IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025

DOI: 10.17148/IJARCCE.2025.14654

# "A Survey Paper on LegalSphere – Your AI-Powered Legal Companion"

**Misba Saba[1], Lekhna L[2], Mohammed Tahir[3], Nawaz Khan[4], Mr. Kumar K[5]**

VI Sem, Dept. of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru[1]

VI Sem, Dept. of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru[2]

VI Sem, Dept. of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru[3]

VI Sem, Dept. of Computer Science and Engineering, K. S. Institute Of Technology, Bengaluru[4]

Associate Professor, Dept. of Computer Science and Engineering, K. S. Institute of Technology, Bengaluru[5]

**Abstract.** The complexity of legal terminology and the limited availability of affordable professional support often hinder access to legal knowledge in India. LegalSphere – Your AI-powered Legal Companion – aims to bridge this gap using Natural Language Processing (NLP) and machine learning techniques. This paper surveys AI and NLP-based approaches to democratize legal access by simplifying legal language, offering intelligent query handling, and providing case-specific responses grounded in Indian laws and regulations. The paper compares transform-based models like BERT and RoBERTa in legal NLP applications and highlights system architecture, dataset strategies, and practical use cases for LegalSphere.

**Keywords:** Natural Language Processing, Legal AI, Transformer Models, BERT, Indian Laws, Legal Advisor.

## I. INTRODUCTION

Legal systems worldwide are complex, and in India, they are further compounded by linguistic barriers and an overload of legal jargon. Most citizens remain unaware of their rights, leading to legal missteps. LegalSphere addresses this issue by providing a conversational AI legal assistant trained in Indian laws, court judgments, and regulations. Users can input queries in natural language and receive concise, legally sound answers. Through technologies such as Hugging Face Transformers and models like Legal-BERT, the project delivers legal literacy at scale.

India's legal system is one of the largest and most complex in the world, governed by an intricate network of constitutional articles, central and state laws, judicial precedents, and administrative rules. Unfortunately, this complexity makes legal information hard to access and even harder for the average citizen to understand. The majority of the population lacks the legal literacy required to understand their rights, obligations, or navigate processes such as filing a complaint, drafting a contract, or responding to a legal notice.

Moreover, access to legal support is often hindered by high costs, geographical constraints, and overloaded judicial systems. LegalSphere seeks to bridge this gap by introducing an AI-powered legal assistant that can interpret user queries in natural language and respond with meaningful, context-aware legal insights. Built on advanced Natural Language Processing (NLP) techniques and trained on Indian legal datasets, LegalSphere offers an accessible, scalable, and intelligent legal advisory platform.

This paper surveys the core technologies behind LegalSphere, including transformer-based language models like BERT and RoBERTa, the challenges of legal text processing, and real-world applications in democratizing legal knowledge.

## II. LITERATURE REVIEW

2.1 Legal NLP: A Specialized Domain: Legal NLP (Natural Language Processing) is a growing research field focused on processing legal texts such as statutes, contracts, and case laws. Unlike general NLP tasks, legal NLP faces challenges such as:

- Archaic language and complex sentence structures
- Long-range dependencies
- Frequent use of domain-specific jargon
- Contextual importance of punctuation, clauses, and citations

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025
DOI: 10.17148/IJARCCE.2025.14654

2.2 Previous Work in Legal AI Early systems relied on rule-based engines, which were rigid and required continuous manual updating. Later, statistical NLP models improved classification and document retrieval but failed in nuanced understanding. The introduction of deep learning and particularly transformers revolutionized legal NLP.

- LEGAL-BERT (Chalkidis et al., 2020): Pretrained on EU and US legal corpora, it shows that domain-specific models vastly outperform general models in legal tasks.

- Indian Legal NLP Benchmarks ( Kalamkar et al., 2021): This study benchmarks various models on Indian legal datasets, emphasizing the need for domain adaptation.

2.3 Key Techniques and Tools
- Transfer Learning with pre-trained models like BERT, RoBERTa
- Fine-tuning on legal corpora such as India Code, court judgments
- NER (Named Entity Recognition) to extract acts, sections, and parties
- QA Systems to handle "What is Section 498A?" type queries
- Summarization Models to condense long judgments into readable summaries

## III. METHODOLOGY

3.1 Dataset Collection and Annotation
The effectiveness of LegalSphere depends heavily on the quality and quantity of legal data used to train it. Datasets are collected from:

- Government websites (e.g., https://indiacode.nic.in/, Supreme Court portal)
- Open-source repositories (e.g., Legal datasets on Kaggle)
- Law journals and academic papers

After collection, the data undergoes:
- Cleaning: Remove metadata, formatting noise, and irrelevant footnotes
- Segmentation: Break long judgments into paragraphs and sections
- Labeling: Annotate legal entities such as Acts, Sections, and Parties using spaCy

3.2 NLP Model Development

LegalSphere uses Hugging Face's transformer models such as:
- RoBERTa: For legal question-answering and contextual understanding
- DistilBERT: For faster inference on low-power machines
- LEGAL-BERT: Where legal-specific understanding is needed

Tasks include:
- Intent Classification: Understanding what type of legal help the user needs
- Document Retrieval: Finding relevant case laws or articles
- Answer Generation: Generating meaningful answers using pre-trained QA models

3.3 Architecture
- Frontend: React-based chatbot/web interface
- Backend: Python FastAPI server for handling requests
- Database: MongoDB for user logs, PostgreSQL for structured law references
- Model Hosting: Azure/GCP/AWS with GPU support

3.4 Evaluation Metrics
- F1 Score, Precision, and Recall for classification tasks
- BLEU / ROUGE Scores for text generation and summarization
- User Feedback Loop for reinforcement learning and model refinement

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.471 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 14, Issue 6, June 2025
DOI: 10.17148/IJARCCE.2025.14654

## IV. CONCLUSION

LegalSphere stands as a promising solution to one of India's most pressing civic problems—accessible legal help. Integrating cutting-edge NLP techniques and India-specific legal data, enables citizens to seek clarity on legal questions without being overwhelmed by technical jargon or procedural complexities.

This survey reveals that while generic NLP models offer a foundation, legal-specific training and fine-tuning significantly boost the relevance and accuracy of the system. Furthermore, the application potential of LegalSphere extends across sectors: from law education and public legal awareness to government services and NGO aid delivery.

Looking ahead, LegalSphere can evolve by incorporating:

- Multilingual capabilities to support regional languages
- Speech-to-text for voice-based legal queries
- Real-time document summarization for lawyers and courts

By democratizing legal access, LegalSphere has the potential to empower millions with the knowledge they need to protect their rights and fulfill their responsibilities—transforming how India interacts with its legal system.

## REFERENCES

[1]. Chalkidis, I., Malakasiotis, P., Aletras, N. (2020). LEGAL-BERT: The Muppets straight out of Law School
[2]. Kalamkar, P., Venugopalan, J., Raghavan, V. (2021). Indian Legal NLP Benchmarks: A Survey, JSAI-isAI Workshops
[3]. Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
[4]. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need
[5]. Hugging Face Documentation – https://huggingface.co/docs
[6]. Supreme Court of India Open Data Portal – https://main.sci.gov.in

  