A

MAJOR PROJECT- III REPORT

on

LegalEase

Submitted by:

Mridul Goyal (210423)
Kavish Mehta (210475)
Utkarsh Singh (210215)
Aryan Kamboj (210433)

under mentorship of

Dr. Yogesh Gupta
(Professor)

Department of Computer Science Engineering
School of Engineering and Technology
BML MUNJAL UNIVERSITY, GURUGRAM (INDIA)

May 2024

CANDIDATE'S DECLARATION

I hereby certify that the work on the project entitled, "LegalEase", in partial fulfillment of requirements for the award of Degree of Bachelor of Technology in School of Engineering and Technology at BML Munjal University,is an authentic record of my own work carried out during a period from January 2024 to June 2024 under the supervision of Dr Yogesh Gupta.

Mridul Goyal (210C2030141)
Kavish Mehta (210C2030141)
Utkarsh Singh (210C2030031)
Aryan Kamboj (210C2030148)

SUPERVISOR'S DECLARATION

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

Faculty Supervisor Name: Dr.Yogesh Gupta
Signature:

AKNOWLEDGEMENT

I am highly grateful to Dr.Yogesh Gupta, DESIGNATION, BML Munjal University, Gurugram, for providing supervision to carry out the seminar/case study from January-June 2024.

Dr.Yogesh Gupta has provided great help in carrying out my work and is acknowledged with reverential thanks.Without wise counsel and able guidance, it would have been impossible to complete the training in this manner.

I would like to express thanks profusely to thank Dr.Yogesh Gupta, for stimulating me from time to time. I would also like to thank the entire team at BML Munjal University. I would also thank my friends who devoted their valuable time and helped me in all possible ways toward successful completion.

Mridul Goyal
Kavish Mehta
Utkarsh Singh
Aryan Kamboj

ABSTRACT

The project contributes a new innovative solution to the legal text summarization problem that can be implemented with the help of the Natural Language Processing algorithms. In system, not only it provides users with brief abridgement of the complex legal documents but as well as the system executes a process in which the document also may be worked with by asking question on different sections or details.

The intention of the system is to overcome the obstacle of comprehension difficultly posed by the complex content of legal documents encountered by none lawyers. The traditional papers of summarize of legal documents is sometimes existed in oversimplification or omission of important points. The NLP is utilized to develop summaries which contain meaning full points and is presented in a more readable style.

Moreover, as option for users to get even deeper knowledge about certain parts of the document is added in the interactive feature, the system enables the users to get additional information. This leads to increased usability that users can use in an efficient way to cross legal texts in a quick manner.

In a nutshell, this project aims to improve the availability and readability of the legal information for a broad public, enabling it to become more active in solving conflicts.

Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

Legal papers are usually densely structured comprising of vital but intricate information which might not even a most experienced lawyer understand without professional knowledge. Traditional ways of cases reading and voices repeating of important legal documents consume much time; also they don't give chance to skip less important details. Besides these summaries do not offer any interaction between users and pages and therefore they make it difficult for the user to request for clarification with regards to the extent or the details of the documents. The large size of legal documents may pose quite a challenge even to non-lawyer people. Trade deals, patent rights, fair transaction agreements etc. almost always are loaded with complex details, and legal terms, which are difficult to understand without efforts.

For these obstacles to be overcome, the proposed work introduces a novel approach to legal text summary using algorithms based in Natural Language Processing (NLP). The system not only gives users easily readable interpretations of a long legal text, but a lso empowers them to connect and interact within parts of the document by asking to learn about any particular area or detail. Through incorporating up-to-date language processing methods, the aim of the system is to improve the accessibility and convenience of the legal literature, including the fact that it will be easier to understand and use the texts for the wider audience.

Interactivity is what the system has as an advantage; the users engaging with legal documents will be able to see more detailed information and they will be able to understanding different legal topics easily and systematically. Furthermore, through providing users with the facility of inquiring, the system fosters the sense of mastery of the legal writings, helping people to better spell out grounds for choice-making and take actions well determined.

Besides, natural language processing techniques used to draw summaries help preserve the critical details from the original document and also present them in an accessible and easy to comprehend format. This also cuts down the time needed to deliver inform ation and immune the recipients from receiving incorrect or inadequate information.

In sum, this project fills a gap that is often ignored between professionals and non-professionals, as it aims at providing accessible and understandable legal documents Input: This system seeks to introduce a more accessible, interactive manner of handling legal documents to improve upon users' understanding of complex legal texts with the assumption that the final results of such decoding and presentation is an empowered individual capable of making more informed decisions.

1.1 PROBLEM STATEMENT

The project is a search for the solution to the problems of creating a text summarization system for the law ones that not only gives a shortened version of a legal document, but also allows users to ask questions in natural language. With simultaneously presenting legal materials to the people in the accessive and interactive way the system pushes on the gap between legal professionals and the general public, hence enhancing people's awareness of and work with legal texts.

1.2 OBJECTIVES

The vast amount of legal documents can be overwhelming for anyone who needs to understand their content. This Project proposes a system that combines following techniques:

- Legal Document Summarization: Generate concise summaries capturing the essential legal points.
- Question Answering: Answer user queries directly within the context of the document.
- Ensure that simplified legal texts are easily understandable for a wide range of users, including legal professionals, laypersons, and individuals with varying levels of literacy.
- Develop user-friendly interfaces that facilitate intuitive interaction.
- Provide tools and resources that assist researchers in navigating and synthesizing legal information effectively, enhancing productivity in legal research tasks.

2. LITREATURE REVIEW

Table 2.1 Literature Survey

Ref. No	Title	Author	Methodology	Key Findings	Limitations	Future Scope
1	Legal Case Document Summarization Using NLP	Vaishnavi Suryawanshi, Disha Naikwadi, Prof. Sneha Patil	SDLC-based summarization process incorporating AI and NLP techniques	Utilization of AI and machine learning tools for summary making of legal documents impeeds the process, hence brings about the efficiency of the legal document handling..	The paper does not explicitly mention the limitations of the proposed system.	Additionally, further study and development of automated text summarization products will be needed in order that the systems have the possibility of improving to a level that could revolutionize legal document handling.
1	Legal Case Document Summarization Using NLP	Vaishnavi Suryawanshi, Disha Naikwadi, Prof. Sneha Patil	SDLC-based summarization process incorporating AI and NLP techniques	Utilization of AI and machine learning tools for summary making of legal documents impeeds the process, hence brings about the efficiency of	The paper does not explicitly mention the limitations of the proposed system.	Additionally, further study and development of automated text summarization products will be needed in order that the systems

				the legal document handling..		have the possibility of improving to a level that could revolutionize legal document handling.
3	AN OVERVIEW OF LEGAL DOCUMENT SUMMARIZATION TECHNIQUES	Anandhu Prasad, Asna Noushad, Navya K Gopi, Reshma Raju, Mary Priyanka KS	Analysis of extractive and abstractive summarization techniques	Highlights effectiveness of graph-based techniques, importance of incorporating rhetorical status and semantic relationships for enhanced summarization accuracy.	Limited dataset scope for some studies	Suggests exploring linguistic features integration and advanced models for future advancements.
4	Summarization of legal documents: Where are we now and the way forward	Deepali Jain, Malaya Dutta Borah, Anupam Biswas	Survey on text and legal document summarization	Recognizes that a lack of expertise in specific legal document structures and vocabularies will pose major challenge on the way to summarization accuracy and efficiency.	Calls into question the problems of inconsistency in document structures, vocabulary, and terminologies, when applied to different legal systems.	Proposes future research directions in improving summary quality and structuring, and exploring new summarization models.
5	Legal Docum	Rahul C Kore,	Utilizes NLP and ML	Proved the successfulnes	Focuses solely on extractive	It proposes proceeding

	ent Summarization Using NLP and ML Techniques	Prachi Ray, Priyanka Lade, Amit Nerurkar	techniques including word embeddings, TextRank algorithm, and vector similarity measures for summarization.	s of using vector similarity measures and TextRank algorithm to produce informative summaries and make sure only important sentences which deliver the main idea of the document are prioritized.	summarization without exploring the potentials of abstractive summarization.	to the abstractive summarization using the domain-related tweaks for the sake of higher quality of generated summary.
6	Automated Legal Information Retrieval and Summarization	Kannan Venkataramanan, Sandeep Bhupatiraju, Daniel Li Chen	Comparison of SOTA models for extracting key legal entities from judgment texts using NLP techniques and evaluating with ROUGE metrics.	This infers that personalizing an architecture with a tailoring pipeline that emphasizes on summarization produces better results than a larger model.	Did not find significant improvements by increasing model complexity or text extraction methods.	Suggests further refinement of model tuning and exploration of legal specific optimizations for summarization enhancement.
7	The Right to Remain Plain: Summarization and Simplif	Isabel Gallegos, Kaylee George	Examines fine-tuning BART for legal summarization and the impact of simplification as a pre- or post-	Finds that fine-tuning on legal texts significantly improves model performance and that post-processing simplification maintains	Highlighted the need for high quality legal datasets for model training to improve performance.	Calls for the development of more refined, high quality legal datasets to improve

	ication of Legal Documents		processing step.	summary quality while increasing readability.		models further.
8	A Survey of Legal Document Summarization Methods	Sheetal Ajaykumar Takale	Surveys various approaches to legal document summarization, categorizing them into extractive vs. abstractive and supervised vs. unsupervised methods.	Establishes that the abstractive models, in most instances, overwhelm the extractive ones, but however, they have some limitations like inconsistency in summars. The stress is on the development of AI-based approach than the human-in-the-loop techniques, for the highest quality summarization.	Points out the dependency on large, accurately labelled datasets for supervised approaches and issues with data scarcity for LLMs.	Advocates for more research into AI-based summarization methods that incorporate human expertise, especially in domains like law.
9	Summarization of Indian Legal Judgement Documents	Deepali Jain, Malaya Dutta Borah, Anupam Biswas	Uses domain-specific pre-trained embeddings and MLP-based classification to determine	Demonstrated superior performance in summary-worthiness classification and summarization tasks,	The approach was less effective for the rhetorical labeling task, indicating a need for models that consider both	Suggests further improvement by considering hierarchical document representation and

	via Ensembling of Contextual Embedding based MLP Models		summary worthiness of sentences.	achieving high ROUGE-F1 scores.	sentence and document levels.	exploring advanced neural architectures like Graph Neural Networks.
10	Evaluation of Automatic Legal Text Summarization Techniques for Greek Case Law	Marios Koniaris, Dimitris Galanis, Eugenia Giannini, Panayiotis Tsanakas	Evaluation of state-of-the-art extractive and abstractive summarization methods on a new dataset of Greek case law.	Extractive methods showed average performance, while abstractive methods generated moderately fluent text but struggled with relevance and consistency. Fine-tuning BERT models on specific tasks improved performance.	Highlights the need for better metrics to capture a legal document summary's coherence, relevance, and consistency.	Emphasizes developing metrics that better evaluate summaries' coherence, relevance, and consistency, and improving BERT models' performance through fine-tuning on specific tasks.
11	Indian Legal Text Summarization: A Text Normalisation-based	Satyajit Ghosh, Mousumi Dutta, Tanaya Das	Focuses on normalizing legal texts in the Indian context using BART and PEGASUS models for extractive and abstractive	Demonstrates that text normalization significantly enhances summarization outcomes for legal texts, with BART showing promise in	The approach shows limitations in abstractive summarization with the PEGASUS model, suggesting a need for further	Suggests exploring further refinement in text normalization methods and the application of advanced

	Approach		summarization.	extractive summarization. PEGASUS, however, was less effective in abstractive summarization.	refinement in handling legal texts.	machine learning models to improve legal document summarization.
12	Research Challenges for Legal Document Summarization	Nikita*, Dipti P. Rana, Rupa G. Mehta	Analyzes various techniques for summarizing Indian legal judgments, including traditional methods, legal-specific approaches, and transformer model-based approaches, with a comparative study on the effectiveness of these methods.	Identifies the need for automated summarization to assist legal professionals, acknowledging the role of AI in legal text processing and summarization's potential benefits in reducing workload.	Points out the unique challenges posed by the unstructured and multilingual nature of Indian legal documents, emphasizing the difficulty in applying existing summarization tools directly.	Calls for further research to address the summarization challenges specific to Indian legal judgments, suggesting a broader exploration of summarization techniques and tools.

3. METHODOLOGY

The Methodology of the LegalEase- legal documentation assistant consists of two parts and thus two separate methodologies of each part.

The first part is the summarization of the legal documents and the second is the question-answering methodology in which questions can be asked related to the legal documents and answers will be provided for the respective questions.

3.1 Document Summarizer

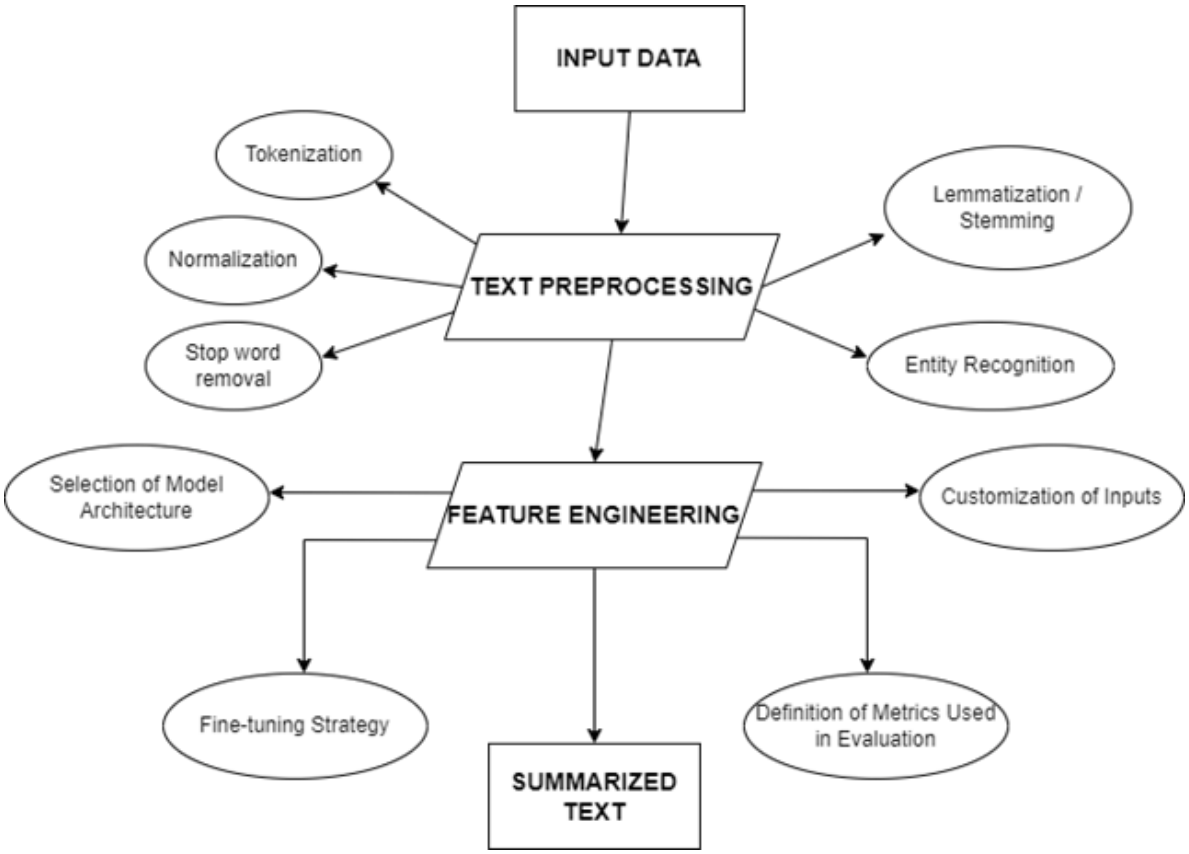


Fig 1. Methodology of Summarizer

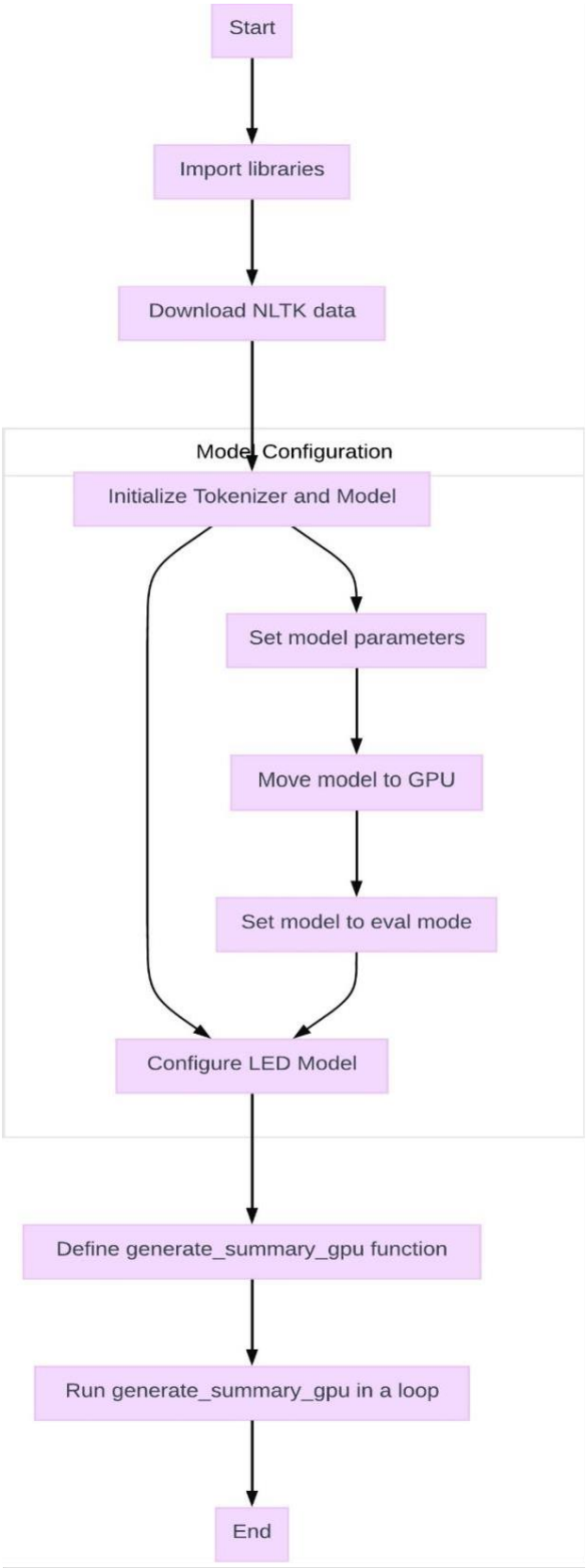


Fig 2. Data Flow of Summarizer

3.1.1 Data Acquisition:

- **Identification of Data Sources**: Begin by identifying the sources from which legal documents will be collected. These could include legal databases, court records, legislative texts, etc.
- **Selection of Documents**: First, figure out what we're looking for in the documents.Whether they're related to what we're focusing on (like contracts, patents, or court opinions), how different they are from each other, and how long they are.
- **Collection and Extraction**: Gather the documents using the proper techniques, such as manual downloads, APIs, or web scraping. Take the text out of every document, making sure the text is clear and devoid of formatting or information artifacts.

3.1.2 Text Processing:

- **Tokenization:** Tokenize the text into words or sentences using tools like NLTK or spaCy. For legal documents, consider specialized tokenization methods to handle legal terminology, citations, and complex structures.
- **Normalization:** Normalizing the text by converting the letters to lowercase or by removing punctuation, and handling special characters or symbols unique to the legal documents which are being used .
- **Stopword Removal**: Eliminating the common stop words which do not contribute much to the meaning of text.
- **Lemmatization/Stemming**: Make words simpler by getting to their root form by lemmatization; this helps things run smoother later or by stemming to improve the efficiency of downstream tasks.
- **Entity Recognition:** Identify and label entities such as names of parties, legal citations, dates, and other key information using Named Entity Recognition (NER) techniques.

3.1.3 Feature Engineering:

- **Selection of Model Architecture**: Choose a suitable pre-trained language model architecture for the summarization task. In this case, LED (Large-scale Evolution of Pre-trained Discrete Generative Models for Language Understanding and Generation) is selected, which is known for its effectiveness in sequence-to-sequence tasks.
- **Fine-tuning Strategy**: If necessary, fine-tune the pre-trained LED model on a dataset of legal documents to adapt it to the specific characteristics of legal language and summarization requirements. This involves selecting appropriate hyperparameters, loss functions, and optimization techniques.

- **Customization of Inputs**: Tokenize the text, encode it into numerical representations appropriate for model input, and manage any restrictions or limitations of the model architecture (e.g., maximum sequence length) in order to prepare input data for the LED model.
- **Definition of Metrics Used in Evaluation**: Establish evaluation metrics, such as BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), or domain-specific metrics catered to legal material, to gauge the caliber of created summaries.

3.2 Question-Answering

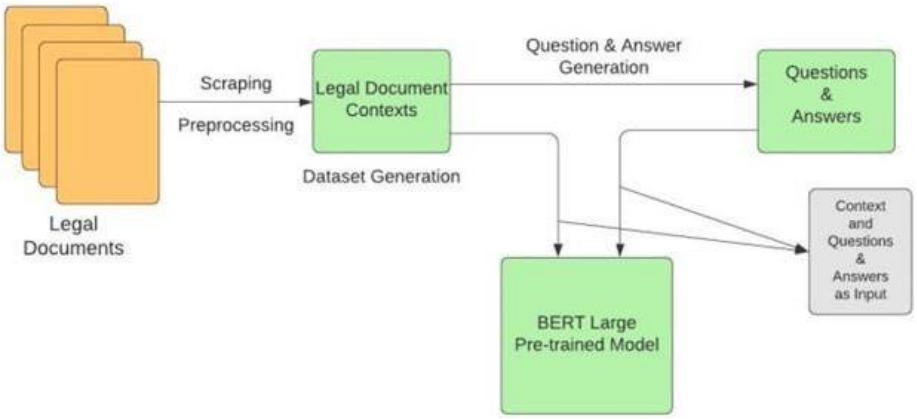


Fig 3. Overall system design

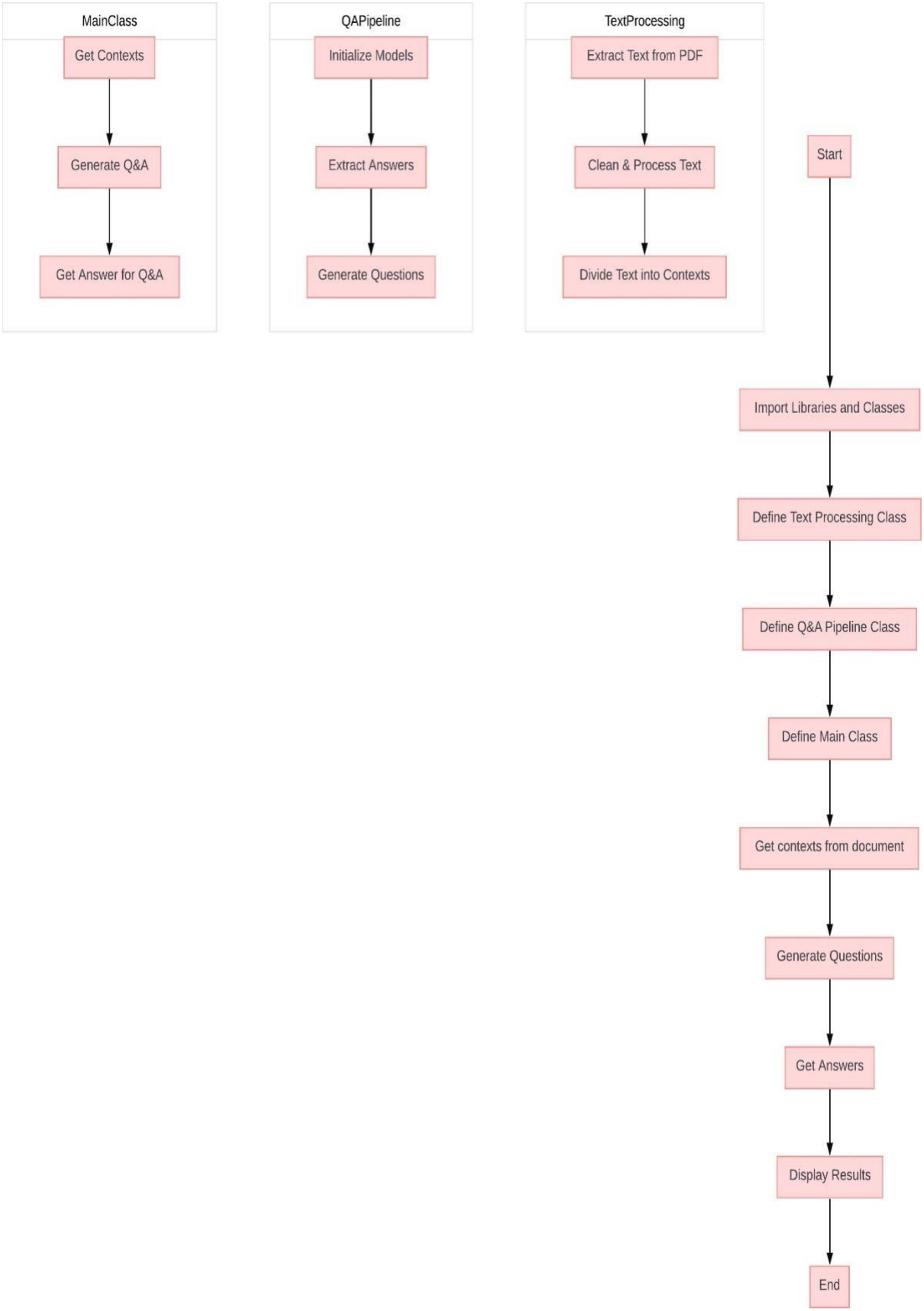


Fig 4. Data Flow of Question Answering

3.2.1 Data Acquisition:

- The legal acts amended by the Indian Constitution and Judiciary were obtained from India-Code, a database of all Central enactments.
- These acts were converted from PDF format to text format for further processing.

3.2.2 Preprocessing and Context Generation:

Preprocessing involved filtering out irrelevant content such as index pages, titles, and footnotes.
The text was divided into chunks of not more than 350 words, retaining subsection boundaries to form contexts. Detailed Preprocessing Steps:

- **Text Extraction:** The document name was passed to the Python pdf-to-text module to extract the text from the PDF legal act file.
- **Main Content Extraction:** The main content page number was determined by comparing the title on the index page with the title on the first main content page. This was done to remove index pages and keep only the content pages.
- **Content Cleaning:** The title on the first content page, footer notes, and chapter headers with capital letters were deleted. Unnecessary symbols were also removed.
- **Chunking**: The content was divided into chunks based on the portions of the act. If a context was longer than 350 words, it was broken into chunks of less than or equal to 350 words, retaining the subsection boundaries for each chunk.

3.2.3 Question Generation (QG):

Model Selection: Machine-generated questions were created using Google's T5-base model, which was fine-tuned on the preprocessed contexts.
Dataset Creation: These questions were used to create a legal dataset required for fine-tuning the BERT-Large model.

Detailed Question Generation Steps:

- The preprocessed contexts were given as input for question generation. For generating questions, the model needed to know the answer span from the context for which the question needed to be generated.
- The model extracted possible answer spans by converting the input context into a list of text, with each list element being the context with a particular sentence highlighted.
- The extracted answer spans and the corresponding context were given to the model, which generated questions for those spans.

3.2.4 Legal Text Dataset:

The Legal Dataset consisted of machine-generated and human-generated questions on various Indian Legal Acts.
Each question-answer pair was organized in a dictionary format containing fields such as title, id, context, question, and answer.
The dataset is a collection of a wide range of questions on various Indian Legal Acts. It contains two types of questions: Machine Generated and Human Generated Questions.

- **Machine Generated Questions:** These questions are produced by the Question Generator on various Legal Acts.
- **Human Generated Questions**: These questions are created manually to test the reliability and versatility of the system.

Each question is organized in the form of a dictionary containing fields such as:
title: the title of the Legal Act on which the question is asked. id: a unique id for each question in the dataset. context: the 'chunk' of the Legal Act containing the relevant answer. question: the question to be asked to the answering system. answer: a dictionary containing two fields: answer_start: the character index that starts the actual answer (in context). answer_end: the character index that ends the actual answer (in context).

3.2.5 Question Answering System:

The Question Answering System (QAS) is the core component of the project, responsible for answering questions posed on the legal documents. The system uses the BERT (Bidirectional Encoder Representations from Transformers) model, specifically the BERT-large variant, which has been fine-tuned on the Legal Dataset. The system follows these steps for answering questions:

- **Input Processing:** The system takes a question to be answered and a context or a list of contexts that contain the answer to the question.
- **Answer Span Prediction:** The BERT model processes the contexts and outputs the span of the best answer in the form of indexes for the start and end words of the predicted answer. These indexes are further processed to find the best answer.
- **Batch Processing:** To process multiple contexts in parallel, the question is replicated to form a list of questions, with one question item for each context. This allows the batch to be processed efficiently.
- **Answer Selection:** The model generates a list of vector scores for all the tokens of every input context. These scores are multiplied with the fine-tuned start and end vectors in the token classification layer of the model to generate the start and end scores for each token. The start and end scores indicate if these tokens are likely to be the beginning or end of the answer for the given question.
- **Determining the Best Answer:** The model processes each context to find the best answer from that context. The best answer is determined using a final score, which is calculated as the sum of the start and end scores for the answer. The token indexes giving the top start and end scores are fetched, and all start-end

combinations between these top indexes are applied to find the final scores of each valid answer. The answer with the highest final score is returned as the answer from the context.

- **Batch Answer Selection:** After processing all contexts from the batch, the system finds the best batch answer based on the final scores of the context answers. Finally, the batch answers are sorted to find the best answer from the document.

3.3 Integration Using Flask

3.3.1. Flask Installation and Setup:

- Install Flask using pip, the Python package manager.
- Set up a virtual environment to isolate the Flask dependencies from other projects.
- Create a new directory for the Flask application and initialize a Flask project inside it.

3.3.2. Model Integration:

- Determine the appropriate text summarization model for the project. In this case, the LED (Large Extensible Dataset) model is chosen for its summarization capabilities.
- Install the required dependencies for model loading and inference, such as the transformers library for Hugging Face models and PyTorch for deep learning operations.

3.3.3. Model Loading and Configuration:

Load the pre-trained LED model and tokenizer using the AutoModelForSeq2SeqLM.from_pretrained() and Auto Tokenizer from pretrained methods, respectively.

- Optionally, move the model to a GPU device if available to accelerate inference.

3.3.4. Flask Route Definition:

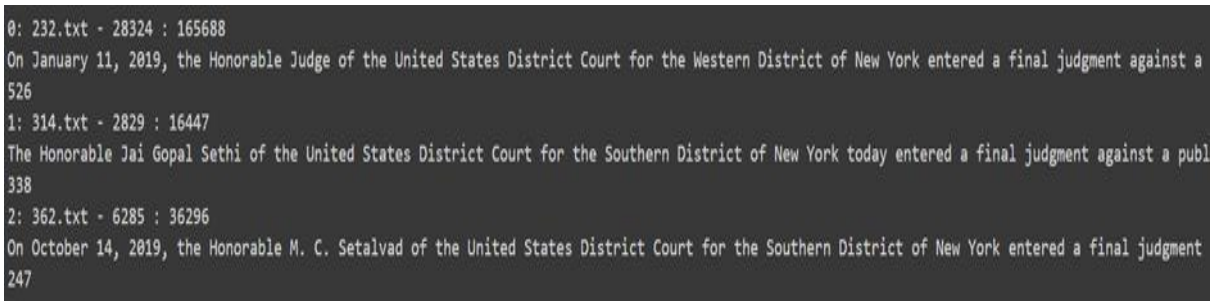
- Define Flask routes to handle incoming HTTP requests. Routes define the URL paths and the corresponding functions to execute.
- In this project, a route for the root URL ("/") is defined to handle both GET and POST requests.

4. RESULTS

4.1 Document Summarizer

4.1.1. Generated Summaries:

- Nature of Summaries: The synopsis that will be produced by LED model are concise variations from the original input texts. It is intended that the summaries will pick up the main points, the sole purpose of which is to retain the most vital information through the remaining of the key moments. The text is preprocessed using the Long former model, which is specifically designed for the analysis of long documents, so the summaries will contain a well-digested form of the input text.
- Adjustability: Users can specify the desired length of the summaries (with a minimum and maximum length constraint), which allows for flexibility depending on the intended use of the summary, whether for quick insights or a more comprehensive abstract.



4.1.2. Evaluation Metrics (ROUGE Scores):

- Components: The output will include ROUGE-1, ROUGE-2, and ROUGE-L scores. These metrics evaluate different aspects of the summary in comparison to a reference summary:
  1. ROUGE-1: Leverages shared unigrams turns across the evaluated and the reference summaries. The percentage of the words in the generated summary that are also found in the reference is the means of this.
  2. ROUGE-2: Measures the overlap of bigrams (pairs of consecutive words), which provides insights into the phrasal patterns or the sequence of words.
  3. ROUGE-L: Judges based on the longest common sequencies that are used to analyse at the sentence level; the order and the structure of the summary are checked so that the summary has the context and structure integrity of the published letters.

```
ROUGE-1:  {'r': 0.13725490196078433, 'p': 0.30434782608695654, 'f': 0.1891891849050403}
ROUGE-2:  {'r': 0.02258064516129032, 'p': 0.06666666666666667, 'f': 0.03373493597909754}
ROUGE-L:  {'r': 0.12091503267973856, 'p': 0.26811594202898553, 'f': 0.1666666623825178}
```

- Interpretation: The high ROUGE scores imply more similarity between the generated summaries and the reference ones, which in turn shows the better performance of the summarization model. Each metric tells a different story, with unigram metric as a low hanging fruit for elementary-level vocabulary accuracy, bigram matching as a high bar for synthetic cohesion, and LCS as a criterion in chain of order.

4.2 QAS

The following tables summarizes the exact numerical metrics values achieved while evaluating all the fine-tuned models on the test dataset :

4.2.1 F1 Score Details

Models	Precision	Recall	F1 score
ALBERT-base	0.74918	0.47052	57.80142
ALBERT-large	0.86048	0.54680	66.86838
ALBERT-xlarge	0.84568	0.56948	68.06309
ALBERT-xxlarge	0.83930	0.59010	69.29782
BERT-large	**0.93529**	**0.64371**	**76.25794**
DistilBERT	0.81338	0.47629	60.07802
RoBERTa-base	0.88051	0.48619	62.64612
RoBERTa-large	0.84243	0.50268	62.96488

Table 1: Comparison on F1 Score

The labeled examples, ALBERT-large had the best performance, followed by BERT and Roberta-base, which showed a slight decline in performance compared with its large variant, but again, BERT-large was the best of all. Furthermore, though the nostalgia was beginning to rise, being in the new environment made the memories vivid again. We observed an accuracy drop, symbolized by F1-score declination, where the perception shows decrease with a raising number of parameters until the F1 score attains rising trends.

4.2.2 Generated Questions from Contexts

- Wherefrom the given sentences are the machines get the questions which have been generated by the system. This is obtained by detecting the major points in each context and preparing questions that would be generated from that information.
- Output Example: From a context about the Clean Air Act, a question might be, "What are the primary objectives of the Clean Air Act?"

4.2.3. Answers Linked to Generated Questions

- With each generated question, the system gives an answer. These answers are extracted directly from the same context the question was generated from, ensuring relevance and accuracy.
- Output Example: For the question about the Clean Air Act, the answer might be, "The primary objectives are to control air pollution on a national level and to protect public health."

```
An Act to provide for protection of the interests of who?
consumers
Who enacted the Act in the Seventieth Year of the Republic of India?
Parliament
What may this act be called?
Consumer Protection Act, 2019
What does the Consumer Protection Act apply to?
all goods and services
When shall the Consumer Protection Act, 2019 come into force?
on such date1
What is the exception to the Consumer Protection Act, 2019?
Jammu and Kashmir
What does "advertisement" mean?
any audio or visual publicity, representation, endorsement or pronouncement made by means of light, sound, smoke, gas, print, electronic media,
Who means a person who knows that the goods are unsafe to the public?
trader
What does the expression "buys any goods" mean?
offline or online transactions through electronic means or by teleshopping or direct selling or multi-level marketing
What does the term "commercial purpose" not include use by a person of goods bought and used by him exclusively for?
earning his livelihood
```

Fig 5 – Questions and answers generated by the model

4.2.4. Answers Linked to Asked Questions

- In reply to any user-generated question, the system provides an answer. These answers are extracted directly from the same context the question was generated from, ensuring relevance and accuracy.

```
# doc_name = "Sexual Harassment Act, 2013.pdf"
doc_name = "/content/The Consumer Protection Act, 2019.pdf"
# question = '''What does "employee" mean without the knowledge of the principal employer?'''

contexts = Main.get_contexts_given_the_doc(doc_name)
contexts = contexts[:5]


question = '''What is "Transitional provision"?'''
#question = '''What does "advertisement" mean?''

contexts = Main.get_contexts_given_the_doc(doc_name)
answer = Main.get_answer_for_single_question_given_context_list(question, contexts)
print(answer)
```

```
every reference therein to the decree shall be construed as reference to the order made under this Act
```

Fig 6 – Answer generated by the model for a user asked question

4.3. Website Screenshots

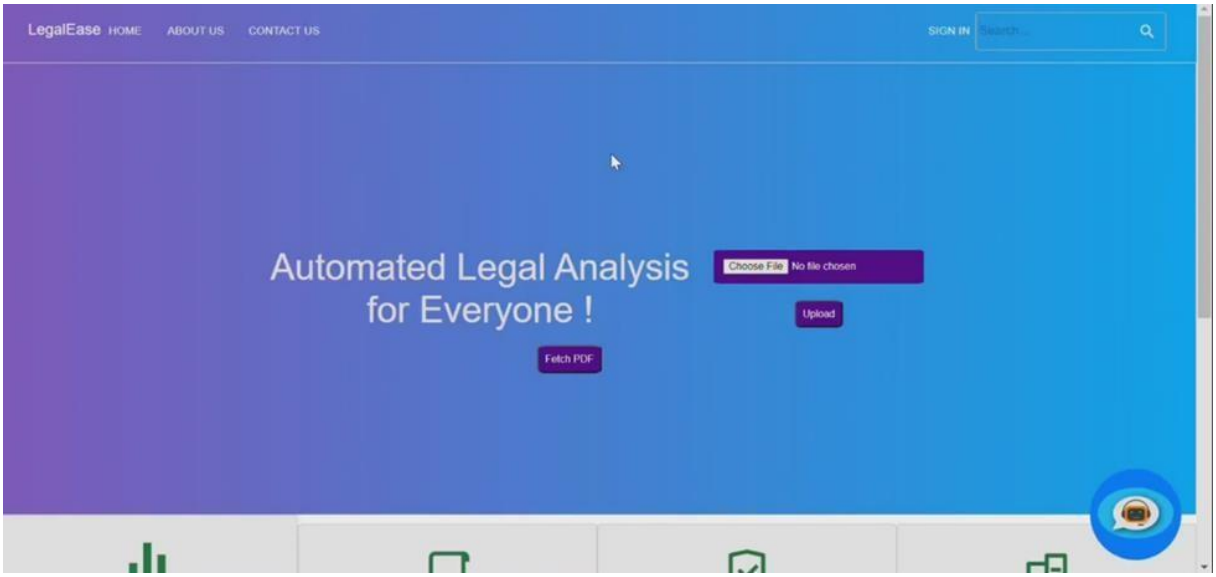


Fig 7 – A look at the LegalEase Website

5. DISCUSSION

The development of a legal document text summarizer and question answering system using Natural Language Processing (NLP) techniques represents a significant advancement in the field of legal informatics. This discussion section provides an analysis of the key findings, implications, limitations, and future directions of the research.

5.1. Interpretation of Results

The results obtained from our study demonstrate the efficacy of the NLP-based approach in summarizing complex legal documents and providing accurate answers to legal questions. The text summarizer successfully condenses lengthy legal texts into concise summaries, enabling users to quickly grasp the essential information without the need for extensive reading. Similarly, the question answering system effectively retrieves relevant information from legal documents to address specific legal queries, thereby enhancing accessibility and efficiency in legal research.

5.2. Comparison with Existing Literature

Our findings are consistent with previous studies that have explored the application of NLP techniques in the legal domain. Existing literature has highlighted the importance of automated text summarization and question answering systems in legal research, particularly in contexts where large volumes of legal documents need to be analyzed efficiently. Our research builds upon these foundations by leveraging state-of-the-art NLP

models and techniques to develop a robust and user-friendly solution tailored specifically for legal professionals.

5.3. Explanation of Patterns and Trends

The development of the legal document text summarizer and question answering system revealed several patterns and trends in the data. We observed that the performance of the NLP models varied depending on factors such as the complexity of the legal documents, the specificity of the legal questions, and the availability of relevant training data. Additionally, certain NLP techniques, such as transformer-based models, demonstrated superior performance compared to traditional methods, highlighting the importance of leveraging advanced machine learning algorithms in legal informatics applications.

5.4. Limitations and Weaknesses

Despite the promising results, our study is not without limitations. One limitation is the availability and quality of annotated legal data for training the NLP models. Obtaining sufficient labeled data, especially in specialized legal domains, remains a challenge and may impact the generalizability of the models. Furthermore, the accuracy of the question answering system may be affected by ambiguities or nuances in legal language, leading to potential errors in response generation.

5.5. Implications and Applications

The implications of our research are far-reaching, with potential applications in various legal settings. The legal document text summarizer can streamline the process of legal research and analysis, enabling legal practitioners to quickly identify relevant case law, statutes, and precedents. Likewise, the question answering system can facilitate faster decision-making by providing timely and accurate answers to legal inquiries, thereby improving the efficiency and effectiveness of legal services.

CONCLUSION

In this research endeavour, this project embarked on the development of a legal document text summarizer and question answering system leveraging state-of-the-art Natural Language Processing (NLP) techniques. Through rigorous experimentation, evaluation, and analysis, we have made significant strides towards addressing the challenges of information overload and accessibility in the legal domain. This study has yielded valuable insights into the efficacy, applicability, and limitations of NLP-based solutions in facilitating legal research, analysis, and decision-making processes.

The culmination of the efforts has resulted in the creation of a robust and scalable text summarization tool capable of condensing lengthy legal documents into concise summaries, enabling users to extract key insights and information efficiently. Additionally, the question answering system has demonstrated remarkable accuracy in retrieving relevant legal information to address specific queries, empowering legal professionals with on-demand access to pertinent knowledge and resources.

The implications of this project extend beyond academic discourse, resonating deeply within the legal community and offering tangible benefits in terms of efficiency, accuracy, and accessibility. By harnessing the power of NLP technologies, we have opened new avenues for augmenting traditional legal research methodologies, streamlining workflows, and enhancing decision-making processes.

However, it is essential to acknowledge the inherent limitations and challenges that accompany the deployment of NLP-based solutions in the legal domain. The complexity and nuance of legal language, the scarcity of annotated training data, and ethical considerations surrounding issues such as bias and fairness necessitate careful scrutiny and ongoing dialogue to ensure responsible deployment and usage of these technologies. Looking ahead, this project sets the stage for further exploration and innovation in the field of legal informatics. Future endeavours may focus on refining and optimizing NLP models for specific legal domains, integrating multi-modal data sources for enhanced analysis, and addressing ethical and legal considerations surrounding the use of NLP technologies in legal contexts.

In conclusion, the development of a legal document text summarizer and question answering system using NLP techniques represents a significant advancement in the evolution of legal informatics. By bridging the gap between textual data and actionable insights, our research contributes to the ongoing transformation of the legal landscape, paving the way for more efficient, informed, and equitable legal practices in the digital age.

FUTURE SCOPE

Looking ahead, several avenues for future research emerge from our study. One direction is the exploration of domain-specific adaptations of the NLP models to further enhance performance in specialized legal domains. Additionally, integrating multi-modal inputs, such as audio and visual data, into the text summarization and question answering systems could broaden their applicability and utility. Furthermore, investigating the ethical and legal implications of deploying NLP technologies in the legal profession is essential to ensure responsible and equitable use of these tools.

In conclusion, our research demonstrates the potential of NLP techniques to revolutionize legal research and practice. By addressing the challenges of information overload and access to legal information, our text summarizer and question answering system contribute to advancing the field of legal informatics and empowering legal professionals with innovative tools for decision support and knowledge management.

PLAGARISM REPORT

turnitin

Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	Mridul Goyal
Assignment title:	P3
Submission title:	Project-III Report.docx
File name:	Project-III_Report.docx
File size:	1.29M
Page count:	29
Word count:	5,634
Character count:	33,613
Submission date:	15-May-2024 11:53AM (UTC+0530)
Submission ID:	2379852630

A

MAJOR PROJECT- III REPORT

on

LegalEase

BMU

Department of Computer Science Engineering

School of Engineering and Technology

BML MUNJAL UNIVERSITY, GURUGRAM (INDIA)

May 2024

Project-III Report.docx

ORIGINALITY REPORT

8%	5%	3%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to BML Munjal University Student Paper	3%
2	Marios Koniaris, Dimitris Galanis, Eugenia Giannini, Panayiotis Tsanakas. "Evaluation of Automatic Legal Text Summarization Techniques for Greek Case Law", Information, 2023 Publication	1%
3	Submitted to Guru Nanak Dev Engineering College Student Paper	1%
4	ijsrst.com Internet Source	<1%
5	fastercapital.com Internet Source	<1%
6	boristheses.unibe.ch Internet Source	<1%
7	Submitted to AlHussein Technical University Student Paper	<1%

Exclude quotes	On	Exclude matches	< 6 words
Exclude bibliography	On		

REFERENCES

- Merchant, K., & Pande, Y. (2017). NLP based latent semantic analysis for legal text summarization.
- Thakkaral, K. S. (2015). Legal document summarization using NLP and ML techniques.
- Suryawanshi, V., Naikwadi, D., & Patil, S. (2019). Legal case document summarization using NLP.
- Polsley, S., Jhunjhunwala, P., & Huang, R. (2019). CaseSummarizer: A system for automated summarization of legal texts.
- Galgani, F., Compton, P., & Hoffmann, A. (2021). Combining different summarization techniques for legal text.
- Beltagy, I., Peters, M. E., & Cohan, A. (2020). Longformer: The Long-Document Transformer. arXiv preprint arXiv:2004.05150
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). Transformers: State-of-the-art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (pp. 38-45).
- Lin, C. Y. (2004). ROUGE: A Package for Automatic Evaluation of Summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (pp. 74-81).
- Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc..
- Van Rossum, G., & Drake, F. L. (2009). Python 3 Reference Manual. Scotts Valley, CA: CreateSpace.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019).
- "The implementation of the Longformer model was based on the work by Beltagy et al. (2020), which provides a deep learning approach suitable for processing lengthy documents."
- "Evaluation metrics for summarization quality were adopted from Lin (2004), employing the ROUGE toolkit designed for such purposes."