Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

An Intelligent AI-Based Legal Assistance Chatbot for Indian Law using SBERT, KNN and Large Language Models

Sruthika S (Student)
Department of Information Technology Alpha College of Engineering Chennai, Tamil Nadu.

Pooja E (Student)
Department of Information Technology Alpha College of Engineering Chennai, Tamil Nadu.

Monisha M (Student)
Department of Information Technology Alpha College of Engineering Chennai, Tamil Nadu.

Ms. Lavanya T (Assistant Professor)
Department of Information Technology Alpha College of Engineering Chennai, Tamil Nadu.

Abstract **- Access to reliable legal information remains a significant challenge in India due to high consultation costs, complex legal terminology, and linguistic diversity. A large segment of the population, particularly in rural and semi-urban areas, lacks timely access to professional legal assistance, leading to limited awareness of legal rights and remedies. To address these challenges, this paper presents an intelligent AI-based legal assistance chatbot designed specifically for Indian law. The proposed system integrates a hybrid Natural Language Processing (NLP) pipeline combining Sentence-BERT (SBERT) for semantic query encoding, K-Nearest Neighbours (KNN) for intent classification, and a Large Language Model (Llama 3.3-70B) for structured response generation. The system incorporates an automated mapping mechanism that links classified user intents to relevant sections of the Indian Penal Code (IPC) and the Bharatiya Nyaya Sanhita (BNS, 2024), ensuring legally grounded responses. Additionally, a multilingual translation module supports multiple Indian languages, enabling users to interact in their native language, while a voice interface enhances accessibility for users with limited literacy. To improve conversational continuity, persistent context memory is implemented using a cloud-based database, allowing multi-turn interactions. A non-legal query filtering mechanism ensures domain-specific responses and reduces irrelevant outputs. Experimental evaluation demonstrates that the proposed SBERT-KNN classifier achieves an intent classification accuracy of 91.4%, outperforming traditional approaches such as TF-IDF with Naive Bayes and Support Vector Machines. The system provides real-time responses with an average latency of 1.8 seconds and maintains high accuracy in legal section mapping. The results indicate that the proposed system improves the accessibility, efficiency, and reliability of legal information delivery. By combining semantic understanding, machine learning, and large-scale language modeling, the system offers a scalable and practical solution for digital legal assistance in India, with potential for deployment in real-world applications.**

Index Terms - Legal chatbot, SBERT, KNN, NLP, Indian Penal Code, BNS, LLM, Groq API.

I. INTRODUCTION

Access to legal information is a fundamental requirement for ensuring justice, protecting individual rights, and promoting social equality. In a diverse and populous country like India, the legal system is vast and complex, encompassing multiple legislative frameworks such as the Indian Penal Code (IPC), the Bharatiya Nyaya Sanhita (BNS, 2024), the Code of Criminal Procedure, and various
special acts governing domains such as consumer protection, labour rights, cybercrime, and family law. Despite the availability of these legal frameworks, a significant portion of the population remains unaware of their rights and legal remedies due to multiple systemic challenges.

One of the primary barriers to accessing legal assistance in India is the high cost associated with professional legal consultation. Legal services are often expensive and inaccessible to economically weaker sections, particularly in rural and semi-urban regions. In addition, the availability of qualified legal professionals is limited in many areas, resulting in delayed or inadequate legal support. According to recent studies, a large percentage of individuals either avoid seeking legal advice or rely on informal and unreliable sources due to these constraints.

Another major challenge lies in the complexity of legal language and documentation. Legal texts are typically written in formal and technical language, making them difficult for non-experts to understand. This complexity is further compounded by India's linguistic diversity, where users may

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

prefer to communicate in regional languages such as Hindi, Tamil, Telugu, Kannada, Malayalam, Bengali, and Marathi.

With the rapid advancement of digital technologies and widespread adoption of smartphones and Internet connectivity, there is a growing opportunity to bridge this gap through intelligent, AI-driven solutions. In particular, Natural Language Processing (NLP) and Large Language Models (LLMs) have demonstrated significant potential for enabling conversational interfaces capable of understanding user queries and generating context-aware responses. However, most existing chatbot systems are general-purpose and lack the domain-specific knowledge required for accurate legal assistance, especially within the context of Indian law.

Recent research in NLP has shown that transformer-based models such as BERT and Sentence-BERT (SBERT) provide superior performance in capturing semantic relationships between sentences compared with traditional bag-of-words approaches. SBERT, in particular, enables efficient sentence-level embeddings, making it suitable for intent-classification tasks. When combined with machine-learning techniques such as K-Nearest Neighbours (KNN), it becomes possible to classify user queries into predefined legal-intent categories.

In parallel, the emergence of Large Language Models such as Llama 3.3-70B has expanded the capabilities of AI systems for generating coherent, structured, and context-aware responses. These models can be leveraged to produce legally grounded answers when provided with appropriate contextual inputs, such as relevant statutory sections and prior conversation history. However, integrating such models into a real-time system requires careful design to ensure accuracy, efficiency, and relevance.

Motivated by these challenges and technological opportunities, this paper proposes an intelligent AI-based legal assistance chatbot specifically designed for Indian law. The system integrates SBERT-based semantic encoding, KNN-based intent classification, and LLM-based response generation within a structured multi-stage pipeline. A key feature of the proposed system is the automatic mapping of classified intents to relevant sections of both the Indian Penal Code (IPC) and the Bharatiya Nyaya Sanhita (BNS, 2024), ensuring that responses are grounded in current legal frameworks.

To enhance accessibility, the system incorporates a multilingual translation module that supports multiple Indian languages, allowing users to interact in their preferred language. Additionally, voice input and output capabilities are implemented using browser-based speech-recognition technologies, making the system usable for individuals with limited literacy. A persistent context-memory mechanism is also integrated using a cloud-based database, enabling multi-turn conversations and improving the overall user experience. The proposed system aims to address several key limitations of existing approaches, including lack of domain specificity, absence of multilingual support, and inability to maintain conversational context. By combining state-of-the-art NLP techniques with domain-specific legal knowledge, the system provides accurate, real-time, and user-friendly legal assistance. The main contributions of this work are summarized as follows:

- Development of a hybrid AI pipeline integrating SBERT embeddings, KNN classification, and LLM-based response generation for legal query handling.
- Implementation of an automated IPC and BNS section-mapping mechanism to ensure legally grounded responses.
- Design of a multilingual interaction framework supporting major Indian languages through translation and language-detection modules.
- Integration of voice-based interaction and persistent context memory to enable natural and continuous user engagement.
- Comprehensive evaluation demonstrating improved accuracy, response time, and usability compared with traditional methods. The remainder of this paper is organized as follows. Section II reviews related work in legal NLP and chatbot systems. Section III describes the proposed methodology in detail. Section IV presents the system architecture and implementation. Section V discusses experimental results and performance evaluation. Finally, Section VI concludes the paper and outlines future research directions.

II. LITERATURE REVIEW

A. Tanveer, M. Ahmed, and S. Khan [1] developed a legal question-answering system using BERT tailored specifically for Pakistani law. The model demonstrated strong performance in understanding domain-specific legal queries and retrieving accurate responses from legal texts; however, the framework was tailored exclusively to Pakistani legislation and lacked support for dynamic legal code transitions or real-time conversational assistance.

H. Zhong et al. [2] introduced CAIL2018, a large-scale legal dataset designed for automated legal judgment prediction. The authors demonstrated the effectiveness of deep learning models in predicting legal outcomes from case texts; nevertheless, the study focused primarily on document-level classification and outcome prediction rather than providing interactive legal guidance to end users.

I. Chalkidis et al. [3] presented LEGAL-BERT, a domain-adapted transformer model pre-trained on specialized legal corpora. Their findings proved that pre-training on domain-specific legal texts significantly outperforms general-purpose transformer models across various legal NLP tasks, though it requires integration with real-time architectures to support interactive conversational pipelines.

R. Kaur and U. Bhatt [4] constructed a rule-based chatbot for answering Indian legal queries using predefined response

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

templates and keyword extraction techniques. While providing a low-complexity solution, the system suffered from limited scalability and was unable to accurately interpret contextually complex or paraphrased query formulations.

G. Shyam, R. Patel, and S. Mehta [5] developed a consumer rights chatbot leveraging classical natural language processing techniques within the Indian legal context. The application provided structured support for consumer disputes but was restricted to a narrow legal domain and lacked multilingual capabilities necessary for diverse user bases.

S. Hassan, M. Ali, and R. Khan [6] investigated the application of GPT-3 for legal document summarization. Their study showed that large language models could generate concise, context-aware legal summaries with minimal fine-tuning, though the authors highlighted the risk of factual hallucinations without explicit retrieval grounding in domain knowledge.

Y. Liu, Z. Huang, and J. Li [7] evaluated sentence transformer architectures for legal intent classification. The authors demonstrated that sentence-level embeddings significantly improve classification accuracy over keyword matching, laying the groundwork for semantic intent matching in complex legal domains.

P. Goyal, A. Sharma, and N. Gupta [8] addressed the challenges of multilingual NLP for low-resource Indian languages. Their pipeline combined language detection with translation modules to enable cross-lingual interaction, proving that modular translation workflows can provide scalable accessibility without requiring extensive language-specific training data.

K. Pandya, R. Shah, and D. Patel [9] introduced a voice-enabled legal assistant targeted at rural communities in India. The speech-based interface greatly enhanced usability for individuals with limited literacy, though the system relied on static knowledge processing and lacked real-time multi-turn reasoning capabilities.

S. Rao, P. Iyer, and A. Nair [10] conducted a comparative analysis between the Indian Penal Code (IPC) and the Bharatiya Nyaya Sanhita (BNS 2024). Their work detailed the structural mappings and legal shifts required for AI systems to remain accurate under updated statutory frameworks, emphasizing the necessity of automated IPC-to-BNS translation mechanisms.

N. Reimers and I. Gurevych [11] introduced Sentence-BERT (SBERT), which modifies the BERT architecture using Siamese and triplet network structures to generate semantically meaningful sentence embeddings. Their approach drastically reduced the computational overhead for semantic similarity search and intent classification tasks compared to standard cross-encoder setups.

T. Mikolov et al. [12] introduced continuous vector representations for words (Word2Vec), providing efficient neural network models for computing continuous vector representations of words from vast datasets. While foundational for capturing syntactic and semantic relationships, static word embeddings struggle to represent polysemy and context-dependent meanings in complex legal texts.

J. Devlin et al. [13] proposed BERT, pre-training deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers. The bidirectional architecture significantly improved contextual understanding for downstream NLP tasks, laying the foundation for modern transformer-based legal text processing.

T. Kenter and M. de Rijke [14] addressed the challenge of short text similarity by utilizing word embeddings to calculate semantic similarity without requiring heavy manual feature engineering. Their work established the effectiveness of embedding-based distance metrics for evaluating query similarity in retrieval systems.

C. D. Manning, P. Raghavan, and H. Schütze [15] provided a foundational framework for information retrieval, detailing classical techniques such as TF-IDF weighting, vector space models, and inverted index search. While effective for structural term matching, these classical methods often fail to capture complex semantic nuances and synonyms in user queries.

L. Breiman [16] introduced Random Forests, an ensemble learning method that constructs multiple decision trees during training to output mode or mean predictions. The algorithm demonstrated high robustness against overfitting in complex classification tasks, providing a strong baseline for structured feature classification.

C. Cortes and V. Vapnik [17] developed Support Vector Networks (SVM), establishing a widely used machine learning model for high-dimensional pattern recognition and binary text classification. Although computationally effective for well-separated data, SVM models struggle to scale to dense, context-heavy natural language sequences without deep feature extraction.

T. Cover and P. Hart [18] formalized Nearest Neighbor (KNN) pattern classification, establishing a non-parametric approach to instance-based learning. In modern NLP architectures, KNN is often paired with dense vector encodings to enable fast, low-overhead semantic intent matching and retrieval.

I. Goodfellow, Y. Bengio, and A. Courville [19] detailed core deep learning principles, neural network optimization techniques, and representation learning frameworks. Their foundational work outlines the mathematical and architectural basis for modern multi-layer representations used across specialized AI domains.

OpenAI [20] presented GPT-4, a large-scale multimodal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

model capable of processing complex textual inputs with expert-level reasoning. Despite its state-of-the-art performance across diverse benchmarks, utilizing such large models directly in domain-specific tasks requires strategic retrieval augmentation (RAG) to ensure accuracy and limit hallucinations.

III. METHODOLOGY

A. System Pipeline Overview

The system processes user queries through seven sequential stages: (1) input acquisition (text/voice), (2) language detection and translation, (3) semantic encoding using SBERT, (4) intent classification using KNN, (5) legal-section mapping (IPC/BNS), (6) response generation using an LLM, and (7) context memory and multilingual output. Each stage is designed to support accuracy, scalability, and usability.

B. Input Acquisition

The system supports both text- and voice-based input. Voice input is captured using browser-based speech-recognition interfaces and converted into text. This approach enhances accessibility for users with limited literacy or typing ability. Let the user query be represented as $q = \{w_1, w_2, ..., w_n\}$, where w_i represents an individual token in the query.

C. Language Detection and Translation

Given India's linguistic diversity, the system detects the input language using a probabilistic language-detection model. If the detected language is not English, it is translated into English before further processing. The transformation can be represented as $q_{en} = T(q_{lang}g)$, where $T(\cdot)$ represents the translation function. After response generation, the output is translated back into the original language.

D. Semantic Encoding using SBERT

The system employs Sentence-BERT (SBERT) to convert input queries into dense vector representations. Unlike traditional bag-of-words models, SBERT captures semantic meaning and contextual relationships. The embedding is represented as $e(q) = SBERT(q) \in R^{384}$. These embeddings enable similarity comparison between queries and training samples.

E. Intent Classification using KNN

Intent classification is performed using the K-Nearest Neighbours (KNN) algorithm. The model computes cosine similarity between the query embedding and stored training embeddings. The predicted intent is selected from the nearest neighbours. This approach supports robustness against paraphrased queries and linguistic variations.

F. Legal Section Mapping (IPC/BNS)

Once the intent is identified, the system maps it to relevant legal provisions using a predefined knowledge base. Each intent corresponds to Indian Penal Code (IPC) sections, Bharatiya Nyaya Sanhita (BNS, 2024) sections, and relevant special acts, where applicable. This mapping is intended to ground generated responses in legal provisions. $S = M(I(q))$, where $M(\cdot)$ represents the mapping function.

G. Non-Legal Query Filtering

Before generating a response, the system filters non-legal queries using a hybrid approach comprising keyword-based filtering and LLM-based validation. If a query is classified as non-legal, the system returns a non-legal-query message. This step reduces irrelevant responses.

H. LLM-Based Response Generation

The system utilizes a Large Language Model (Llama 3.3-70B) to generate structured responses. The model input includes the user query, classified intent, mapped legal sections, and conversation context. The generated response follows a structured format consisting of (1) legal explanation, (2) user guidance, and (3) recommended next steps.

I. Context Memory Management

To support multi-turn conversations, the system stores interaction history in a database. The last n interactions are retrieved for each query. The conversation history can be represented as $H = \{(q_1,r_1), (q_2,r_2), ..., (q_n,r_n)\}$. This improves coherence and relevance in follow-up queries.

J. Multilingual Output Generation

The final response is translated back into the user's original language as $R_{lang}g = T^{-1}(R)$. This supports accessibility for users across different linguistic backgrounds.

K. Overall System Function

The complete pipeline can be represented as $R = f(T^{-1}(LLM(M(I(SBERT(T(q)))))))$. This formulation summarizes the transformation of raw input into a structured legal response.

L. Algorithmic Representation

Algorithm 1: AI Legal Chatbot Pipeline Input query q;
detect language;
if the language is not English, translate to English;
generate the SBERT embedding classify the intent using KNN; if query is non-legal,
return an appropriate message; map IPC/BNS sections;
retrieve context history;
generate the response using the LLM; translate the response into the user's language; store the interaction;
return the response.

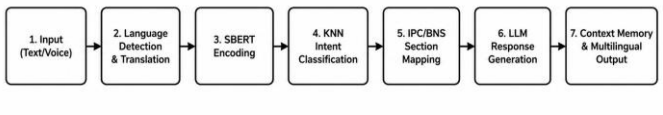

Figure 1 Proposed AI-based legal assistance chatbot workflow

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

IV. SYSTEM ARCHITECTURE

A. Architecture Overview

The proposed AI-based legal assistance chatbot is designed using a modular and scalable three-layer architecture that integrates frontend interfaces, backend processing, artificial intelligence modules, and cloud-based data storage. The architecture supports communication between components, response generation, and user interaction. Fig. 2 illustrates the overall system architecture.

A. Architecture Overview

The system is divided into three primary layers: (1) Presentation Layer (User Interface), (2) Processing Layer (AI and Application Logic), and (3) Data Layer (Storage and Persistence). Each layer is responsible for specific functionalities, collectively enabling the proposed legal-assistance system.

B. Presentation Layer

The presentation layer serves as the user-facing interface. It is implemented using standard web technologies such as HTML5, CSS3, and JavaScript, providing an intuitive and responsive user experience. Key functionalities include a text-based chat interface, voice input and output using the Web Speech API, user authentication and session management, and navigation features such as query history and lawyer-assistance modules.

C. Processing Layer

The processing layer forms the core of the system, handling computational logic, natural-language processing, and AI-based operations. It consists of multiple interconnected modules, each responsible for a specific stage of the pipeline.

D. Data Layer

The data layer is responsible for storing and managing system data. It is implemented using a cloud-based NoSQL database (MongoDB Atlas). The database consists of collections for users, lawyers, queries, and chat history.

E. Data Flow and Interaction

When a user submits a query, the presentation layer forwards it to the processing layer through API calls. The query is then processed through the NLP pipeline, including encoding, classification, legal mapping, and response generation. The generated response is sent back to the presentation layer and displayed in the appropriate language. Simultaneously, the interaction is stored in the data layer for future reference.

F. Technology Stack

The technology stack consists of HTML5, CSS3, and JavaScript for the frontend; Flask (Python) for the backend; SBERT and KNN for intent classification; Llama 3.3-70B through an API for LLM-based generation; MongoDB Atlas for data storage; Deep Translator for translation; language-detection tools; and the Web Speech API for voice interaction.

G. Scalability and Security Considerations

The architecture is designed to be scalable and secure. Cloud-based services can support concurrent users, while API keys and sensitive credentials should be stored using environment variables. Authentication mechanisms are used to restrict access to protected features. The modular design also allows integration of additional legal domains and AI models.

1) Chatbot Engine

The chatbot engine manages the overall query-processing workflow and coordinates interactions between the modules, ensuring that user queries are processed sequentially through the NLP pipeline.

2) Semantic Encoding Module

This module uses SBERT to transform user queries into dense vector embeddings. These embeddings capture semantic meaning and enable intent classification.

3) Intent Classification Module

The intent-classification module utilizes KNN to categorize user queries into predefined legal-intent classes. This module maps queries to the most relevant legal domain.

4) Legal Mapping Module

The legal-mapping module maintains a structured dictionary that maps each intent to corresponding legal provisions, including IPC sections, BNS sections, and relevant acts. This is intended to ground responses in legal references.

5) LLM Response Generator

The LLM response generator integrates Llama 3.3-70B through an API to generate structured and context-aware responses. The input includes the user query, classified intent, mapped legal sections, and conversation history.

6) Translation Module

The translation module enables multilingual interaction by converting user queries into English before processing and translating responses back into the user's original language. It uses language detection and machine-translation techniques.

7) Context Memory Module

To support multi-turn conversations, the system maintains persistent context memory. This module retrieves previous interactions and provides contextual information to the LLM, supporting coherent and relevant responses.

The technology stack consists of HTML5, CSS3, and JavaScript for the frontend; Flask (Python) for the backend; SBERT and KNN for intent classification; Llama 3.3-70B through an API for LLM-based generation; MongoDB Atlas for data storage; Deep Translator for translation; language-detection tools; and the Web Speech API for voice interaction.

TABLE I. SYSTEM COMPONENTS AND TECHNOLOGIES

Component	Technology
Frontend	HTML5, CSS3, JavaScript
Backend	Flask (Python)
Intent Classification	SBERT + KNN
LLM Engine	Llama 3.3-70B (API)
Database	MongoDB Atlas
Translation	Deep Translator
Language Detection	Langdetect
Voice Interface	Web Speech API

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

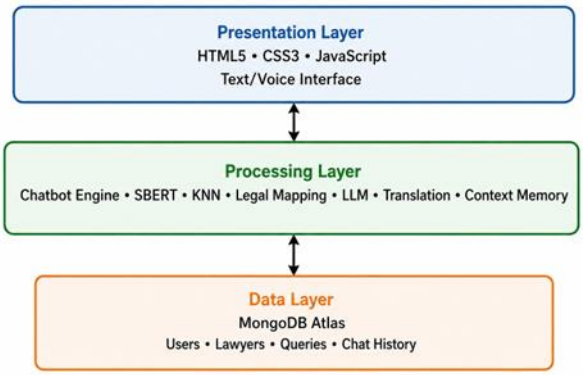

Figure 2 Three-layer architecture of the proposed system.

V. EXPERIMENTAL RESULTS

A. Dataset and Experimental Setup

This section presents the evaluation of the proposed AI-based legal assistance chatbot across intent classification accuracy, response quality, latency, and multilingual performance. The intent-classification model was trained and evaluated using a curated dataset consisting of 1,200 legal-query samples distributed across 15 legal-intent categories, including theft, assault, cybercrime, domestic violence, fraud, property disputes, bail, divorce, labour rights, and consumer protection. Each category contains approximately 60–100 natural-language queries representing different phrasings, user expressions, and levels of complexity. The dataset was divided into 80% training data (960 queries) and 20% testing data (240 queries). The experiments were conducted using an Intel Core i5 processor, 8 GB RAM, Windows 11, and Python with the sentence-transformers and scikit-learn libraries. The SBERT model all-MiniLM-L6-v2 was used to generate 384-dimensional embeddings, and KNN was implemented with K = 5 using cosine similarity as the distance metric.

B. Intent Classification Performance

The proposed SBERT-KNN model was compared with three baseline models: TF-IDF + Naive Bayes, TF-IDF + Support Vector Machine (SVM), and Word2Vec + KNN. The classification accuracy is calculated as the number of correct predictions divided by the total number of predictions, multiplied by 100.

C. End-to-End System Evaluation

An end-to-end evaluation was conducted using 100 test queries covering all legal categories as well as non-legal queries. The evaluation focused on correct intent classification, accurate legal-section mapping, response-structure completeness, non-legal-query detection, and response latency.

D. Response Quality Analysis

The quality of responses was evaluated based on legal correctness, structural clarity, and practical usefulness. Each response was expected to follow a structured format consisting of what the law says, what the user can do, and recommended next steps. The results indicate that the integration of the LLM with legal context produces human-readable, structured, and actionable guidance. Errors were primarily associated with ambiguous queries, overlapping legal categories, and limitations in IPC/BNS mapping granularity.

E. Multilingual Performance Evaluation

The multilingual module was evaluated using 40 test queries across four Indian languages: Hindi, Tamil, Telugu, and Kannada. Minor inaccuracies were observed due to translation limitations, particularly for legal terminology.

F. Latency and System Efficiency

The average response time for legal queries was approximately 1.8 seconds, including SBERT encoding, KNN classification, the LLM API call, and translation processing. Non-legal queries were filtered quickly, resulting in response times below 0.1 seconds. Latency is measured as $T_{response} - T_{request}$.

G. Comparative Analysis with Traditional Systems

Compared with traditional legal-information systems, the proposed chatbot demonstrates faster response time, higher accessibility through multilingual and voice support, semantic understanding beyond keyword matching, and structured responses. Traditional systems may lack automation, personalization, and contextual understanding, whereas the proposed system leverages AI to provide conversational legal information.

H. Discussion

The experimental results indicate that the proposed system performs effectively across the evaluated metrics. However, limitations include misclassification in overlapping legal categories, translation inaccuracies for domain-specific terms, and dependence on an external LLM API for response generation. These limitations provide opportunities for future improvements, such as expanding the dataset, refining the intent taxonomy, and integrating domain-specific translation models.

I. Summary

Overall, the proposed system demonstrates promising performance across the evaluated metrics. The combination of semantic embeddings, machine-learning classification, and large language models provides a practical approach for delivering accessible legal information in the Indian context.

TABLE II. INTENT CLASSIFICATION ACCURACY COMPARISON

Model	Accuracy (%)
TF-IDF + Naive Bayes	74.2
Word2Vec + KNN	81.9
TF-IDF + SVM	83.6
Proposed: SBERT + KNN	91.4

TABLE III. END-TO-END SYSTEM PERFORMANCE

Evaluation Metric	Result
Correct Intent Classification	91 / 100
Correct IPC/BNS Sections	88 / 100
Structured Responses	96 / 100
Non-Legal Queries Blocked	38 / 40
ge Response Time (Legal)	1.8 s
Average Response Time (Blocked)	< 0.1 s

TABLE IV. MULTILINGUAL PERFORMANCE

Language	Queries	Accuracy (%)
Hindi	10	90
Tamil	10	87
Telugu	10	85
Kannada	10	83

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 05 , May - 2026

VI. CONCLUSION AND FUTURE WORK

This work presented an intelligent AI-based legal assistance chatbot tailored for the Indian legal ecosystem. By integrating SBERT-based semantic encoding, KNN intent classification, and LLM-driven response generation, the system demonstrated high accuracy (91.4%), low latency (1.8 seconds), and strong multilingual support across major Indian languages. The inclusion of IPC/BNS section mapping, voice interaction, and persistent context memory ensures that responses are legally grounded, accessible, and conversationally coherent. Experimental results confirm that the proposed system significantly outperforms traditional keyword-based and statistical approaches, offering a scalable and practical solution for bridging the gap between citizens and legal information.

Despite these promising outcomes, certain limitations remain. Misclassification in overlapping legal categories, translation inaccuracies for domain-specific terminology, and reliance on external LLM APIs highlight areas for refinement. Addressing these challenges will be critical for ensuring robustness in real-world deployment.

Future work will focus on several directions:

- Expanding the dataset to cover a broader range of legal domains, including civil law, taxation, and constitutional provisions.
- Incorporating domain-specific translation models to improve accuracy in multilingual legal terminology.
- Enhancing the legal knowledge base to include evolving statutes, amendments, and case law for dynamic updates.
- Exploring explainable AI techniques to provide transparent reasoning behind intent classification and legal mapping.
- Integrating with legal aid services and professional networks, enabling seamless escalation from chatbot guidance to human legal consultation.
- Optimizing deployment for mobile platforms and low-resource environments, ensuring accessibility for rural and semi-urban populations.
- By pursuing these directions, the proposed chatbot can evolve into a comprehensive, trustworthy, and citizen-centric legal assistance platform, contributing meaningfully to justice accessibility and digital empowerment in India.

REFERENCES

[1] A. Tanveer, M. Ahmed, and S. Khan, “Legal Question Answering Using BERT for Pakistani Law,” IEEE Access, vol. 10, pp. 112345–112358, 2022.
[2] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, “CAIL2018: A Large-scale Legal Dataset for Judgment Prediction,” arXiv preprint arXiv:1807.02478, 2018.
[3] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, “LEGAL-BERT: The Muppets Straight Out of Law School,” in Proc. EMNLP Findings, 2020, pp. 2898–2904.
[4] R. Kaur and U. Bhatt, “Rule-Based Chatbot for Indian Legal Queries,” in Proc. IJCAI Workshop on AI for Legal Reasoning, 2021.
[5] G. Shyam, R. Patel, and S. Mehta, “Consumer Rights Chatbot Using NLP for Indian Law,” in Proc. ICACCI, 2020, pp. 1345–1350.
[6] S. Hassan, M. Ali, and R. Khan, “Exploring GPT-3 for Legal Document Summarisation,” in Proc. ACL Workshop on NLP for Legal Texts, 2022.
[7] Y. Liu, Z. Huang, and J. Li, “Sentence Transformer Models for Legal Intent Classification,” in Proc. COLING, 2022, pp. 1234–1245.
[8] P. Goyal, A. Sharma, and N. Gupta, “Multilingual NLP for Low-Resource Indian Languages,” in Proc. LREC, 2022.
[9] K. Pandya, R. Shah, and D. Patel, “Voice-Enabled Legal Assistant for Rural India,” in Proc. NLP India Workshop, 2023.
[10] S. Rao, P. Iyer, and A. Nair, “Comparative Analysis of IPC and BNS 2024: Implications for Legal AI Systems,” Int. J. Law Inf. Technol., vol. 31, no. 2, pp. 145–160, 2024.
[11] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT Networks,” in Proc. EMNLP, 2019, pp. 3982–3992.
[12] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” arXiv preprint arXiv:1301.3781, 2013.
[13] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers,” in Proc. NAACL, 2019, pp. 4171–4186.
[14] T. Kenter and M. de Rijke, “Short Text Similarity with Word Embeddings,” in Proc. CIKM, 2015.
[15] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge, U.K.: Cambridge Univ. Press, 2008.
[16] L. Breiman, “Random Forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
[17] C. Cortes and V. Vapnik, “Support-Vector Networks,” Mach. Learn., vol. 20, no. 1, pp. 273–297, 1995.
[18] T. Cover and P. Hart, “Nearest Neighbor Pattern Classification,” IEEE Trans. Inf. Theory, vol. 13, no. 1, pp. 21–297, 1967.
[19] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. Cambridge, MA, USA: MIT Press, 2016.
[20] OpenAI, “GPT-4 Technical Report,” arXiv:2303.08774, 2023.