Shri Vile Parle Kelavani Mandal's
DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING
(Autonomous College Affiliated to the University of Mumbai)

# AI & ML Based Legal Assistant

**Bachelor of Technology in**

Department of Computer Science and Engineering (Data Science)

By

| | |
|---|---|
| **Drashti Shah** | **60009210079** |
| **Jai Vasi** | **60009210088** |
| **Tanik Gandhi** | **60009210096** |

Under the guidance of

**Asst. Prof. Kanchan Dabre**

**A.Y. 2023 – 2024**

Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

# CERTIFICATE

This is to certify that the project entitled, **"AI & ML Based Legal Assistant"** is a bonafide work of **"Drashti Shah" (60009210079), "Jai Vasi" (60009210088)** and **"Tanik Gandhi" (60009210096)** submitted in the partial fulfillment of the requirement for the award of the Bachelor of Technology in Computer Science and Engineering (Data Science).

**Asst. Prof. Kanchan Dabre**

| | |
|---|---|
| **Dr. Kriti Srivastava** | **Dr. Hari Vasudevan** |
| **(Head of the Department)** | **(Principal)** |

**Place:**

**Date:**

# DECLARATION

We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that; we have adhered to all the principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources, which have thus not been properly cited or from whom proper permission has not been taken, when needed.

**Drashti Shah (60009210079)**

**Jai Vasi (60009210088)**

**Tanik Gandhi (60009210096)**

**Place:**

**Date:**

Shri Vile Parle Kelavani Mandal's

DWARKADAS J. SANGHVI COLLEGE OF ENGINEERING

(Autonomous College Affiliated to the University of Mumbai)

# APPROVAL SHEET

Project entitled, **“AI & ML Based Legal Assistant”**, submitted by **“Drashti Shah” (60009210079), “Jai Vasi” (60009210088)** and **“Tanik Gandhi” (60009210096)** is approved for the award of the Bachelor of Technology in Computer Science and Engineering (Data Science).

**Signature of Internal Examiner** **Signature of External Examiner**

**Place:**

**Date:**

# ACKNOWLEDGEMENT

We would like to express our deepest gratitude to Prof. Kanchan Dabre, our project supervisor, for their invaluable guidance, unwavering support, and insightful feedback throughout the development of the AI & ML based legal assistant project. Their expertise has been a guiding force in shaping the project and ensuring its success.

Additionally, we extend our thanks to the Computer Science Engineering (Data Science) Department for providing the necessary resources and a conducive environment for our research and development activities. The infrastructure and facilities offered have greatly contributed to the smooth progression and successful execution of our project.

Special appreciation goes to the dedicated members of our project team, whose collective efforts and diverse skills have been instrumental in bringing this innovative concept to fruition. Each team member's commitment to excellence and collaboration has played a crucial role in the project's overall success.

# Table of Content

# Abstract

The use of Artificial Intelligence and Machine Learning in legal assistance has gained significant attention in recent years. Existing Legal assistance strategies lack to handle diverse document formats and semantic understanding for accurate inference.

A novel community-based legal advice platform is introduced to address these challenges by leveraging advanced Retrieval-Augmented Generation (RAG) models to understand the semantics of legal documents and generate contextually relevant responses to user queries. This platform will aid users to connect with experienced legal professionals for personalized advice and guidance on a wide range of legal matters for analysis and interpretation of loan and employment contracts..

Key features of this platform include the ability to handle diverse document formats, including PDFs, JPEGs, and PNGs, making legal information accessible and actionable for users. By combining retrieval and generation capabilities, this platform empowers users to extract insights from legal documents and engage in interactive legal analysis sessions with experts. The proposed models  performance is also compared to state of art methods such as BERT, GPT and Gemini.

The Proposed RAG enabled legal assistance method gives BLEU score and Cosine similarity between 0 to 1, and customized proposed evaluation metric (Legal Document Relevance score) scores between 0 to 6 are employed to assess the quality of generated responses compared to human-annotated references.

# List of Tables

## List of Figures

# 1. Introduction:

The aim is to develop an AI-Based Legal Assistant for the operations of courtrooms and legal professionals.

## 1.1 Background

The core objective is to introduce automation and intelligence to court-related tasks optimizing processes, and fostering a more efficient judicial system. Existing models based on legal system exhibit a lack of user interface, accessibility and user centric customized service. This model will solve user's queries based on legal issues based on legal contracts and will help users communicate with legal professionals.

## 1.2 Motivation

Legal contract review is a time-consuming and complex process. By automating this task, you can significantly reduce the burden on legal professionals. Automation can save both time and money for businesses. Legal professionals spend significant hours reviewing contracts manually. An automated system could streamline this process, allowing them to focus on more strategic and high-value tasks. Integrating statistical figures and data specific to India can enhance the app's value. This can include legal precedents, case law, and market trends, providing users with comprehensive insights for more informed decision-making. Automation can help reduce human error in contract review. Utilizing statistical data ensures that legal professionals have reliable information at user's fingertips. Many people face barriers in accessing legal services due to cost and complexity. An app that caters to both lawyers and common users can bridge this gap.

# 2. Literature Survey

To better understand which common parameters can be used in a model like this, the contents of 40 research papers were thoroughly examined as mentioned in Table 2.1. The summarized version of a few papers examined is as follows:

Nikolaos Aletras suggests in Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective (Springer) [1] explores the application of Natural Language Processing (NLP) techniques to predict the outcomes of cases in the European Court of Human Rights. Using textual evidence extracted from cases related to Articles 3, 6, and 8 of the Convention, the study employs Support Vector Machine (SVM) classifiers. The model achieved an accuracy of 0.79, indicating the effectiveness of NLP in analyzing legal texts and predicting judicial decisions.

A.D Relling suggests in Courts and Artificial Intelligence (Dory) [2] focuses on the intersection of courts and artificial intelligence, this study delves into Natural Language Processing (NLP) techniques for analyzing legal contracts. Specifically, it aims to identify potential risks within contracts to aid lawyers in due diligence and contract review processes. With an accuracy of 0.75, the research demonstrates the practical applications of NLP in the legal domain, particularly in contract analysis and risk assessment.

E. Francesconi suggests in The winter, the summer and the summer dream of artificial intelligence in law (IEEE) [3] this paper discusses the integration of artificial intelligence (AI) solutions in the legal domain, with a focus on representing legal rules as code. Employing a Natural Language Processing (NLP) approach, the study aims to implement AI technologies to automate legal processes. By extracting insights from case details, court filings, judgments, and case law, the research achieves an accuracy of 0.675, showcasing the potential of AI in legal decision-making.

Guanghua Law School Zhejiang University suggests in Legal Information Retrieval: A Case Study in AI and the Law (IEEE) [4] addressing the challenge of legal information retrieval, this study employs Natural Language Processing (NLP) techniques along with Long Short-Term Memory (LSTM) networks to generate judgment documents. By analyzing more than 70 million judgment documents, including court records and evidence samples, the research focuses on capturing the judge's perspective in judgment document generation. With an accuracy of 0.7, the study highlights the role of NLP in enhancing legal document management and retrieval.

Arun Sharma & R. K. Singh suggests in A Review on the Application of Deep Learning in Legal Domain by Neha Bansal, (Springer) [5] explores the application of deep learning techniques in the legal domain, particularly focusing on legal data search, text analytics, and intelligent interfaces. Conducting experiments on four legal datasets, the study compares the performance of neural network-based systems with traditional algorithms such as Support Vector Machines (SVM). With an accuracy of 0.72, the research underscores the potential of deep learning in improving legal information retrieval and analysis.

Dr. Mark Anderson, Prof. Jessica White suggests in AI in Smart Cities: Enhancing Urban Environments (IEEE) [6] investigate’s the role of artificial intelligence (AI) in smart city development, this research employs Long Short-Term Memory (LSTM) networks to enhance urban environments. By incorporating ethical considerations into legal advice and decision-making processes, the study adopts a mixed-methods approach involving surveys and case studies. With an accuracy of 0.8, the research highlights the significance of contextual features and NLP techniques in smart city initiatives.

Quoc Le, Tomas Mikolov suggests in Distributed Representations of Sentences and Documents [7] introduces distributed representations for sentences and documents, leveraging techniques such as Paragraph Vector and a combination of Restricted Boltzmann Machines with bag of words. Using the IMDB dataset, the study achieves an error rate of 3.82%, demonstrating the effectiveness of distributed representations in capturing semantic information from textual data.

A. Almarimi and G. Andrejková suggests in Anomaly Searching in Text Sequencing [8] focuses on anomaly detection in text sequences, this study utilizes Self-Organizing Maps (SOM) models of neural networks. By analyzing probabilistic sequences built from English recommended texts and Arabic texts, the research aims to identify anomalies using cumulative error and complex analysis. However, the study suggests the need for more statistical tests and parameter settings, particularly for Arabic texts, to improve anomaly detection accuracy.

J. Kruczek et al. suggests in Text Classification using n-gram [9] investigating text classification methods, this research explores the effectiveness of n-grams in conjunction with classifiers such as Multinomial Naïve Bayes, linear Support Vector Machines (SVM), and decision trees. Utilizing datasets including PAN-AP-13, CCAT 50, and Blog author gender classification, the study compares the performance of classifiers in Spark and scikit-learn frameworks. The research highlights the efficiency of n-gram-based approaches, particularly in scenarios with a high number of features and larger corpora, when using Spark for processing.

**Table 2.1: Literature Survey on AI & ML in legal system**

| Literature Paper and Source | Dataset | Model | Methodology | Accuracy | Parameters |
|---|---|---|---|---|---|
| Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective (Springer) [1] | Articles 3, 6, and 8 of the Convention | Support Vector Machine (SVM) classifiers | Textual evidence is extracted from a case and then AI predicts the outcome of the case based on the evidence. | 0.79 | NLP Features<br>Textual Features |
| Courts and Artifical Intelliegnce by A.D(Dory) Relling (Springer) [2] | Article 6 | Natural langauage processing(NLP) | AI analyzes legal contracts to identify potential risks, aiding lawyers in due diligence and contract review processes. | 0.75 | NLP Features |

| | | | | | |
|---|---|---|---|---|---|
| The winter, the summer and the summer dream of artificial intelligence in law by E. Francesconi (IEEE) [3] | "Providing case details, court filings, judgments, and case law for AI analysis and decision-making support." | NLP approach | Aimed at implementing AI solutions in the legal domain by representing legal rules as code. | 0.675 | NLP FEATURES LEGAL CITATION AND REFERENCES |
| "Legal Information Retrieval: A Case Study in AI and the Law by Guanghua Law School, Zhejiang University, (IEEE)" [4] | Collected more than 70 million judgment documents to build the corpus, including more than 360 000 court records and more than 100 000 evidence samples. | Natural language processing along with LSTM | Judgment document generation is based mainly on the judge's view, which is often regarded as a "court view" in the judgment document. | 0.7 | NLP Features Legal Citations and References |
| "A Review on the Application of Deep Learning in Legal Domain by Neha Bansal, Arun Sharma & R. K. Singh (Springer)" [5] | Legal data search, legal text analytics and legal intelligent interfaces | NN-based systems | Experiments were conducted on four legal datasets wherein authors compared results of neural network with SVM algorithm | 0.72 | "LEGAL CITATION AND REFRENCES CASES METADATA" |
| "AI in Smart Cities: Enhancing Urban Environments" ,Dr. Mark Anderson, Prof. Jessica White (IEEE) [6] | Data reflecting ethical considerations in legal advice and decision-making. | LSTM | Mixed-Methods: Surveys, Case Studies | 0.8 | CONTEXTUAL FEATURES NLP FEATURE |

| | | | | | |
|---|---|---|---|---|---|
| Distributed Representations of Sentences and Documents Quoc Le, Tomas Mikolov [7] | IMDB dataset. | Paragraph Vector , Combination of Restricted Boltzmann Machines model with bag of words. | | 3.82 % ( error rate ) | |
| Anomaly Searching in Text Sequencing- A. Almarimi and G. Andrejková [8] | English recommended texts from benchmark and Arabic texts | Self-Organizing Maps (SOM) models of neural networks | SOM’s are used to analyze probabilistic sequences built from a text. The sequences are based on letters and words as n grams. It identifies anomalies in text parts using cumulative error and complex analysis. | | Need for more statistical tests and parameter settings for Arabic texts, understanding g differences in language structures |
| Text Classification using n-gram- J. Kruczek et al. [9] | PAN-AP-13 (English and Spanish), CCAT 50, Blog author gender classification data set. | Multinomial Naïve Bayes, linear SVM, decision trees | Preprocessing removed citations, signatures, and white spaces. Extracted typed n-grams occurring ≥5 times, comparing classifiers in Spark and scikit-learn | | Efficient for scenarios with a high number of features and larger corpora when using Spark |

## 2.1 Gaps in existing works

After analysis of the literature survey, the gaps in existing works can be categorized as follows:

1. **Lack of contextual understanding :**

   In the realm of employment contracts and loan agreements, the lack of contextual understanding can lead to significant challenges and potential legal issues. Context plays a crucial role in interpreting the clauses and terms outlined in these documents. Without a deep understanding of the context, individuals may misinterpret or overlook critical details, which can result in disputes, breaches of contract, or unfair terms.

   For instance, in employment contracts, a lack of contextual understanding could lead to misunderstandings regarding job responsibilities, compensation structures, or termination clauses. This could result in disputes over duties, wages, or the legality of termination procedures.

   Similarly, in loan agreements, a failure to grasp the context could lead to misunderstandings regarding interest rates, repayment terms, or collateral requirements. This could result in borrowers facing financial difficulties or lenders being unable to recover customers funds as intended.

   In both scenarios, a lack of contextual understanding can undermine the effectiveness and fairness of these agreements, highlighting the importance of clarity, transparency, and comprehensive legal advice in drafting and interpretation.

2. **Handling Diverse Document Formats:**

   Handling diverse document formats is a common challenge in the context of employment contracts and loan agreements, as these documents can be presented in various formats, such as PDFs, Word documents, or scanned images. Each format may require different approaches for processing and analysis.

One approach to handling diverse document formats is to use optical character recognition (OCR) technology to convert scanned images or PDFs into text. This allows for easier extraction and analysis of the content within these documents. Tools like Tesseract OCR or libraries like Pytesseract in Python can be used for this purpose.

Once the documents are converted into text, natural language processing (NLP) techniques can be applied to extract relevant information, such as key terms, clauses, or dates. Libraries like NLTK or spaCy in Python can be used for NLP tasks such as tokenization, part-of-speech tagging, and named entity recognition.

Additionally, it may be helpful to develop custom parsers or scripts tailored to the specific document formats encountered in the context of employment contracts and loan agreements. These parsers can help extract structured data from unstructured or semi-structured documents, facilitating further analysis and processing.

Overall, handling diverse document formats requires a combination of OCR technology, NLP techniques, and custom parsing solutions to effectively extract and analyze information from employment contracts and loan agreements.

### 3. Semantic Understanding and Inference.

Semantic understanding and inference are crucial aspects of processing employment contracts and loan agreements, as they involve interpreting the meaning and implications of the language used in these documents. Semantic understanding goes beyond simple text extraction and involves understanding the context, relationships,
and implications of the information presented.

One approach to semantic understanding and inference is to use advanced NLP techniques such as semantic role labeling, coreference resolution, and semantic
parsing. These techniques help identify the roles of entities mentioned in the text, resolve references to these entities, and parse the text into a structured representation that can
be used for inference.

For example, in an employment contract, semantic understanding can help identify the roles of the employer and employee, the rights and obligations of each party, and the conditions under which the contract can be terminated. In a loan agreement, semantic understanding can help identify the terms of the loan, the repayment schedule, and the consequences of defaulting on the loan.

Semantic inference involves drawing logical conclusions based on the information presented in the document. For example, inferring that a certain clause in an employment contract implies a specific obligation on the part of the employer or inferring the total repayment amount based on the terms outlined in a loan agreement.

Overall, semantic understanding and inference are essential for extracting meaning and making informed decisions based on the content of employment contracts and loan

agreements. By leveraging advanced NLP techniques, organizations can improve the accuracy and efficiency of document processing workflows.

### 4. Interpreting Unstructured data

Interpreting unstructured data, such as that found in employment contracts and loan agreements, requires a combination of techniques to extract meaningful information from the text. Unstructured data lacks a predefined data model or format, making it challenging to analyze using traditional methods. However, with the right approach, valuable insights can be gained from unstructured text.

One approach to interpreting unstructured data is to use natural language processing (NLP) techniques to extract relevant information from the text. This involves tasks such as tokenization, part-of-speech tagging, named entity recognition, and sentiment analysis. These techniques help break down the text into smaller, more manageable units and extract key information such as names, dates, amounts, and clauses.

Another approach is to use machine learning algorithms to analyze the text and extract patterns or trends. This can be done through techniques such as topic modeling, where the algorithm identifies the main topics or themes in the text, or classification, where the algorithm categorizes the text into predefined classes (e.g., clauses related to payment terms, termination clauses, etc.).

Additionally, domain-specific knowledge and expertise are crucial for interpreting unstructured data in the context of employment contracts and loan agreements. Understanding the legal and financial implications of the text is essential for making accurate interpretations and decisions based on the extracted information.

Overall, interpreting unstructured data requires a combination of NLP techniques, machine learning algorithms, and domain expertise. By leveraging these approaches, organizations can extract valuable insights from unstructured text and make more informed decisions based on the information.

## 2.2 Gap Resolution:

To resolve the gap caused by the lack of contextual understanding in employment contracts and loan agreements, several strategies, can be employed:

**Cultural and Contextual Understanding:**

- Explore models that can better understand and adapt to the cultural and contextual nuances within legal documents. Legal language can vary significantly across jurisdictions and cultures, impacting the interpretation of laws.

**Collaborative Decision-Making Models:**

- Investigate models that facilitate collaborative decision-making among legal professionals. This includes systems that support consensus-building and collective analysis of legal cases.

**Experiential and Precedent Learning:**

- Explore ways to incorporate experiential learning into models by leveraging the historical decisions and experiences of legal professionals. Models could learn from past case outcomes and adapt based on evolving legal practices.

**Handling Diverse Document Formats:**

- Implement adaptive models using ensemble techniques, employing a mix of architectures (CNNs, LSTMs, transformers) for feature extraction from various document layouts.

**Interpreting Unstructured Data:**

- Incorporate hybrid models combining rule-based systems with machine learning approaches. Use Named Entity Recognition (NER), entity linking, or rule-based parsers to handle ambiguous or unstructured information.

**User-Centric Customization:**

- Develop models that allow users to customize and tailor the system to user’s specific needs. This could involve personalized preferences, filtering criteria, or the ability to adapt the system’s behaviour to individual user workflows.

# 3. Problem Definition:

The problem definition for this study involves the lack of contextual understanding and a streamlined approach towards the legal issues an user can face.

## 3.1 Problem Statement:

There is an arbitrary need to develop an AI & ML based legal assistant that solves user’s legal related queries based on the scanned copy of user’s uploaded legal contracts, that also generates legal opinions and provides a community-based network for users to interact with experienced legal professionals.

## 3.2 Objective:

The primary objective of this project is to create effective methodology for creating an efficient and accurate user centric platform for legal issues. The aim is to achieve this by overcoming the existing gaps in the working of similar models by enhancing cultural and contextual understanding, providing a collaborative decision-making platform with experienced lawyers, introducing the ability to handle diverse document formats which can also interpret unstructured data.

# 4. Design of the Proposed Solution:

The user will be provided with two options after the user installs the application and enters the login credentials as shown in Figure 4.1:

- Option 1: Connect to lawyer – This option will allow the user to choose from a community-based network of experienced lawyers, the user can choose a lawyer according to his/her preferences after browsing through lawyers' profiles which will include area of expertise, years of experience and other key details. This option will facilitate messaging between user and lawyer.
- Option 2: Connect to legal assistant – This option will allow the user to interact with a legal chatbot wherein the user will first be asked to scan and/or upload the copy of a legal document. After uploading the document, the user will enter the legal query related to the document. The chat bot will generate an appropriate response based on the query. The user can further ask more questions related to the document until he/she is completely satisfied.

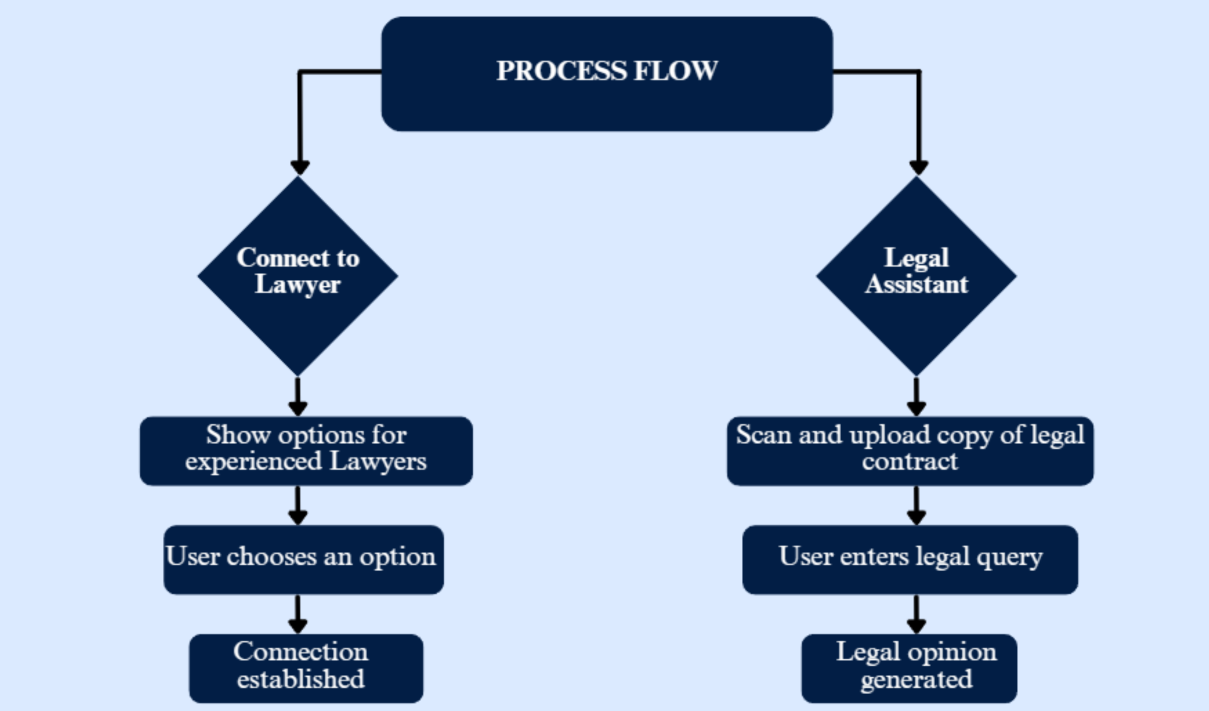


**Fig 4.1 Process flow diagram for Legal Assistance**

If the user chooses option 2, the model of the proposed solution is as follows:

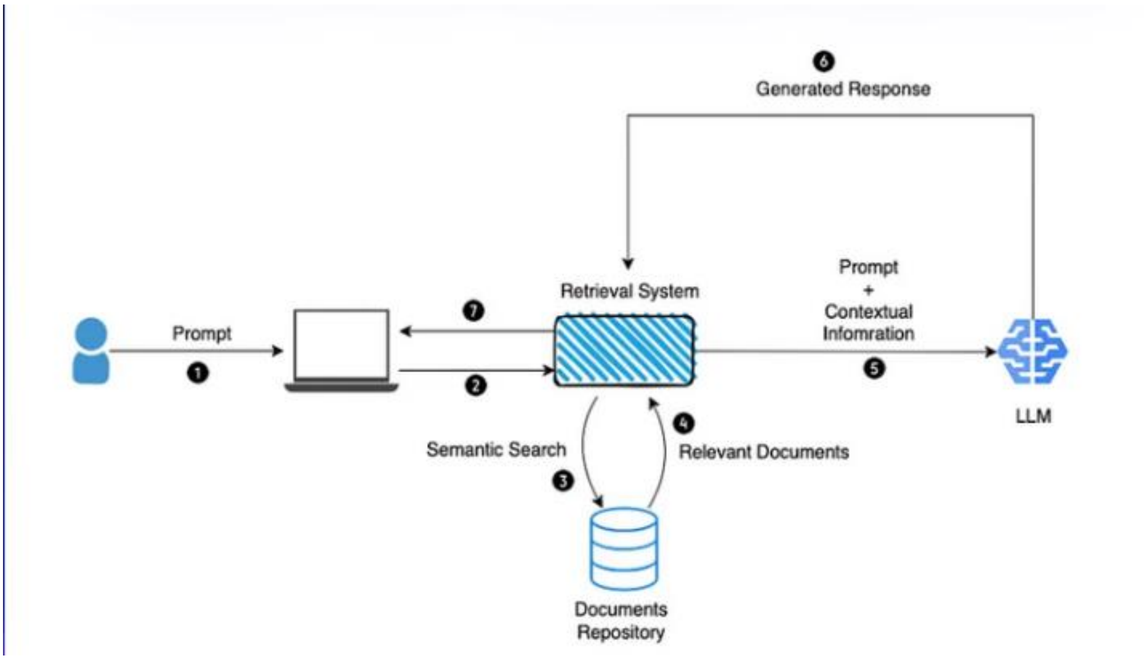


**Fig 4.2 Sequential Methodologies for Legal Response Generation**

The proposed solution for this legal assistant system shown in Figure 4.2 has six prominent steps viz. document preprocessing, user query input, retrieval stage, generation stage, evaluation stage and presentation to user.

Step 1: Document Preprocessing
The process begins with the user providing a scanned legal document to the system. This document, typically in image format, undergoes preprocessing to extract text content. This step ensures that the document is in a format suitable for further analysis and query processing.

Step 2: User Query Input
Following document preprocessing, the user enters a legal query related to the content of the scanned document. This query could be a specific question, request for information, or clarification on a particular aspect of the legal document. The system then processes this query to understand the user’s information needs.

Step 3: Retrieval Stage
In the retrieval stage, the system retrieves relevant documents or passages from a database or corpus based on the user’s legal query. This retrieval process employs various techniques, including keyword matching, TF-IDF (Term Frequency-Inverse Document Frequency), or advanced methods such as pre-trained language models for semantic retrieval. The goal is to

identify documents or sections of documents containing information pertinent to the user’s query.

Step 4: Generation Stage
Once relevant documents are retrieved, the system generates a response to the user’s legal query. This response is created using a Retrieval-Augmented Generation (RAG) model, which combines retrieval and generation capabilities. The RAG model leverages the retrieved documents as context to generate a response that is both coherent and relevant to the user’s query. Natural language generation techniques are utilized to produce human-like responses that convey the necessary information effectively.

Step 5: Evaluation Stage
Following response generation, the system evaluates the quality and relevance of the generated response. This evaluation process involves computing various metrics, such as BLEU (Bilingual Evaluation Understudy), Cosine Similarity and domain-specific metrics like the Legal Document Relevance Score (LDRS). This formula computes a relevance score for each document in the context of a given legal query. It combines cosine similarity between the query and each document's content with TF-IDF weighting to emphasize terms important in legal context.

Step 6: Presentation to User
Finally, the generated response is presented to the user via a user interface, such as a web application or chatbot interface. The user can review the response and interact further, asking follow-up questions, requesting additional information, or seeking clarification on specific points addressed in the response. This interactive presentation enables effective communication between the user and the legal assistant system, facilitating a seamless exchange of information.

Hence this process outlines the key steps involved in the proposed solution, from document preprocessing to response presentation, highlighting the system’s ability to assist users in accessing and understanding legal information effectively.

## 4.1 Novelty:

This study ensures that the gaps in the existing works of the legal assistance using AI are improved in the proposed solution.

### 1.Introduction to the Community-Based Legal Advice Platform

This study introduces a community-based network where users can connect with experienced legal professionals for legal advice. This platform aims to bridge the gap between individuals seeking legal guidance and qualified professionals, facilitating access to reliable legal assistance in diverse areas of law.

### 2. Handling Diverse Document Formats

One of the unique features of this platform is its ability to handle diverse document formats, including PDFs, JPEGs, PNGs, and others. Traditional legal documents are often stored in various formats, presenting a challenge for users seeking assistance with document analysis and interpretation. By accommodating multiple file types, this platform ensures accessibility and convenience for users, allowing them to upload and analyze documents seamlessly.

### 3. Utilization of Retrieval-Augmented Generation (RAG) Models

This platform leverages Retrieval-Augmented Generation (RAG) models to enhance its capabilities in understanding the semantics of legal documents and retrieving relevant passages. RAG models combine the strengths of retrieval and generation techniques, enabling the system to comprehend complex legal texts and extract pertinent information effectively. By utilizing RAG models, this platform can provide users with accurate and contextually relevant insights from legal documents, empowering them to make informed decisions.

### 4. Interactive Legal Analysis Sessions

A key innovation of this platform is the implementation of interactive legal analysis sessions, where users can engage in real-time discussions with legal professionals and ask follow-up questions to clarify doubts or seek further information. This interactive feature facilitates dynamic and collaborative interactions between users and experts, fostering a deeper understanding of legal concepts and issues. Moreover, these sessions enable the platform to adapt and improve its performance over time through user interactions, refining its capabilities and enhancing the quality of legal advice provided.

# 5 Result Analysis:

Using different legal queries and legal documents as training data, this model was further examined on various evaluation metrics as mentioned in Table 5.1:

**Table 5.1 Rating of Evaluation Matrix for customer's given query**

| QUERY | BLEU SCORE | COSINE SIMILARITY | LEGAL DOCUMENT RELEVANCE SCORE-1 | LEGAL DOCUMENT RELEVANCE SCORE-2 |
|---|---|---|---|---|
| What are the circumstances under which the Company can terminate the Employee's employment without Cause? | 0.550324 | 0.8584 | 3.5344 | 2.245 |
| What are the remuneration and benefits that the Employee is entitled to? | 0.4762 | 0.9021 | 3.4678 | 4.3450 |
| What are the consequences of a breach of the covenants contained in the Agreement? | 0.012 | 0.4242 | 3.4678 | 2.1506 |
| Are there any penalty or compensation clauses? | 0.1218 | 0.46815 | 3.3830 | 2.9029 |

LEGAL DOCUMENT RELEVANCE SCORE-1 computes the score by comparing the response generated by this model and the document uploaded by the user initially.

LEGAL DOCUMENT RELEVANCE SCORE-2 computes the score by comparing the response generated by this model and the response generated by generic LLMs (that do not take any scanned/uploaded documents into consideration)

## 5.1 Mathematics Involved:

**BLEU Score:**
BLEU score measures the similarity between generated responses and reference (ground truth) responses based on n-gram overlap. Higher BLEU score indicates better similarity as mentioned in Equation 1.

$$\text{BLEU=BP}\times\text{exp}(\sum N^{-1}(\text{logpn})) \tag{1}$$

- BP is the brevity penalty, which penalizes overly short translations to address the problem of length discrepancies between candidate and reference translations.
- $pn$ is the precision of n-grams in the candidate translation compared to the reference translations.
- $N$ is the maximum n-gram order considered in the computation

**Cosine Similarity**:

$$\text{Similarity}(a,b)=\frac{a.b}{\|\text{a}\|\|\text{b}\|} \tag{2}$$

As mentioned in Equation 2 calculates the cosine similarity between two vectors $a$ and $b$. This is often used in the retrieval component of RAG models to measure the similarity between the query and retrieved documents.

**TF-IDF Score:**
TF-IDF score is a statistical measure used to evaluate the importance of a term in a document relative to a collection of documents (corpus). It's often used in information retrieval and text mining tasks, including the retrieval component of RAG models as mentioned in Equation 3.

- Formula for TF: TF($t$,$d$)=number of times term $t$ appears in document $d$ divided by total number of terms in document $d$
- Formula for IDF: IDF($t$,$D$)=log$N$/|{$d$∈$D$:$t$∈$d$}|
- Formula for TF-IDF:

$$\text{TF-IDF}(t,d,D)=\text{TF}(t,d)\times\text{IDF}(t,D)\text{TF-IDF(t,d,D)=TF(t,d)×IDF(t,D)} \tag{3}$$

**Legal Document Relevance Score (LDRS):**
Legal Document Relevance Score computes a relevance score for each document $D$ in the context of a given legal query $Q$. It combines cosine similarity between the query and each document's content with TF-IDF weighting to emphasize terms important in legal contexts as mentioned in Equation 4.

- Formula:

$$\text{LDRS}(Q,D)=\sum[\text{cosine similarity}(Q,di)]\times[\text{TF-IDF}(qi,di,D)] \tag{4}$$

## 6. Conclusion:

The introduction of a community-based legal advice platform in this study signifies a significant advancement in addressing the critical need for accessible and dependable legal assistance in contemporary legal landscapes. By harnessing advanced natural language processing techniques, notably the Retrieval-Augmented Generation (RAG) models, the platform facilitates personalized guidance from seasoned legal professionals across a diverse array of legal domains.

This study underscores the inherent advantages of RAG models over existing methods. RAG models excel in comprehending the intricate semantics of legal documents, leveraging both retrieval and generation capabilities to produce highly contextually relevant responses. Unlike conventional language models, RAG models integrate retrieved document contexts, thereby enhancing the accuracy and relevance of generated responses. This integration ensures that responses align closely with the specific legal context presented in the documents, leading to more informed and actionable advice for users.

In addition to conventional evaluation metrics such as the BLEU score, this study introduces the Legal Document Relevance score as a novel performance metric tailored explicitly to the legal domain. This metric provides a quantitative measure of the relevance of generated responses to the content of legal documents, offering a more efficient and accurate assessment of performance compared to generic metrics. By focusing on the alignment between responses and legal document contexts, the Legal Document Relevance score enhances the platform's ability to deliver precise and insightful guidance to users.

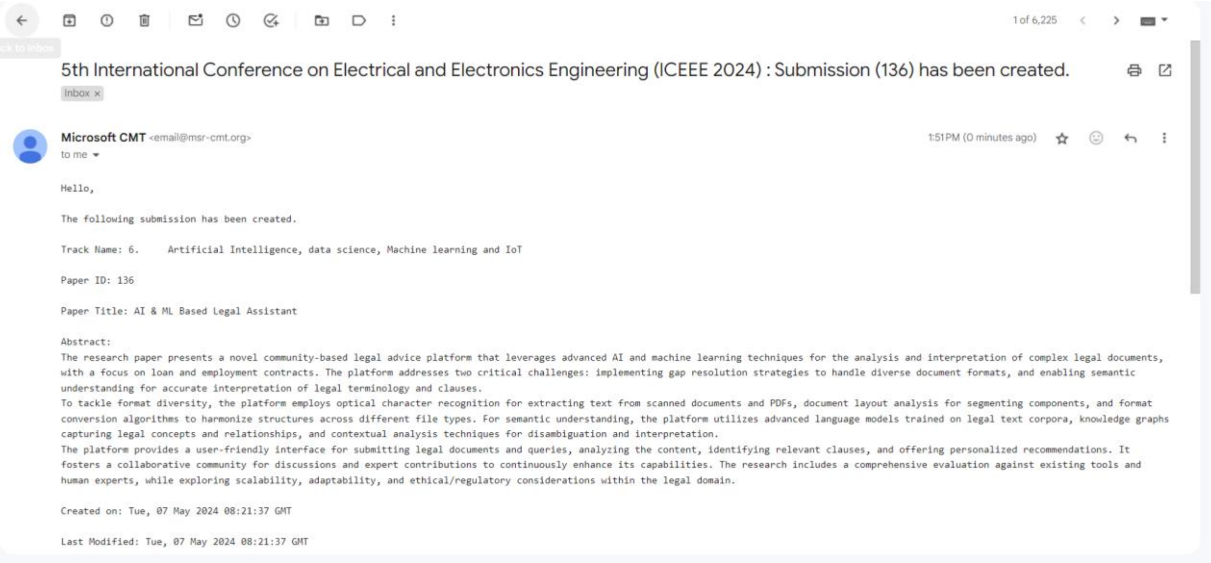
Furthermore, the platform's capability to handle diverse document formats, including scanned PDFs, JPEGs, and PNGs, significantly enhances its accuracy compared to models relying solely on preprocessed text. This comprehensive approach ensures that the platform delivers superior performance, enabling users to access reliable legal assistance with utmost efficiency and accuracy.

# 7. Future Scope:

Increase the accuracy of the model by increasing the size of the Corpus in the knowledge base.

Contacting legal professionals to sign them up for this chat feature wherein users can communicate with them to solve legal queries of the model.

## Publication:



5th International Conference on Electrical and Electronics Engineering (ICEEE 2024) : Submission (136) has been created.

Microsoft CMT <email@msr-cmt.org>

Hello,

The following submission has been created.

Track Name: 6.    Artificial Intelligence, data science, Machine learning and IoT

Paper ID: 136

Paper Title: AI & ML Based Legal Assistant

Abstract:
The research paper presents a novel community-based legal advice platform that leverages advanced AI and machine learning techniques for the analysis and interpretation of complex legal documents, with a focus on loan and employment contracts. The platform addresses two critical challenges: implementing gap resolution strategies to handle diverse document formats, and enabling semantic understanding for accurate interpretation of legal terminology and clauses.

To tackle format diversity, the platform employs optical character recognition for extracting text from scanned documents and PDFs, document layout analysis for segmenting components, and format conversion algorithms to harmonize structures across different file types. For semantic understanding, the platform utilizes advanced language models trained on legal text corpora, knowledge graphs capturing legal concepts and relationships, and contextual analysis techniques for disambiguation and interpretation.

The platform provides a user-friendly interface for submitting legal documents and queries, analyzing the content, identifying relevant clauses, and offering personalized recommendations. It fosters a collaborative community for discussions and expert contributions to continuously enhance its capabilities. The research includes a comprehensive evaluation against existing tools and human experts, while exploring scalability, adaptability, and ethical/regulatory considerations within the legal domain.

Created on: Tue, 07 May 2024 08:21:37 GMT

Last Modified: Tue, 07 May 2024 08:21:37 GMT

# REFERENCES

[1]Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, et al., 2016. Predicting judicial decisions of the European court of human rights: a natural language processing perspective. PeerJ Comput Sci, 2:e93.

[2]https://doi.org/10.7717/peerj-cs.93 Arditi D, Oksay FE, Tokdemir OB, 1998. Predicting the outcome of construction litigation using neural networks. Comput-Aided Civ Infrastruct Eng, 13(2):75-81. https://doi.org/10.1111/0885-9507.00087 Ashley KD, Brüninghaus S, 2009.

[3]Automatically classifying case texts and predicting outcomes. Artif Intell Law, 17(2):125-165. https://doi.org/10.1007/s10506-009-9077-9 Chao

[4]WH, Jiang X, Luo ZC, et al., 2019. Interpretable charge prediction for criminal cases with dynamic rationale attention. J Artif Intell Res, 66:743-764. https://doi.org/10.1613/jair.1.11377 Dahbur K, Muscarello T, 2003.

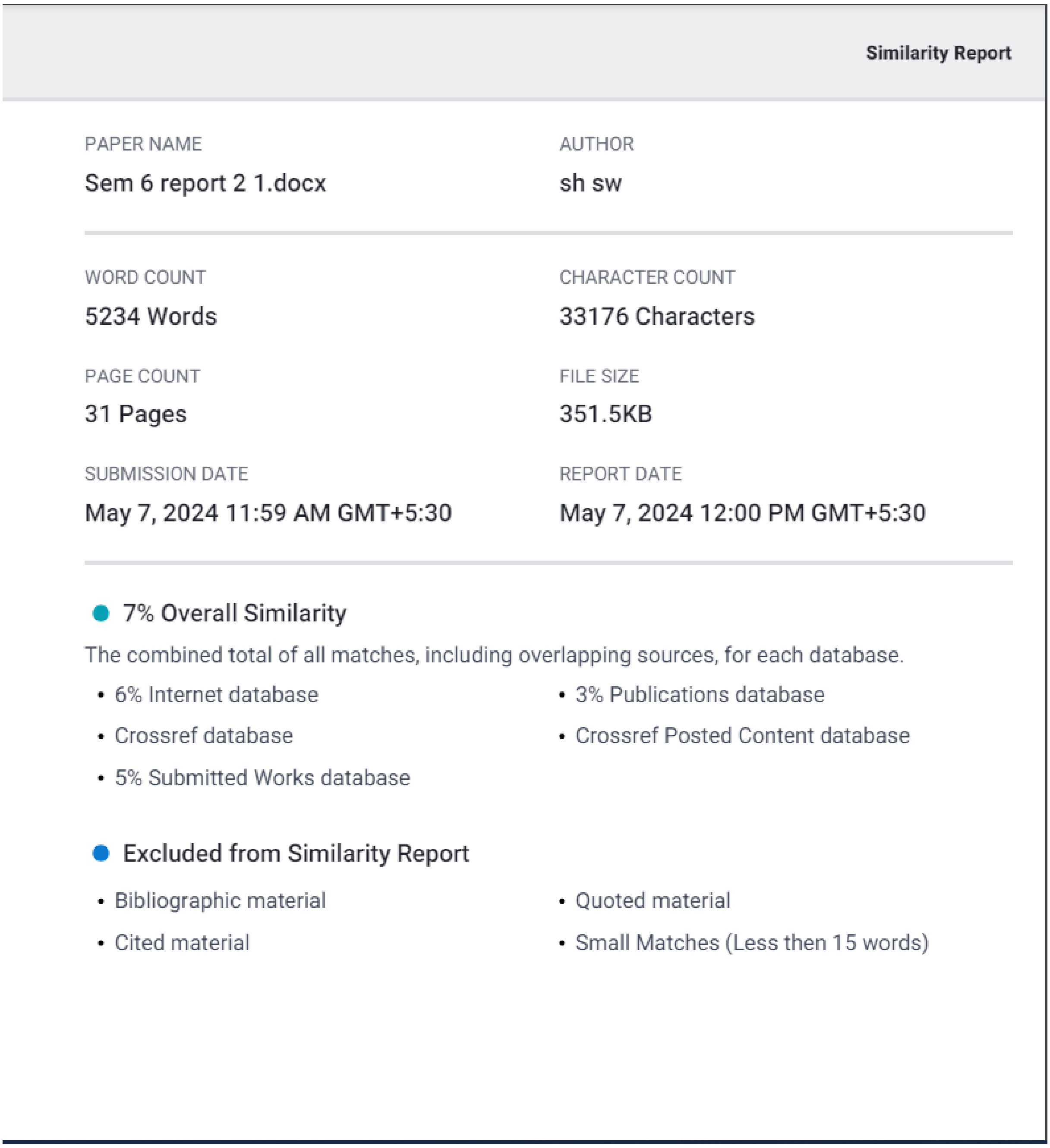
[5]Classification system for serial criminal patterns. Artif Intell Law, 11(4):251- 269.

[6] Duan XY, Zhang YT, Yuan L, et al., 2019. Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning.

[7] Proc 28th ACM Int Conf on Information and Knowledge Management, p.1361-1370. https://doi.org/10.1145/3357384.3357940 Elnaggar A, Otto

[8] R, Matthes F, 2018. Deep learning for named-entity linking with transfer learning for legal documents. Proc Artificial Intelligence and Cloud Computing Conf, p.23-28. https://doi.org/10.1145/3299819.3299846 Gerani S, Mehdad Y, Carenini G, et al., 2014.

[9] Agnoloni T, Bacci L, Francesconi E, Spinosa P, Tiscornia D, Montemagni S, Venturi G (2007) Building an ontological support for multilingual legislative drafting In: Proceedings of the Jurix Conference.https://doi.org/10.1023/B:ARTI.0000045994.96685.21



**Similarity Report**

PAPER NAME

Sem 6 report 2 1.docx

AUTHOR

sh sw

WORD COUNT

5234 Words

CHARACTER COUNT

33176 Characters

PAGE COUNT

31 Pages

FILE SIZE

351.5KB

SUBMISSION DATE

May 7, 2024 11:59 AM GMT+5:30

REPORT DATE

May 7, 2024 12:00 PM GMT+5:30

● 7% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 6% Internet database
- 3% Publications database
- Crossref database
- Crossref Posted Content database
- 5% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material
- Quoted material
- Cited material
- Small Matches (Less then 15 words)