. |
| [25] | H.-L. Nguyen, T.-B. Nguyen, T.-M. Nguyen, H.-T. Nguyen and H.-Y. T. Vuong, "Retrieving legal document and extracting answer with BERT-based model," in *14th International Conference on Knowledge and Systems Engineering (KSE)*, Nha Trang, 2022. |
| [26] | H. N. Van, D. Nguyen, P. M. Nguyen and M. L. Nguyen, "Miko Team: Deep Learning Approach for Legal Question Answering in ALQAC," in *14th International Conference on Knowledge and Systems Engineering (KSE)*, Nha Trang, 2022. |
| [27] | D. Yogish, T. N. Manjunath and R. S. Hegadi, "Survey on trends and methods of an intelligent answering system," in *International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, Mysuru, 2017. |
| [28] | P. Karvelis, D. Gavrilis, G. Georgoulas and C. Stylios, "Topic recommendation using Doc2Vec," in *International Joint Conference on Neural Networks (IJCNN)*, Rio de Janeiro, 2018. |
| [29] | A. Rawat and S. S. Samant, "Comparative Analysis of Transformer based Models for Question Answering," in *2nd International Conference on Innovative Sustainable Computational Technologies (CISCT)*, Dehradun, 2022. |
| [30] | P. Yu and Y. Liu, "Roberta-based Encoder-decoder Model for Question Answering System," in *International Conference on Intelligent Computing, Automation and Applications (ICAA)*, Nanjing, 2021. |
| [31] | P. Yu and Y. Liu, "Roberta-based Encoder-decoder Model for Question Answering System," in *International Conference on Intelligent Computing, Automation and Applications (ICAA)*, Nanjing, 2021. |

|  |  |
| --- | --- |
| [32] | C. Hoppe, D. Pelkmann, N. Migenda, D. Hötte and W. Schenck, "Towards Intelligent Legal Advisors for Document Retrieval and Question-Answering in German Legal Documents," in *IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, Laguna Hills, CA, USA, 2021. |
| [33] | P. O. T. D. S. R. O. S. LANKA, "RIGHT TO INFORMATION ACT," *Gazette of the*  *Democratic Socialist Republic of Sri Lanka,* 2016. |
| [34] | H. Lian, Z. Qin, T. He and B. Luo, "Knowledge Graph Construction Based on Judicial Data with Social Media," in *14th Web Information Systems and Applications Conference (WISA)*, Liuzhou, China, 2017. |
| [35] | S. Sitthithai, B. Manaskasemsak and A. Rungsawang, "Legal Search System using Ontology for Government Agencies in Lao PDR," in *Seventh ICT International Student Project Conference (ICT-ISPC)*, Nakhonpathom, Thailand, 2018. |
| [36] | D. K. Mahto and L. Singh, "A dive into Web Scraper world," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, India, 2016. |
| [37] | S. Gharatkar, A. Ingle, T. Naik and A. Save, "Review preprocessing using data cleaning and stemming technique," in *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, Coimbatore, India, 2017 . |
| [38] | S. K. Vishwakarma, K. I. Lakhtaria, D. Bhatnagar and A. K. Sharma, "An Efficient Approach for Inverted Index Pruning Based on Document Relevance," in *Fourth International Conference on Communication Systems and Network Technologies*, Bhopal, India, 2014. |
| [39] | M. Virkar, V. Honmane and S. U. Rao, "Humanizing the Chatbot with Semantics based Natural Language Generation," in *International Conference on Intelligent Computing and Control Systems (ICCS)*, Madurai, India, 2019. |
| [40] | A. Stranieri and J. Zeleznikow, "Tools for intelligent decision support system development in the legal domain," in *12th IEEE Internationals Conference on Tools*  *with Artificial Intelligence. ICTAI*, Vancouver, BC, Canada, 2000. |

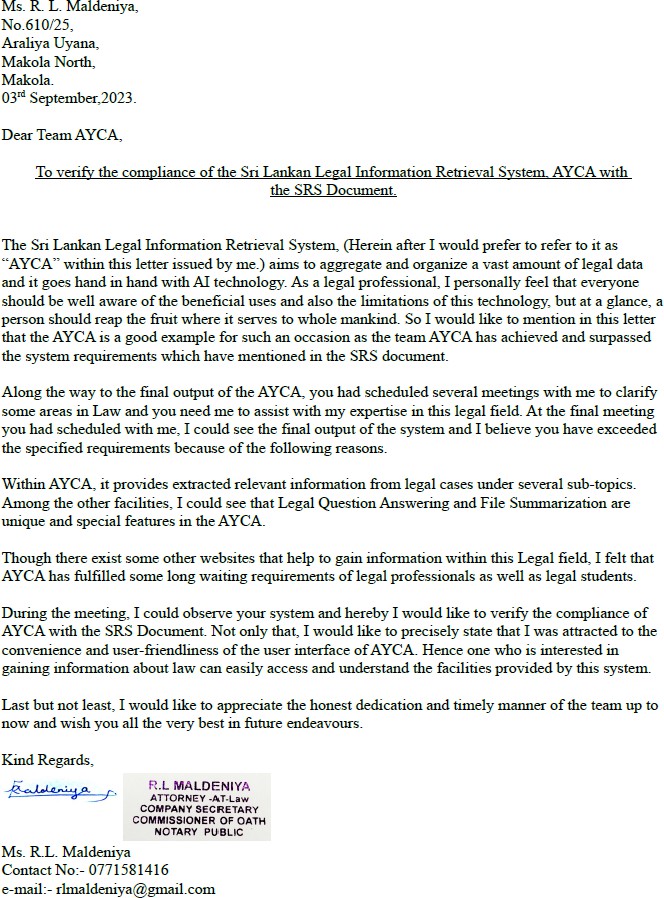
# APPENDIX A: SURVEY RESULTS

|  |  |
| --- | --- |
| 01 | A pie chart with numbers and a diagram  Description automatically generated |
| 02 |  |
| 03 | A diagram of a pie chart  Description automatically generated |

|  |  |
| --- | --- |
| 04 |  |
| 05 |  |
| 06 | A blue and orange pie chart  Description automatically generated |

|  |  |
| --- | --- |
| 07 |  |
| 08 | A graph with numbers and a bar  Description automatically generated |
| 09 |  |

# APPENDIX B: CLIENT APPROVAL LETTER



# APPENDIX C: PERSONNEL AND FACILITIES

As the research project was entangled with the legal domain, which is a distinct educational and technical field beyond the knowledge base of Information Technology, domain specialists were enlisted to aid in the project's development.

Legal domain specialists were consulted to acquire the necessary knowledge about legal concepts, terminology, technical constraints, and ethical considerations that were encountered during the development of the system. The main contributors included:

* Ms. R. L. Maldeniya, Attorney-at-law, LLB (Hons) University of Colombo, Notary Public.
* Mr. Bilaal Marikar, Treasurer of the Law Students' Union of Sri Lanka.

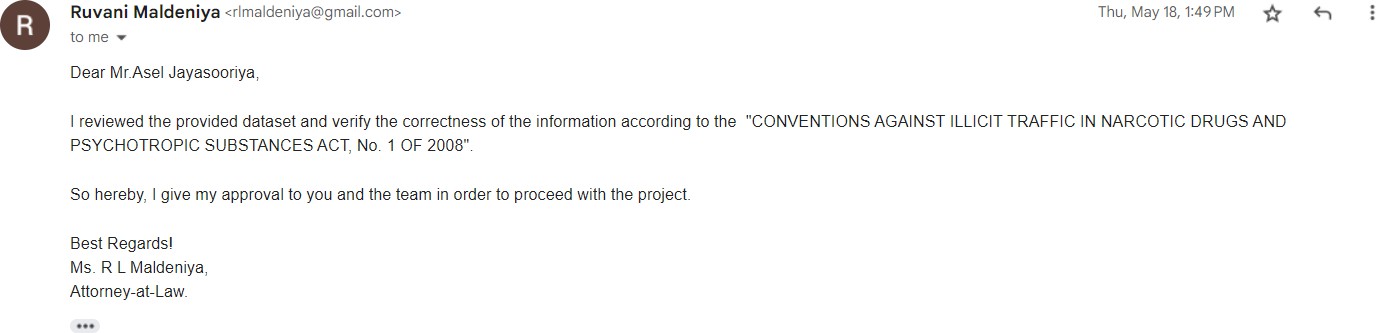
These individuals played a pivotal role in the development of the system, providing domain-specific insights, addressing domain-specific challenges and issues, and ultimately validating the system.

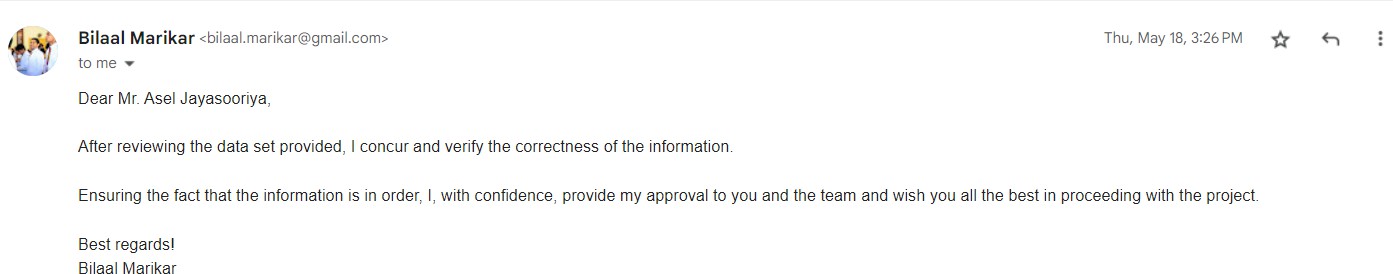
The research project has been completed, and the facilities utilized include government institutions' legal information websites, which were instrumental in gathering relevant data for the system's development.

In addition to the aforementioned personnel, the support of the supervisor and co- supervisor was essential throughout the research. They provided technical supervision and guidance, which proved instrumental in successfully completing the project:

* Dr. Lakmini Abeywardhana (Supervisor)
* Dr. Dharshana Kasthurirathna (Co-supervisor)

# APPENDIX D: DATASET APPROVAL





# APPENDIX E: TURNITIN REPORT

