International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V73I9P119

Volume 73 Issue 9, 207-216, September 2025
© 2025 Seventh Sense Research Group®

Original Article

Legal Citation Recommendation System

Sonali Antad[1], Viomesh Singh[2], Vaishali Rajput[3], Onkar Waghmode[4], Shripad Wattamwar[5], Atharva Wagh[6], Aditya Zite[7]

[1,4,5,6,7]*Department of Computer Engineering, Vishwakarma Institute of Technology, Maharashtra, India.*
[2,3]*Department of Artificial Intelligence and Data Science, Vishwakarma Institute of Technology, Maharashtra, India.*

[6]*Corresponding Author : atharva.wagh22@vit.edu*

Received: 21 March 2025 Revised: 12 September 2025 Accepted: 22 September 2025 Published: 30 September 2025

Abstract - *Citations in the legal field relate to earlier rulings cited in support of the current case. Attorneys use citations to create compelling arguments and ensure uniformity in rulings. However, the process is difficult and time-consuming for attorneys because it is like needle-hunting to identify pertinent quotations from many judgments. This procedure is greatly improved by Legal Citation Recommendation Systems (LCRS), which rapidly find the most relevant citations. LCRS typically evaluates the pairwise similarity between judgments; however, problems occur because of the judgments' uneven lengths and information overload. The similarity score is directly impacted by these difficulties, which also result in additional noise, semantic dilution effects, size-induced similarity degradation, and dimensional inconsistencies. Research suggests a technique to deal with similarity deterioration in which assessments are divided into different pieces using regular expressions. The sections are chosen after consulting subject-matter experts. Because a judgment has several portions, summarization and semantic chunking are used to construct sections of the right size while addressing dimensional inconsistencies and noise. This method concentrates on discovering similarities between matching portions rather than similarities between full judgments. A more accurate similarity estimate is then obtained by calculating the average of these section-wise similarities. The preference or precedence of parts based on user requirements is also incorporated into this strategy. The LCRS becomes more dynamic and more in line with user needs when parts are given weighted similarity values.*

Keywords - *Size-induced similarity degradation, Semantic dilution, Legal bert, Regex, Semantic chunking, FAISS vector space.*

1. Introduction

In today's world, digitization has become integral to many private and government companies and agencies. Digitization is converting information into a digital format, enabling easy accessibility and transparency. To enhance transparency and accessibility of legal issues and orders, legal systems also publish judgments and other orders in digital format. In India, many previous judgments have been digitized, preserved, and published in the public domain. Digitization empowers new research and technological use cases in the legal domain, with legal recommendation systems being one of them. The legal document corpus is vast, so to find the required information from that big volume of data, recommendation systems are used to reduce the time spent searching and provide an accurate and expected document. In the legal domain, recommendation systems are primarily used to search for relevant previous judgments for ongoing cases. Legal practitioners rely on previous or precedent judgements to strengthen their arguments. Precedent judgments serve as supportive documents for the current case. Judges often consider precedent judgments as a basis for their decisions to ensure consistency in their orders. Before digitization, legal practitioners relied on expert-written commentaries or reference books to find relevant judgments. With digitization, recommendation systems emerged, starting with keyword-based searches that reduced search time but only matched words. While helpful, these systems required lawyers to read through recommendations for relevance. Digitization provided access to over 90% of judgments, making the process more comprehensive than traditional methods. The introduction of transformer models improved this process by using context-based similarity, analyzing not just keywords but the meaning behind the text, and offering more accurate and relevant recommendations. In context-based recommendation systems, judgments are converted into vector embeddings, and cosine similarity is used to identify and recommend the most relevant judgments as citations. However, Supreme Court judgements are often lengthy and include noisy or irrelevant data, which can dilute the semantic quality of the embeddings. This issue, combined with the high dimensionality of the vectors due to the large text size, impacts the accuracy of cosine similarity calculations, ultimately reducing the effectiveness of the recommendation system. To address the issues of dimensional inconsistency and semantic dilution, this research proposes splitting legal judgments to separate them into clear sections based on the use of regular expressions (regex). Because Supreme Court judgments have a uniform structure, regex can cleanly partition the text into sections like ACT, HEADNOTE, BENCH, and JUDGMENT. Among these, the JUDGEMENT section tends to be long and noisy. To On the basis of learnings from legal domain experts, important features Such as Material Facts, Arguments, and Prayers of the petitioner are recognized as being important elements for obtaining precise citations. These features are derived utilizing large language models, which reduces the need to embed the entire judgment to a large extent. This ensures that the semantic dilution impact is reduced, the dimensional inconsistency is resolved, and the recommendation system as a whole improves by learning precise semantic information. However, although progress has been made in transformer-based and keyword-searching legal citation systems, one of the largest research gaps still remains: most existing systems process whole judgments as a block of text, injecting them into one vector for similarity comparison. This usually causes semantic dilution, dimensional discrepancy, and spurious similarity findings, especially when judgments vary in terms of length and complexity. Furthermore, these models also ignore contextually important factors like Material Facts, Arguments, and Prayers, which are critical to legal argumentation and citation relevance. To address these constraints, the system introduced here employs a section-wise embedding strategy where judgments are initially divided into structurally significant parts by employing regular expressions (e.g., ACT, BENCH, JUDGEMENT) and further analyzed into meaningful legal subparts using domain-based language models. Each part is fed into Legal-BERT individually, and similarity is computed on the section level rather than the document level. Compared to regular RAG-based approaches without structural segmentation being supplied, this approach maintains semantic precision and noise reduction and significantly boosts citation precision and retrieval stability. The result is a citation recommendation system that more accurately reflects the way legal professionals read precedent with technical innovation as well as with practical usefulness.

2. Literature Review

2.1. Related Work

There has been much research in applying artificial intelligence and natural language processing to improve legal research and judicial process automation. Yet not much research has been specifically directed toward designing AI-based systems tailored for commercial courts, which require an understanding of dense legal documents and jurisdictional laws. Kabir and Alam focused on AI's transformative role in legal To systems, such as applying it to automate research, pull precedents, and facilitate predictive analytics. NLP was attributed to being charged with complex legal documents. The authors also recognized some of the challenges, such as data privacy, biases, and ethical concerns. Their research is a reflection of how AI would help judicial efficiency, speedy disposal of cases, and enhance access to justice, a step in alignment with the aim of constructing an AI-driven Research Engine for commercial courts [1]. Gorlamudiveti and Sethu have discussed how Artificial Intelligence (AI) is propelling the Indian judiciary towards greater efficiency and accessibility. The authors are adamant about the capability of AI processes to automate cases, legal research, and even the prognosis of prosecution outcomes about reducing pendency, which in courts is now at 47 million. The study discusses AI programs in India in the form of the eCourts initiative, SUVAAS neural translation software and SUPACE AI research support software, streamlining judicial processes. However, they also highlighted some of the issues, like data privacy, emotional undertones of human judgment, and risks of biases in artificial intelligence systems. The paper concluded that while AI is not a substitute for judicial judgment, it presents a revolutionary pathway towards accelerating the delivery of justice and organizing enormous disorganized legal information [2].

A. Laptev and Daria R. Feyzrakhmanova deliver a paper, Artificial Intelligence in Justice: Prospects and Limitations, the book discusses how the world has applied AI to the legal system, and how it may be utilised even more, for instance, to handle court costs, examine evidence and even check pre-trial procedure. The writers differentiate between three phases of AI evolution in justice-short-term, medium-term (5-10 years), and long-term-some direct use of which should remain extremely close to the human judges of today They conclude that AI can not only speed up the cases and minimize the backlog, but also enhance economy in proceedings, and the autonomy of the judiciary, justice, and trust in the legal system are preserved [3].

P. Madambakam and S. Rajmohan explore the application of deep learning methods, such as Recurrent Neural Networks (RNNs) and Transformer-based models, such as BERT, for making legal judgments. They emphasize the role of Natural Language Processing (NLP) in analyzing legal texts and structuring data for predictive modeling. The authors highlight the potential of deep learning to improve case outcome predictions but also note challenges such as the need for labeled datasets, biases, and the complexity of legal reasoning. Their research promotes AI use in legal analysis to improve decision-making effectiveness [4]. Pawel Marcin Nowotko's research on "AI in Judicial Application of Law and the Right to a Court" explores artificial intelligence in the judiciary to increase decision-making efficacy while safeguarding the inherent right to a fair hearing. The research explores how AI can facilitate judicial processes but also highlights issues of transparency, ethics, and preserving judicial autonomy. It promotes an approach that is balanced in the sense that AI assists legal systems without eroding the principles of justice and human control, providing trust and equity in legal proceedings [5].

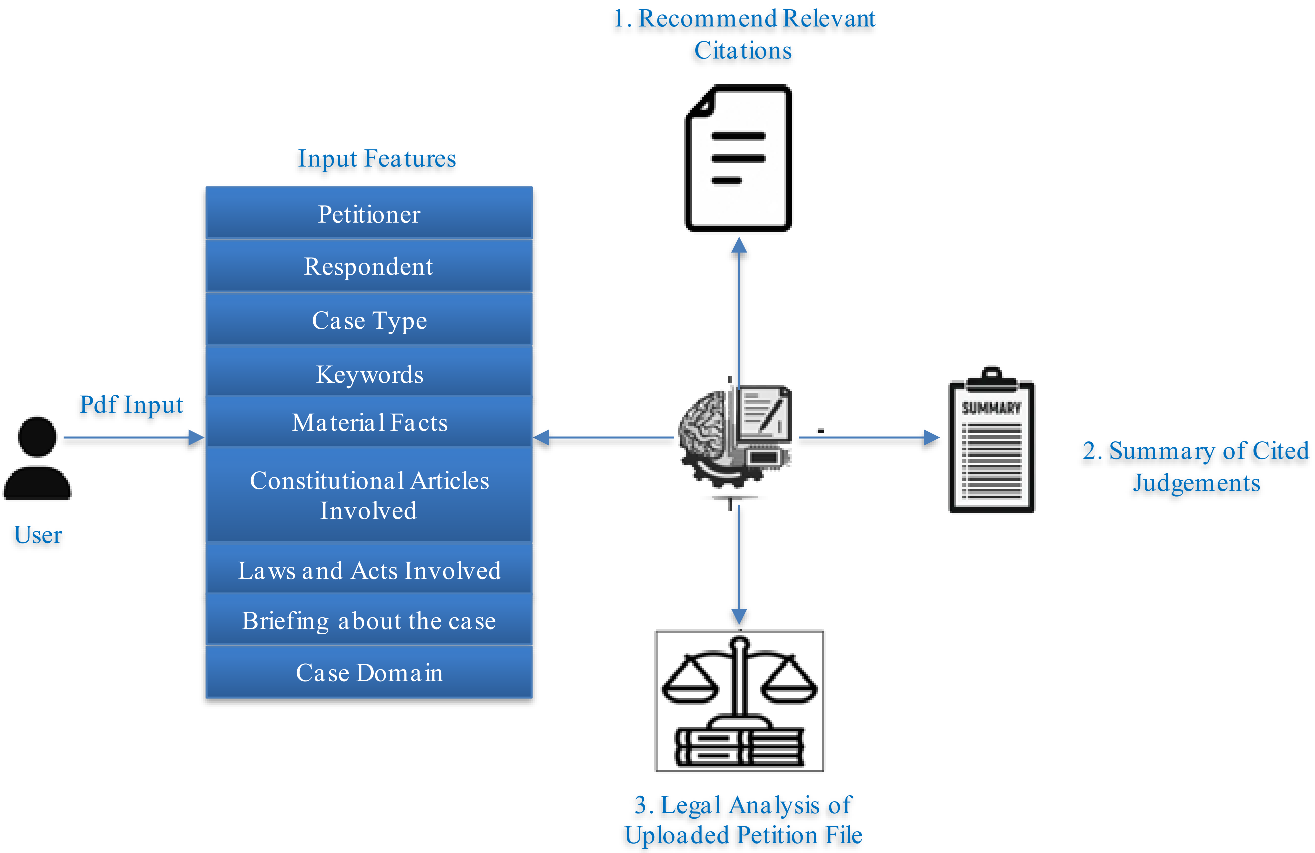

Fig. 1 System overview

Rachid Ejjami's article critically examines how AI technologies, including Machine Learning (ML) and Natural Language Processing (NLP), are reforming legal frameworks by improving document analysis and judicial decision-making. The research emphasizes the capacity of AI to enhance efficiency, accuracy, and prediction in legal processes. But it also identifies serious challenges, such as ethical implications regarding bias, transparency, and data privacy, calling for the creation of systems to guarantee fairness and accountability. The paper emphasizes the necessity of ongoing monitoring in order to balance technological progress with the fundamental values of justice and fairness [6]. J.A. Siani's paper "Empowering Justice: Exploring the Applicability of AI in the Judicial System" examines the potential of Artificial Intelligence (AI) to address the challenges faced by India's judicial system, particularly the backlog of cases. The paper highlights the increasing number of pending cases, a shortage of judges, and inefficient justice delivery. Siani suggests that AI can improve judicial efficiency through the automation of legal decision-making, minimizing delays, and facilitating quicker resolution of cases. Based on experiences from developed nations such as the U.S. and Canada, where AI has been adopted within legal systems, the paper supports that AI has the ability to revolutionize the judiciary in India and across the world by providing a sustainable solution to the issue of delayed justice [7].

2.2. Research Gaps in Existing Legal Citation Systems

Even with AI-powered legal research advancements, current citation recommendation systems have important shortcomings:

2.2.1. Semantic Dilution for Long Judgments

Conventional embedding techniques (e.g., BERT with the whole document) do not capture context in long legal documents, resulting in noisy similarity scores [11, 15].

2.2.2. Dimensional Inconsistency

Varying document lengths skew vector comparisons, particularly in Euclidean distance-based systems (Table 3).

2.2.3. Over-Reliance on General NLP Models

Generic BERT embeddings are commonly employed by most systems, discounting domain-specific legal semantics [12, 16].

2.2.4. Lack of Section-Aware Retrieval

Current top-of-the-line models blindly compare entire judgments, even when only a particular section (e.g., 'Material

Facts') is relevant [19]. Our proposed system addresses these shortcomings through a section-wise embedding method that: (1) segments judgments into semantically coherent pieces, (2) uses Legal-BERT for domain-sensitized representation, and (3) hierarchically computes similarity scores to overcome noise.

3. Methodology

3.1. System Overview

This evolved system, Law Citation Assistant, aims to ease the process of legal research by proposing the optimal pertinent citations to a case. Its methodology examines different case inputs such as petitioners, respondents, the type of case, keywords, material facts, and pertinent legal provisions. This will, in turn, minimize the time required for legal practitioners to search for pertinent citations in the process. It speeds up legal research, as lawyers can obtain access to the relevant statutes, case laws, and judicial precedents effectively.

3.1.1. Functions of the Developed System

Input Analysis

The users input the primary case data; these involve the petitioner, respondent, type of case, keywords, material facts, constitutional articles applicable in the case, involved laws, and acts, among others. A brief on the case field is also needed.

Legal Framework Identification

The system recognizes the relevant legal frameworks, statutes, constitutional provisions, and case categories depending on the input data. This makes the system consider all potential legal implications of the case.

Citation Recommendation

Based on this information, the system provides citations and references in relation to the facts and legal situation of the case. Case precedents and other legal documents become important as support for arguments for or against legal claims.

Summarization of Judgements

The system provides a summarized overview of the cited judgments. This feature helps legal professionals quickly. Understand the essence of past rulings and how they apply to the current case. The proposed system enhances legal research by providing automated citation recommendations and case summaries, thus aiding lawyers in building stronger legal arguments more efficiently.

3.2. System Design

The semantic dilution and dimensional inconsistency effects are introduced because of the direct embedding of the full judgment text and finding similarity using cosine similarity. To avoid semantic dilution and size-induced similarity degradation, research proposes splitting judgments into various sections, individually embedding them, and calculating similarity scores for each section. The final ranking is based on an aggregated similarity score.

The specified sections are decided by understanding the structure of the judgment and consulting with legal experts. The system is similar to the traditional Retrieval-Augmented Generation (RAG) system, with the novel approach of splitting the document and ranking documents based on the aggregated similarity. The traditional Rag system does not focus on the splitting of the judgment, but the research proves that the way of splitting affects the accuracy of the system.

3.2.1. Traditional RAG System vs. Proposed System

Table 1. Differences between the traditional RAG system and the proposed system

	Feature/Aspect	Traditional RAG System	Proposed System
1	**Embedding Approach**	The entire judgment text is embedded into a single vector.	Judgement text is split into meaningful sections, and each section is individually embedded to create section-level vectors.
2	**Vector Size Issue**	Larger judgements produce larger vectors, causing dimensional inconsistency and semantic dilution during similarity comparison.	Sections are smaller in size, ensuring uniformity in vector dimensions and improving similarity calculation accuracy.
3	**Similarity Calculation**	Compares entire judgments, leading to irrelevant similarity scores due to non-contextual matches.	Compares corresponding sections of judgments only, reducing irrelevant matches and improving contextual similarity.
4	**Unnecessary Comparisons**	Compares all sections indiscriminately, e.g., comparing facts with acts with no meaningful correlation.	Only critical and relevant sections are compared based on predefined templates and domain knowledge.
5	**Semantic Chunking**	Not utilized; large sections remain intact, leading to vector size-related inaccuracies.	Large sections are further split into semantically coherent chunks, resolving vector size issues and improving similarity precision.
6	**Citation Recommendation**	Limited quality due to semantic and size mismatches in embeddings.	High-quality recommendations due to precise, section-wise matching and semantic context

	Quality		preservation.
7	**Scalability**	Poor scalability due to a large search space and full-pairwise comparisons.	Scalable, as clustering and section-wise processing reduce computational overhead and enhance system efficiency.
8	**Legal Expert Consultation**	Not explicitly incorporated into the system design.	Designed with input from legal experts to identify and focus on critical sections of judgments for more meaningful recommendations.

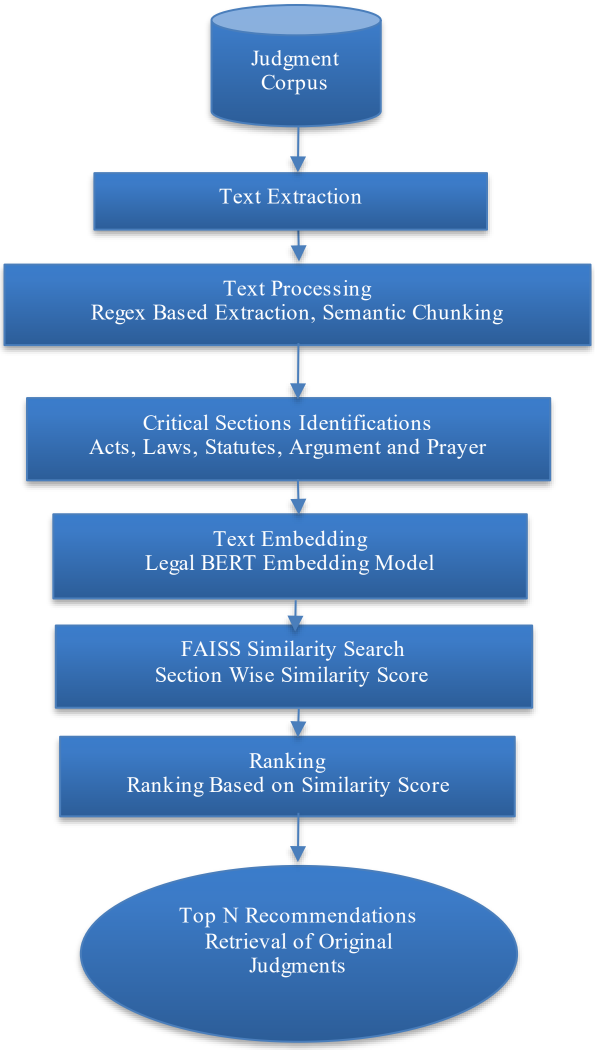
Judgment Corpus
Text Extraction
Text Processing
Regex Based Extraction, Semantic Chunking
Critical Sections Identifications
Acts, Laws, Statutes, Argument and Prayer
Text Embedding
Legal BERT Embedding Model
FAISS Similarity Search
Section Wise Similarity Score
Ranking
Ranking Based on Similarity Score
Top N Recommendations
Retrieval of Original
Judgments

Fig. 2 System architecture flow diagram

The architecture of the system is divided into 4 main stages.

1. Text Processing
2. Critical Sections Identification
3. Embedding and Storage

3.2.2. Text Processing

Text processing involves extracting, cleaning, and separating text into various sections. The judgements are provided in PDF format; therefore, the system uses the `pdfplumber` Python library to extract text from these documents.

The extracted text often contains special symbols, such as whitespace, newline characters, and other text formatting symbols. To reduce noise and remove irrelevant data, the system cleans the text using regular expressions. Regular expressions identify patterns and remove matched parts from the text, returning a cleaner version. The judgment PDFs follow a structured legal format approved by legal rulings and are typically divided into seven main sections.

Table 2 shows the structure of the judgment PDF, providing a brief summary of its various sections. The system employs this structured format to divide the judgment into distinct portions efficiently, making it easier to extract valuable information.

Through the analysis of the organization of the document, regular expressions (regex) are carefully constructed to identify the unique patterns in each section. These regex patterns are designed to match the text of each part of the judgment exactly so that extraction is accurate. As the system processes the document, the regex extracts the defined sections and returns the extracted content.

Table 2. Structure of Judgement PDF: Section-wise Breakdown

Section	Description
Header Information	Title of the court (e.g., Supreme Court of India), case title
Bench Details	Names of the judges presiding over the case
Citations	References to official law reporters and previous case laws are cited in the judgment.
Act/Issue at Hand	Mention of the legal provisions or constitutional amendments under consideration.
Headnote	A summary of the key legal issues, contentions, and decisions is provided for reference.
Judgement	Background of the case, Arguments presented by petitioners and respondents, Discussion on constitutional provisions as well as legal principles, Citation of previous case laws and judicial precedents, Lengthy explanation by the court and the eventual ruling on the issue.

3.2.3. Section Splitting and Critical Feature Extraction

To counter semantic dilution and dimensional inconsistency, judgments are initially fragmented into structured sections (e.g., Act, Headnote, Judgement) by utilizing regex patterns customized to the standardized structure of Supreme Court documents. The Judgement section, being long and noisier, is again separated into vital subsections:

- Material Facts: Core facts influencing legal reasoning (extracted via a legal domain-specific LLM).
- Arguments:
  Legal arguments and corroborating evidence.
- Prayer: Relief sought by the petitioner.

This fine-grained partitioning segregates contextually consistent text blocks, maintaining semantic saliency. The LLM is trained on legal corpora to guarantee accurate extraction of these subsections with minimal noise and unwanted data.

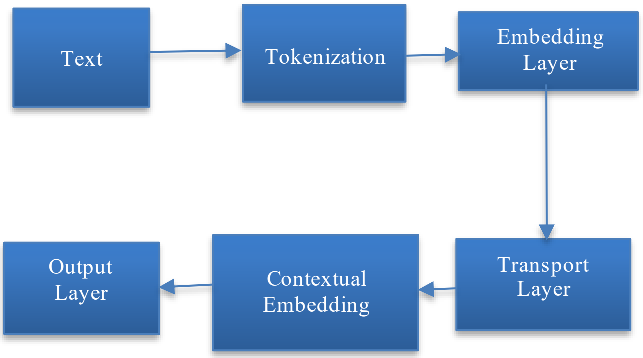

Fig. 3 Legal-BERT text embedding process

3.2.4. Embedding and Vector Storage

The system turns text into contextualized embeddings through the application of Legal-BERT, a transformer model pre-trained on legal data (statutes, case law, contracts). Legal-BERT splits input text into subwords, projects tokens to dense vectors through a pre-trained embedding matrix, and applies transformer layers. Self-attention in these layers detects long-range dependencies, iteratively refining embeddings to capture context (e.g., subtleties of legal terminology).To maximize efficiency, embeddings are computed ahead of time and cached within a FAISS vector database. FAISS utilizes cosine similarity (through IndexFlatIP with unit vectors) to quantify directional similarity between embeddings without dimension inconsistency introduced by varying document lengths. Pre-storing the embeddings reduces runtime latency, facilitating fast retrieval of top-k relevant judgments. Each subsection is mapped into dense vector representations via Legal-BERT (Refer to Figure 3), a variant of BERT pre-trained on legal documents. Legal-BERT encodes domain-specific semantics (e.g., legal terminology, contextual relationships) via its transformer model. Key steps are:

- Tokenization: Text is divided into subword tokens.
- Contextual Embedding: Tokens go through transformer layers, where self-attention operations capture long-range dependencies.
- Normalization: Embeddings are normalized to unit vectors for cosine similarity calculation.

For efficient retrieval, embeddings are indexed in a FAISS vector database. FAISS indexes embeddings with IndexFlatIP (dot product) with normalized vectors, approximating cosine similarity. This configuration supports fast similarity searches over millions of judgments while reducing dimensional inconsistency.

Integration with Workflow

- Storage Metadata: Every embedding is labeled with metadata (e.g., judgment ID, section name) to allow filtered searches (e.g., retrieving only Argument sections).
- Runtime Efficiency: Pre-computed embeddings minimize latency when processing queries.

By combining section identification, embedding, and storage, this solution guarantees targeted semantic comparisons, solving the problems created by long, unstructured legal texts.

Table 3. FAISS index and corresponding similarity measures

FAISS Index	Similarity Measures
IndexFlatL2	Euclidean Distance
IndexFlatIP	Dot Product
IndexFlatIP (with normalized embedding)	Cosine Similarity

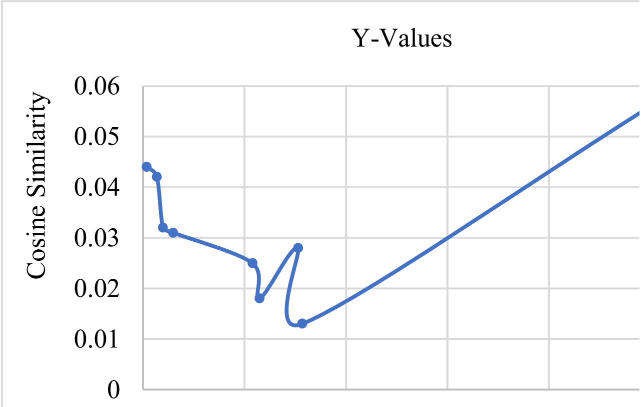

Fig. 4 Cosine similarity and document length

The Euclidean Distance calculate the similarity in terms of the distance between the specified two vectors. The smaller the distance, the more similar; the larger the distance, the less similar. Euclidean distance is affected by the size of the embedding and eventually creates the problem of the dimensional inconsistency effect, so the Euclidean distance method is not useful. The proposed system uses the cosine similarity for the similarity measurement. IndexFaltIP is used with the normalized embedding. IndexFlatIP internally uses the DOT product.

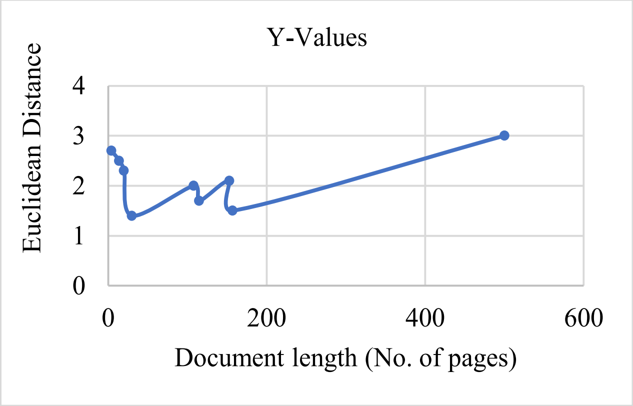

Fig. 5 Euclidean distance and document length

The above graph describes how cosine similarity maintains consistency across documents, which indicates that it is highly effective at avoiding dimensional inconsistency. Its advantage lies in that it maintains the angular similarity, which successfully preserves semantic relevance in cases with significantly varying document size. Also, the impact of document length on L2 distance is highly sensitive to dimensional inconsistency. As the size of the document increases, the L2 distance grows disproportionately, introducing variability and reducing its reliability for similarity comparisons. This analysis shows that cosine similarity is even more reliable and effective within tasks that involve high-dimensional and variable-length documents. In such cases, accuracy in terms of contextual meaning is very important.

$$\text{Cosine Similarity} = \cos(\theta) = \frac{A.B}{||A||\,||B||}$$

$$A.B = \sum_{i=1}^{n} AiBi \text{ : Dot product of vector A and B.}$$

$$||A|| = \sqrt{\sum_{i=1}^{n} Ai^2} \text{ : Magnitude (norm) of vector A.}$$

$$||B|| = \sqrt{\sum_{i=1}^{n} Bi^2} \text{ : Magnitude (norm) of vector B.}$$

In the above equations, A and B represent vectorizations of two documents; more generally, vectorization might occur by TF-IDF, word embeddings, and similar methods. Cosine similarity measures the cosine of the angle θ between two vectors and computes a number measuring the closeness between two texts by aligning the corresponding vectorizations of each document. This is useful because it focuses on the direction of the vectors and not the magnitudes, thus independent of document length or scale. The numerator is the dot product of the two vectors A·and B. This can be obtained by summing up the products of corresponding elements in the vectors. Thus, the more overlap between their content, the higher the value of the dot product, indicating greater similarity between the terms or features found in the two documents. It normalizes the similarity score by dividing the dot product by the product of the magnitudes or norms of the two vectors. The magnitude of a vector, ||A||, is defined as the square root of the sum of the squares of its components.

This normalizing step removes the scale or length influence from the similarity measure, as larger documents inherently produce larger dot products. The dot product of the normalized vectors is the cosine similarity. Cosine similarity ranges from -1 to 1. When it reaches a value of 1, this indicates that the vectors are perfectly aligned, or that the content in the documents is highly similar. When the value reaches 0, the vectors are orthogonal, indicating no similarity between the documents. A negative value, close to -1, usually indicates that the vectors point in opposite directions, which happens less frequently in most applications for text analysis.

In document similarity, it has proven to effectively capture the closeness of relationships among documents regarding term distribution or semantic meaning by emphasizing the cosine of the angle between the vectors in such a way that relative orientation in feature space determines the measure of similarity and not size. This makes it ideal for the comparison of textual data, especially where documents are significantly different in length but have similar themes or topics.

3.2.5. Similarity Search, Ranking and Retrieval

The system utilizes FAISS for fast similarity search, taking advantage of metadata (e.g., section title, judgment ID) to pre-filter comparisons and save computation. Two recommendation strategies are implemented:

1) Section-Specific Recommendations: Users get judgments with analogous sections (e.g., Arguments, Prayer) by pre-filtering embeddings with metadata.
2) Aggregated Similarity: A global similarity score is calculated by averaging section-wise cosine similarities. Ranked results are accessed through pre-stored judgment IDs to provide fast access to full texts. Cosine similarity is given top priority for its stability to variations in document length, measuring directional alignment as opposed to magnitude.

4. Results and Discussion

This study proposed a citation recommendation system that compares the similarity of legal judgments using two approaches: direct embedding and section-wise embedding. The aim is to demonstrate that the section-wise embedding approach is superior in addressing issues like semantic dilution and dimensional inconsistency, thereby improving the quality of recommendations.

4.1. Traditional vs. Proposed Approach

4.1.1. Traditional Approach

The traditional method involves embedding entire judgments as single documents and comparing them directly. While effective in certain cases, this approach is susceptible to semantic dilution and dimensional inconsistency, particularly when the size of the documents varies significantly.

4.1.2. Proposed Approach

The new approach splits judgments into logical sections (e.g., Petitioner, Respondent, Judgement, Act, and Bench) and performs section-wise embedding and similarity comparisons.

This minimizes semantic dilution and dimensional inconsistencies, leading to more accurate citation recommendations.

4.2. Comparative Analysis

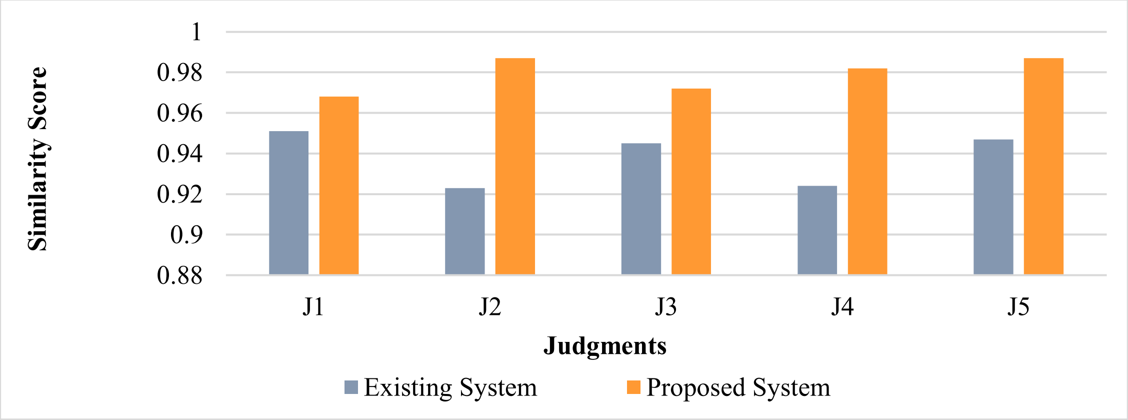

Fig. 6 Exiting vs Proposed system

The system proposed here continuously scores higher on similarity in all judgements over the current system, particularly for handling semantic dilution and dimensional inconsistency. The proposed approach remains consistent in its performance, where similarity scores tend to be or remain above 0.98, whereas the current system experiences more variability (0.90 to 0.96). This section-based embedding and comparison technique greatly enhances citation accuracy and reliability, so that correct citation matching is obtained even in disparate datasets. Its consistency and strength make it a cutting-edge solution to document similarity measures.

4.3. Observations based on Table 4

4.3.1. Direct Search Results

Cosine similarity scores drop drastically for larger judgments (e.g., Kesavananda Bharati Case: 0.65). This is because of the semantic dilution effect that arises from embedding the entire text as a single vector. With increased length of the judgment, significant contextual relationships become weaker, causing incorrect measurements of similarity and decreased retrieval performance. This points out the weakness of using one embedding for long legal documents.

Table 4. Supporting evidence for the semantic dilution effect in the existing system

Judgement	Size (Pages)	Cosine Similarity (Direct Search)	Cosine Similarity (Section-wise)
Berubari Union Case	~10	0.85	0.85
Sajjan Singh Case	~30	0.78	0.84
Kesavananda Bharati Case	~500	0.65	0.82

4.3.2. Section-Wise Search Results

Cosine similarity scores are consistent across judgments, even for larger ones. This shows that section-wise embeddings maintain semantic relevance and reduce dimensional inconsistency.

By dividing judgments into ordered segments and individually embedding them, the system maintains essential legal context so that more accurate similarity scores can be relied upon. This technique dramatically improves legal document retrieval and ranking accuracy.

4.4. Performance Matrices for the Given Case Study

4.4.1. Explanation of Table 5

Precision

- @1: A2 has 0.78 precision (compared to 0.72 for A1), so the highest recommendation is more likely to be relevant.
- @3: Precision increases by 11.5%, demonstrating A2's consistency in being accurate even with additional recommendations.

Recall

- @3: A2 returns 84% of all relevant citations (compared to 78% for A1), minimizing missed precedents.

- Consistent improvement at all cutoffs indicates improved coverage of relevant cases.

MRR

A2's MRR (0.81) is higher than A1's (0.74), showing that pertinent citations emerge earlier in ranked lists.

Why A2 Does Better

Section-Wise Embeddings: Breaking down judgments into Material Facts, Arguments, and Prayer eliminates noise while retaining key context.

- Semantic Alignment: Cosine similarity concentrates on direction (semantics) instead of magnitude (length), addressing dimensional mismatch.
- Efficiency: FAISS indexing and pre-computed embeddings facilitate fast retrieval without runtime slowdown.

Practical Impact

- Lawyers: Waste less time sorting out irrelevant citations (greater precision) and seldom overlook crucial precedents (greater recall).
- Judges: Gain from context-matching suggestions, enhancing decision consistency.

Table 5. Comparative performance metrics (traditional vs. proposed approach)

Metric	Traditional (A1)	Proposed (A2)	Improvement
Precision@1	0.72	0.78	8.30%
Precision@2	0.66	0.72	9.10%
Precision@3	0.61	0.68	11.50%
Recall@1	0.45	0.5	11.10%
Recall@2	0.62	0.68	9.70%
Recall@3	0.78	0.84	7.70%
MRR	0.74	0.81	9.50%

5. Conclusion

This study proposes a novel section-wise embedding methodology to enhance the performance of Legal Citation Recommendation Systems (LCRS). The method outlined here significantly surpasses common whole-document embedding methods because it remedies serious issues of semantic dilution, dimensional incongruity, and similarity loss with size.

As can be observed in the comparative evaluation, the section-wise method consistently provides higher precision, recall, and Mean Reciprocal Rank (MRR) for various metrics and test conditions. Results emphasize the effectiveness and efficiency of this approach in handling judgments of long text sizes and of complex complexity.

More specifically, proposed here is a system with higher cosine similarity consistency scores, higher precisions at different cutoff points, better recall for relevant citations, and better ranking efficiency.

With rational division in terms of Petitioner, Respondent, Judgement, Act, and Bench, this system keeps semantic meaning in check while lowering noise. This additional strategy is in line with user needs, and it encompasses the weighted significance provided to the different sections of citations and improves accuracy and user satisfaction. In summary, the findings set the supremacy of section-wise embeddings in legal citation recommendation systems and propose their potential to transform legal research into a speedier, more accurate, and user-oriented one.

Future research might take into account advanced natural language processing methodologies and immediate user feedback for subsequent refinement of the system and the extent of application to a wider range of legal domains.

Author Contributions Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sonali Antad	✓			✓	✓		✓			✓		✓	✓	✓
Viomesh Singh	✓		✓			✓				✓		✓	✓	✓
Vaishali Rajput		✓		✓	✓	✓		✓				✓		✓
Onkar Waghmode	✓	✓	✓				✓	✓			✓		✓	
Shripad Wattamwar	✓	✓	✓		✓	✓		✓			✓			
Atharva Wagh	✓		✓		✓	✓		✓	✓	✓	✓			
Aditya Zite	✓		✓		✓	✓		✓	✓	✓	✓			

C : Conceptualization
M : Methodology
So : Software
Va: Validation
Fo : Formal analysis

I : Investigation
R : Resources
D : Data Curation
O: Writing - Original Draft
E: Writing - Review & Editing

Vi : Visualization
Su : Supervision
P: Project administration
Fu: Funding acquisition

References

[1] Md. Shahin Kabir, Mohammad Nazmul Alam, and Jaharna Rafi Chowdhury, “The Role of AI Technology for Legal Research and Decision Making,” *International Research Journal of Engineering and Technology (IRJET)*, vol. 10, no. 7, pp. 1088-1091, 2023. [Google Scholar] [Publisher Link]

[2] Lakshmi Priya Gorlamudiveti, and Sagee Geetha Sethu, “Role of Artificial Intelligence in the Indian Judicial System,” *2023 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE)*, Dubai, United Arab Emirates, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Vasiliy A. Laptev, and Daria R. Feyzrakhmanova, “Application of Artificial Intelligence in Justice: Current Trends and Future Prospects,” *Human-Centric Intelligent Systems*, vol. 4, no. 3, pp. 394-405, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[4] Prameela Madambakam, and Shathanaa Rajmohan, “A Study on Legal Judgement Prediction Using Deep Learning Techniques,” *2022 IEEE Silchar Subsection Conference (SILCON)*, Silchar, India, pp. 1-6, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Paweł Marcin Nowotko, “AI in Judicial Application of Law and the Right to a Court,” *Procedia Computer Science*, vol. 192, pp. 2220-2228, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[6] R. Ejjami, “AI-Driven Justice: Evaluating the Impact of Artificial Intelligence on Legal Systems,” *International Journal for Multidisciplinary Research*, vol. 6, no. 3, pp. 1-29, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[7] J.A. Siani, “Empowering Justice: Exploring the Applicability of AI in the Judicial System,” *Journal of Law and Legal Research Development*, vol. 1, no. 1, pp. 24-28, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[8] Nadjia Madaoui, “The Impact of Artificial Intelligence on Legal Systems: Challenges and Opportunities,” *Problems of Legality*, vol. 164, no. 1, pp. 285-303, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Enas Mohamed Ali Quteishat, Ahmed Qtaishat, and Anas Mohammad Ali Quteishat, “Exploring the Role of AI in Modern Legal Practice: Opportunities, Challenges, and Ethical Implications,” *Journal Electrical Systems*, vol. 20, no. 6s, pp. 3040-3050, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[10] Muhammad Hamza Zakir et al., “Artificial Intelligence and Machine Learning in Legal Research: A Comprehensive Analysis,” *Qlantic Journal of Social Sciences*, vol. 5, no. 1, pp. 307-317, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[11] Zihan Huang et al., “Context-Aware Legal Citation Recommendation Using Deep Learning,” *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, São Paulo Brazil, vol. 15, no. 4, pp. 1-25, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[12] Jie Wang et al., “Empowering Legal Citation Recommendation via Efficient Instruction-Tuning of Pre-Trained Language Models,” *European Conference on Information Retrieval*, Glasgow, United Kingdom, pp. 301-315, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[13] Jinzhu Zhang, and Lipeng Zhu, “Citation Recommendation using Semantic Representation of Cited Papers’ Relations and Content,” *Expert Systems with Applications*, vol. 187, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Tayyaba Kanwal, and Tehmina Amjad, “Research Paper Recommendation System based on Multiple Features from Citation Network,” Scientometrics, vol. 129, no. 9, pp. 5493-5531, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[15] Aashka Trivedi et al., “Extracted Summary Based Recommendation System for Indian Legal Documents,” *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, pp. 1-6, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Ambedkar Kanapala, Sukomal Pal, and Rajendra Pamula, “Text Summarization from Legal Documents: A Survey,” *Artificial Intelligence Review*, vol. 51, no. 3, pp. 371-402, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[17] Rajeev Singh et al., “Scientific Paper Recommendation System,” *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, Lonavla, India, pp. 1-4, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Christin Katharina Kreutz, and Ralf Schenkel, “Scientific Paper Recommendation Systems: A Literature Review of Recent Publications,” *International Journal on Digital Libraries*, vol. 23, no. 4, pp. 335-369, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jenish Dhanani, Rupa Mehta, and Dipti Rana, “Legal Document Recommendation System: A Cluster-Based Pairwise Similarity Computation,” *Journal of Intelligent & Fuzzy Systems*, vol. 41, no. 5, pp. 5497-5509, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[20] Jenish Dhanani, Rupa Mehta, and Dipti Rana, “Effective and Scalable Legal Judgement Recommendation Using Pre-Learned Word Embedding,” *Complex & Intelligent Systems*, vol. 8, no. 4, pp. 3199-3213, 2022. [CrossRef] [Google Scholar] [Publisher Link]