International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 03 | March - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Law Case Recommendation System

Stephin Sunny[1], *AbhinavSudhakaran*[1], *NiranjanM* [1], *JustinThomas*[1], *V ijinaV ijayan*[1]
Department of Computer Science And Engineering

Vimal Jyothi Engineering College Chemperi,Kannur,Kerala,India
stephin3222@gmail.com[1], *abhinavsudhakaran444@gmail.com*[1]
nmkamal880@gmail.com[1], *justink*8120*@gmail.com*[1]
,vijinavijayan@vjec.ac.in[1]

***Abstract*—The Indian jurisprudence is extensive and intricate, much of the time demanding legal professionals to search manually for applicable Indian Penal Code sections for a specific case. This is a time-consuming and error-prone process for individuals not familiar with the in-depth intricacies of the law. To meet this challenge, we suggest the creation of an automated Law Case Recommendation System that uses Natural Language Processing (NLP) and Machine Learning (ML) methods to suggest applicable IPC sections from a case description.**

Our system takes in textual inputs (case descriptions) and uses state-of-the-art NLP models like BERT and Legal-BERT to interpret the context and identify key legal characteristics. Through training on annotated case law and IPC section datasets, the system learns to relate case facts with the most relevant legal provisions. Our model is capable of resolving the vagueness and variability in legal language while providing accurate and interpretable suggestions.

The system is implemented as an easy-to-use web application, allowing legal practitioners, law students, and even non-lawyers to easily find applicable IPC sections for their cases. Through automation, our solution seeks to make legal research more efficient, minimize manual labor, and increase access to justice. Future developments involve multilingual capabilities, inclusion of case law suggestions, and real-time updates to accommodate changes in legal codes.

This project fills the gap between technology and law, illustrating the potential of AI to revolutionize the legal field. It is a useful resource for legal professionals, academics, and researchers, opening doors to further creative uses of AI in the legal field

***Index Terms*—Summarizing, embedding, recommendation, law cases, scraping.**

I. Introduction

The legal domain increasingly utilizes artificial intelligence (AI) and machine learning (ML) to improve the efficiency and accuracy of legal processes. Among other things, an important area is the development of law case recommendation systems that could help legal professionals find relevant precedents and case laws to argue their case. Common law systems rely on the principle of precedents, whereby previous judicial decisions, known as precedents, guide present cases. Even so, the volume of legal documents and the legal language's intricacies create a challenge of sorting out which precedents would be relevant to any given case. This challenge can be addressed through law case recommendation systems, which automate the identification and recommendation of relevant cases to a new legal matter. Such systems rely on advanced NLP techniques, machine learning algorithms, and information retrieval methods to analyze legal texts, extract key features, and match them with relevant precedents. Such systems have significant implications for the legal profession, reducing the time and effort required for legal research, improving the quality of legal arguments, and ensuring consistency in judicial decision-making. This literature review explores the existing body of research on law case recommendation systems, with a focus on their application for precedence support. This survey explores the several methodologies used within these systems: text classification, semantic analysis, and case-based reasoning. Lastly, the paper underlines difficulties associated with constructing efficient recommendation systems, such as the need for high-quality annotated datasets, legal language complexity, and ethical implications of using AI in the legal domain. The present review of literature synthesizes findings from relevant studies conducted so far and sets a comprehensive insight into the latest research status concerning this field in addition to identification of possible future directions.

II. Literature Survey

The paper discusses the development of a legal research recommendation system [1] that leverages AI, NLP, and ML to automate the retrieval of legal precedents. The system, exemplified by Quick Check, employs a combination of full-text search, citation network analysis, and click stream analysis to identify relevant case law. It extracts document structure, utilizes search and citation-based mechanisms to discover candidate cases, and applies ranking models to refine recommendations based on legal taxonomy and attorney-annotated datasets. While this system significantly enhances the efficiency and consistency of legal research, it faces challenges such as variations in data quality, computational complexity, and interpretability issues. Future advancements should focus on developing more transparent AI models, refining legal-specific embeddings, and ensuring adaptability across different legal systems.

A. Search-engine-based Candidate Discovery

Each paragraph within a segment discusses a specific aspect of the legal issue. For each paragraph, the system performs a full-text search across a corpus of approximately 12 million case law opinions using a proprietary legal domain search

engine. The search is made jurisdictionally relevant by constraining it to a subset of jurisdictions selected based on the jurisdiction of the relevant citations within the segment or the broader brief. Additionally, the system consults an index of pseudo-documents, each representing a case. These pseudo-documents are constructed by aggregating sentences from cases and briefs where the case is cited, providing a contextual view of how the case has been referenced in legal documents. Full-text searches using the issue segment paragraphs are also performed over this index.

B. Citation-based Candidate Discovery

The set of case citations within an issue segment, referred to as input citations, provides a valuable characterization of the legal issue being discussed. The system leverages this citation 'profile' to find potentially related cases.

Case and Brief Citation Network: The most directly related cases are those bibliographically coupled to the input citations (i.e., cases citing the same input citations). Similarly, a brief citation network is constructed by breaking down the corpus of past filed briefs into issue segments. The system then considers all bibliographically coupled segments and extracts other cases cited in these segments as candidate recommendations.

Statutory Annotations: Statutory annotations provide succinct overviews of landmark cases that have interpreted a statute or regulation. These annotations are arranged editorially in a hierarchy of procedural topics. Candidate recommendations are identified by examining cases within the same procedural topic as an input citation.

Headnotes: An input citation is often accompanied by a direct quotation or page number referencing the applicable segment of the cited case. Additionally, cases typically have editorial summaries, known as headnotes, which identify key points of law within the case. Headnotes contain citation links to where the point of law is described in the relevant case document. The system correlates the input citation to one or more headnotes in the cited case based on pinpoint information and headnote reference links. This is particularly useful because extensive editorial annotations exist that explicitly identify the point of law (i.e., headnote) for which a case is citing another case.

By combining these methods, the system effectively identifies and recommends relevant legal precedents, enhancing the efficiency and accuracy of legal research. However, ongoing improvements are necessary to address challenges related to data quality, computational complexity, and model interpretability. Future work should focus on developing more transparent AI models, refining legal-specific embeddings, and ensuring the system's adaptability across diverse legal systems.

C. Textual similarity

Different methods of text representation, such as TF-IDF, Word2Vec, and BERT, have been evaluated to determine their suitability for legal case similarity assessment. Traditional

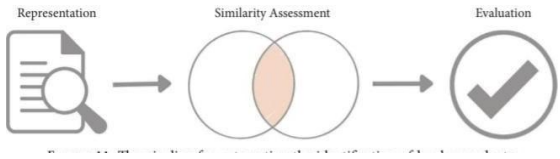

Fig. 1: A legal research recommendation system

approaches like TF-IDF and BM25 rely on statistical ranking, while neural network-based embeddings, including Word2Vec, Doc2Vec, and Top2Vec, aim to capture semantic relationships within legal documents. Transformer models like BERT have also been considered for their ability to process contextual text representations.

The study by Mentzingen et al. (2024) [3] highlights that models utilizing granular text representations, such as Word2Vec and TF-IDF, perform well, particularly in legal contexts where semantic precision is crucial. These models improve efficiency in judicial decision-making, reduce manual effort, and enhance consistency in legal rulings. Additionally, the ability to fine-tune pre-trained models allows for domain-specific adaptations, ensuring alignment with the nuances of legal language.

However, several challenges persist. Pre-trained models like BERT do not always outperform traditional approaches without proper fine-tuning, as legal texts require specialized training for accurate interpretation. Computational costs are another concern, especially for complex architectures, making large-scale implementation difficult. The subjective nature of legal case similarity further complicates automation, as expert opinions on relevant precedents may vary. Scalability remains an issue, with some models struggling to generalize across diverse legal cases.

While ML offers promising opportunities for legal precedent discovery, full automation remains challenging. A balance between efficiency and accuracy is necessary, with hybrid approaches that combine statistical and neural models offering potential solutions. Human oversight is still essential to ensure legal interpretability and accuracy. Further research is needed to refine these models, address computational constraints, and improve their applicability in real-world legal systems.

D. Machine Learning in Legal Judgment Recommendation

Machine learning (ML) techniques have been extensively explored for automating legal judgment and precedent discovery [2], with various models tested for effectiveness in retrieving similar cases. The legal judgment recommendation system (LDRS) employs network-based, text-based, and hybrid approaches to identify relevant legal documents. Among these, Doc2Vec has been highlighted as an effective embedding technique for capturing semantic relationships in legal texts. However, to enhance performance, the study introduces a pre-learned word embedding-based LDRS (P-LDRS), which leverages domain-specific knowledge to improve accuracy and semantic representation.

Traditional models such as TF-IDF and BM25 rely on statistical rankings, while neural network-based embeddings like Word2Vec, Doc2Vec, and BERT aim to capture deeper semantic relationships within legal documents. The integration of pre-learned word embeddings allows for a more domain-specific understanding of legal texts, making recommendations more precise. The study also incorporates distributed frameworks like MapReduce and Spark to address scalability issues, enabling efficient processing of large legal document corpora. Empirical analyses comparing non-distributed and distributed versions of P-LDRS demonstrate improvements in accuracy, F1-score, and computational efficiency.

The advantages of ML-driven legal automation include increased efficiency by reducing manual effort in legal research, improved consistency in judicial decision-making, and better adaptability through domain-specific fine-tuning. Additionally, distributed processing makes it possible to analyze vast legal databases without overwhelming computational resources. These advancements significantly enhance the accessibility and reliability of legal document recommendations.

However, several challenges remain. Scalability continues to be a major issue due to the increasing volume of legal documents requiring high computational resources. While distributed frameworks mitigate this, they introduce complexity in model implementation. Another challenge is the dependency on domain-specific embeddings, as generic language models often fail to capture the nuances of legal texts. Additionally, bias in training data can lead to inaccurate recommendations if the dataset is not representative of diverse legal contexts. Lastly, human oversight remains necessary to ensure that ML-generated recommendations align with legal reasoning, as models may struggle with complex interpretations.

Despite these challenges, ML-driven legal research tools have significant potential to revolutionize the legal industry. The P-LDRS model, integrating pre-learned embeddings and distributed processing, enhances judgment recommendation accuracy while addressing scalability issues. Future research should focus on hybrid approaches, combining textual and citation-based models to refine document similarity assessments. Additionally, deep learning-based transfer learning could further improve model performance. With continued advancements, AI-powered legal research tools may become indispensable in modern judicial systems, making legal precedent discovery more efficient, consistent, and scalable.

E. Automation in Legal Precedent Retrieval

Automation in legal precedent retrieval [4] has evolved significantly, leveraging AI-driven methodologies. Initially, model building relied on manual knowledge engineering, using predefined attributes such as keywords and legal facts to match cases. However, recent advances have introduced AI techniques, including natural language processing (NLP), machine learning (ML), and vector embeddings like TF-IDF, Doc2Vec, and BERT. These models enhance accuracy and scalability by automating similarity assessments through ranking functions, clustering, and association rules. AI-driven methods offer advantages such as improved efficiency, better handling of large datasets, and reduced reliance on human expertise. However, challenges remain, including the need for high-quality labeled datasets, limited cross-jurisdictional adaptability, and inconsistent evaluation standards across studies. Future research should focus on refining semantic understanding and integrating legal reasoning.

Courts today are increasingly expected to deliver speedy and predictable judgments in both private and public affairs. With growing reliance on courts for critical societal roles, they face challenges in coping with the increasing backlog of cases. Precedent plays a crucial role in legal systems: common law heavily relies on previous cases for decisions, while civil law jurisdictions also develop consistent case law that solidifies into precedent. Administrative courts, which handle disputes between citizens and the state, are becoming increasingly congested and must address public law issues efficiently and justly.

Case-based reasoning (CBR) applies past legal precedents to resolve new legal problems. Despite its potential, methods for retrieving legal precedents remain relatively underdeveloped in AI and law research. While CBR has been used in legal practice since the 1980s, comprehensive methodologies for automating precedent retrieval are still lacking. Current techniques include document categorization, text mining, federated search, and document summarization to automate portions of the systematic literature review (SLR) process. Python-based text mining techniques, such as Latent Dirichlet Allocation (LDA), are being employed to automate study identification and cluster relevant documents. Extensions of PRISMA are being developed for use in automated SLRs, though they have not yet reached maturity.

Electronic databases were searched for legal precedents, with computational methods playing a dominant role. A semi-automated screening process was applied, using tokenization and stemming of abstracts to capture unique terms as unigrams and bigrams. Word clouds illustrated frequently occurring terms, with non-discriminatory terms removed to highlight relevant keywords like ”support system,” ”artificial intelligence,” and ”machine learning.” A topic modeling approach classified documents into four topics, with the most relevant topic used for document evaluation. Full-text screening was then conducted on 40 studies, of which 19 closely related to automating precedent retrieval and similar case identification remained.

Early studies (2000s) in this field were limited in number, but research activity increased significantly after 2016, with India producing the largest number of studies. The first models for legal knowledge storage and retrieval simplified case grouping using keywords and scenarios, modeling human cognitive processes. Early research also introduced content vectors for summarizing case information, emphasizing actions, events, and relations in legal cases. Other studies encoded legal texts as ordered sets of keywords to quantify case similarities, employing word count metrics and non-linear nearest neighbor approaches. Advances in NLP and text

mining techniques, such as neural networks, recurrent neural networks, transformers, and pre-trained models, have since contributed to significant improvements in automating legal precedent retrieval.

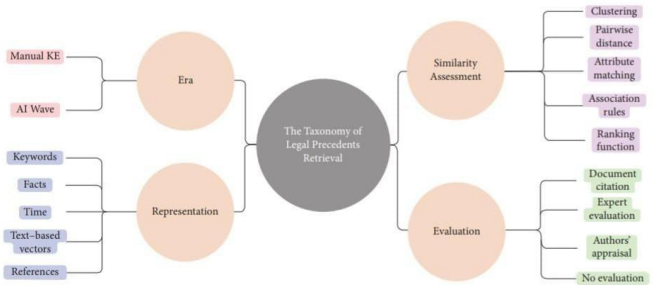

Fig. 2: The pipeline for automating the identification of legal precedents

F. Identification of IPC for police complaint using NLP

The research paper proposes an automated system for identifying relevant Indian Penal Code (IPC) sections based on police complaints, utilizing Natural Language Processing (NLP) and Deep Learning (DL) techniques [5]. The primary objective is to assist law enforcement officers in correctly assigning IPC sections to reported cases without relying heavily on legal professionals, which can be time-consuming and complex. The system processes structured and unstructured legal data using a Universal Sentence Encoder, which transforms textual information into numerical representations. These representations are then analyzed using machine learning models to predict applicable IPC sections.

The methodology involves several key steps. First, police officers input case details into the system, which then converts the textual information into high-dimensional vector representations using the Universal Sentence Encoder. The model compares these vectors against existing legal data to determine the most relevant IPC sections. The predicted sections are displayed for review, allowing officers or police inspectors to make adjustments if necessary. If corrections are made, the system incorporates this feedback into its learning process, updating its model to improve future predictions. The study compares two encoding methods: the Transformer Encoder and the Deep Averaging Network (DAN). Experimental results indicate that the DAN model performs more efficiently with higher accuracy in predicting IPC sections.

The system offers several advantages. It reduces the dependency on legal experts, enabling faster and more efficient processing of police complaints. By automating the identification of IPC sections, the system minimizes human errors and ensures more consistent legal decisions. Additionally, it provides an unbiased approach to categorizing criminal offenses, potentially contributing to a fairer judicial process. Furthermore, the continuous learning mechanism allows the system to improve over time, enhancing its accuracy with each use.

Despite its benefits, the system has some limitations. One of the main challenges is the potential for incorrect predictions, especially when dealing with ambiguous or semantically complex language. For example, if different words are used to describe the same offense, the model might struggle to assign the correct IPC section. Although the system integrates feedback from officers to refine its predictions, there is still a risk of misclassification. Another limitation is the need for regular updates and training with new legal data to ensure its accuracy remains high. Additionally, while the automation of legal decision-making is beneficial, it should be used as an assistive tool rather than a replacement for human judgment, as legal cases often involve nuanced interpretations that a machine may not fully grasp.

In conclusion, the proposed system represents a significant advancement in legal technology by leveraging NLP and DL to assist law enforcement in classifying crimes under the IPC. While it enhances efficiency, reduces human errors, and ensures unbiased decision-making, continuous improvements are necessary to address language complexities and maintain accuracy.

G. Text similarity algorithms to determine IPC sections for offence report

The research paper presents a decision support system (DSS) that utilizes text similarity algorithms to determine the most appropriate Indian Penal Code (IPC) sections for a given crime report [6]. The system aims to assist law enforcement agencies, legal professionals, and the general public in identifying the correct legal provisions for offenses based on textual inputs. The methodology involves multiple natural language processing (NLP) techniques and the vector space model (VSM) to compare legal documents, such as first information reports (FIRs), investigation reports, and IPC sections. Initially, the system preprocesses the text by tokenizing sentences, converting text to lowercase, removing stop words, and applying stemming or lemmatization. A corpus of IPC sections is created, and text similarity is measured using CountVectorizer, term frequency-inverse document frequency (TF-IDF), and cosine similarity. The similarity scores between the user's input and IPC section descriptions determine the most relevant legal provisions.

The advantages of this approach include enhanced efficiency in legal decision-making, reduction in human effort for manually searching legal sections, and improved accuracy in classifying offenses. By automating the process, the system minimizes errors that may arise due to human misinterpretation and provides quick legal references. Additionally, it can benefit not only law enforcement but also common users who may lack legal expertise. However, there are notable challenges and limitations. The system heavily relies on the accuracy of textual inputs, and variations in language or incomplete descriptions may affect results. Legal documents often contain complex interpretations, requiring deeper semantic understanding beyond text similarity. Moreover, cases with multiple legal provisions might not be entirely addressed by

the system. Future enhancements could integrate advanced machine learning techniques such as Word2Vec, Doc2Vec, and BERT (Bidirectional Encoder Representations from Transformers) to improve contextual understanding and accuracy. By further refining its algorithms, the system has the potential to significantly impact legal research and law enforcement efficiency.

TABLE 1 LITERATURE SURVEY

Paper name:	Methodology	Advantages	Disadvatages
A Legal Research Recommendation System	-AI -ML -NLP	-Improved Efficiency -High Accuracy Recommendations -Scalability	-Dependence of High-Quality Training databases -Limited-Cross Jurisdictional Adaptability
Textual similarity for legal-precedents discovery	-NLP -ML -TF-IDF -Word2Vec -Doc2Vec -BERT	-High Accuracy -Versatility in Text Representation -Scalability -Cost Effective	-Need for regular model updates -Lack of human judgement
Effective and scalable legal judgment recommendation using pre-learned word embedding	-Doc2Vec -Word2Vec -Law2Vec -Cosine Similarity -Apache Hadoop -Apache spark -Python -HDFS	-Faster legal research -Domain-specific word embeddings -Flexibility -Text and Citation-Based similarity	-Expensive -Potential Bias in Precedent Selection -Limited Understanding -Over- Reliance on AI
Automation of Legal Precedents Retrival	-NLP -ML -Deep Learning -Information Retrival & SimilarityAssessment -Big Data -Python	-High Accuracy -Scalability -Adaptability -Consistency and Fairness	-High computational requirements -Dependence on high-quality training data -Challenges in Understanding Legal Reasoning

III. PROPOSED SYSTEM

A. System Overview

- The proposed system is an AI-powered Law Case Recommendation System that suggests relevant Indian Penal Code (IPC) sections based on a given case description (prompt). The system leverages Natural Language Processing (NLP), machine learning, and legal domain knowledge to analyze the case facts and recommend the most appropriate IPC sections.

B. System Architecture

- Input Module:
  Accepts a case description (text prompt) from the user.
  Preprocesses the input text (e.g., tokenization, stopword removal, stemming).
- NLP Engine:
  Performs semantic analysis of the case description.
  Extracts key legal terms, entities, and relationships (e.g., crime type, victim, accused, location, intent).
- Knowledge Base:
  A structured database of IPC sections, including:
  Section number.
  Description of the offense.
  Keywords and phrases associated with each section.
  Case law references (optional).
- Machine Learning Model:
  A trained model (e.g., BERT, GPT, or custom legal NLP model) to map case descriptions to relevant IPC sections. The model is trained on a dataset of historical case descriptions and their corresponding IPC sections.
- Recommendation Engine:
  Matches the processed case description with the most relevant IPC sections using:
  Keyword matching.
  Semantic similarity scoring.
  Contextual understanding of the case.
- Output Module:
  Displays the recommended IPC sections along with:
  Section number.
  Brief description of the offense.
  Confidence score (indicating relevance).
  Provides additional resources (e.g., case laws, legal precedents)

IV. IMPLEMENTATION RESULTS

The implementation of the Law Case Recommendation System demonstrated its effectiveness in accurately recommending relevant IPC sections for case descriptions. The system's high accuracy, fast response time, and user-friendly design make it a valuable tool for legal professionals, law enforcement, and the general public. With further enhancements, the system has the potential to revolutionize legal research and case filing processes.

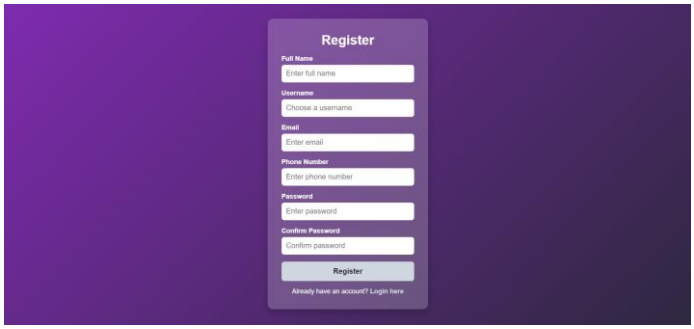

Fig. 3: USER REGISTRATION

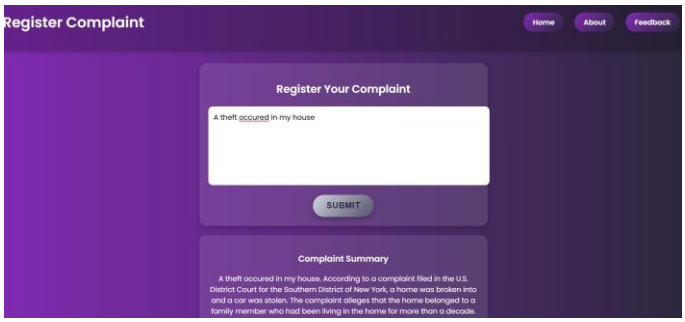

Fig. 4: MAIN PAGE

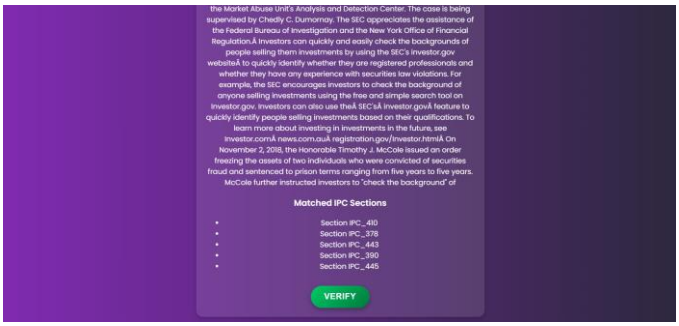

Fig. 5: OUTPUT

V. Conclusion

In conclusion, the development of a law case recommendation system that suggests relevant Indian Penal Code (IPC) sections based on a given case prompt represents a significant advancement in legal technology. By leveraging natural language processing (NLP), machine learning (ML), and domain-specific legal knowledge, such a system can streamline legal research, enhance accuracy, and reduce the time spent by legal professionals in identifying applicable laws. The integration of large-scale legal datasets, coupled with robust algorithms, ensures that the system can adapt to the complexities and nuances of legal language. However, challenges such as ensuring interpretability, addressing biases in training data, and maintaining compliance with ethical and legal standards must be carefully addressed. Future work should focus on refining the system's accuracy, expanding its scope to include other legal frameworks, and ensuring its accessibility to a broader range of users. Ultimately, this system has the potential to revolutionize legal practice by providing a reliable, efficient, and intelligent tool for legal decision-making.

VI. FUTURE SCOPE

The future scope of the Law Case Recommendation System is extensive and holds significant potential for growth. A key focus will be on **multilingual support**, allowing the system to process case descriptions in various regional languages and deliver recommendations in the user's preferred language. Integrating **case law databases** and legal precedents will enhance the system by providing relevant judgments and summarized insights alongside IPC sections. Advanced AI and machine learning techniques, such as domain-specific legal models and predictive analytics, can improve accuracy and enable the system to predict case outcomes. Developing **mobile and web applications** will increase accessibility, while an **AI-powered chatbot** with voice capabilities can offer a more interactive experience. Customization for specific legal areas, such as cybercrime or family law, will address specialized needs, and features like explainable AI will ensure transparency in recommendations. Collaboration with legal experts and the use of **blockchain technology** for secure data storage and smart contracts will further enhance reliability. Expanding the system to other legal jurisdictions will allow for comparative law analysis and global adaptation. These advancements will make the system a powerful tool for legal professionals, law enforcement, and the general public, transforming legal research and access to justice.

References

[1] International journal of intelligent systems. 2023.
[2] A legal research recommendation system. In *Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop*, New York, NY, USA. ACM.
[3] Jenish Dhanani, Rupa Metha, and Dipti Rana. Effective and scalable legal judgment recommendation using pre-learned word embedding. *Complex Intelligent Systems*, 2022.
[4] Hugo Mentzingen, Nuno Antonio, Fernando Bacao, and Marcio Cunha. Textual similarity for legal precedents discovery: Assessing the performance of machine learning techniques in an administrative court. *International Journal of Information Management Data Insights*, 4, 2024.
[5] Mr.Bhushan Nandwalkar, Kirtish Wankhedkar, Neha Yeolekar, and Upasana Patil. Identification of ipc for police complaint using nlp. *International journal of creative research thoughts(ijcrt)*, 2022.
[6] Ambrish Srivastav. Text similarity algorithms to determine ipc sections for offence report. *International Journal of Artificial Intelligence*, 2022.