AI-Based Crime Prediction and FIR Automation

PROJECT SYNOPSIS

**SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE
DEGREE OF BACHELOR OF TECHNOLOGY
IN
COMPUTER SCIENCE & ENGINEERING**

BY

NAME	Registration No:	Roll No:
PUSHKAR PAN	221420110157	14200122139
RANJITA DE	221420110161	14200122146
SUKANYA GHOSH	221420110226	14200122164
SRIJON BANERJEE	221420110218	14200122144
DEBMALYA GHOSH	221420110114	14200122080
ABHISHEK MAITRA	221420110059	14200122173

**Under The Supervision
of
Md. Hamid Islam, Assistant Professor**

**DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
MEGHNAD SAHA INSTITUTE OF TECHNOLOGY
Techno Complex, Beside NRI Complex, Post- Utchepota, Kolkata-700150
Affiliated by**

MAULANA ABUL KALAM AZAD
UNIVERSITY OF TECHNOLOGY,
WEST BENGAL

Session 2025-26

Meghnad Saha Institute of Technology

Nazirabad, P.O. :Utchepota, Via Sonarpur, Kolkata 700150

CERTIFICATE OF APPROVAL

The foregoing project entitled **AI-Based Crime Prediction and FIR Automation** is hereby approved as a creditable study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as prerequisite for the degree for which it has been submitted. It is to be understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein but approve the thesis only for the purpose for which it has been submitted.

Project Guide

Project Coordinator

Head of the Department

Date:

Place:

[Departmental & College Seal]

TABLE OF CONTENTS

Chapters	**Topics**	**Page No.**
Chapter 1	**Introduction**	
	1.1 Purpose	**Page- 3**
	1.2 Domain Definition	**Page- 3**
	1.3 Motivation	**Page- 3&4**
Chapter 2	**Preliminaries**	
	2.1. Basic Natural Language Processing (NLP)	**Page- 5&6**
	2.1.1. Pre-processing and Text Normalisation 2.1.2. Vector Representation Generation (Embeddings) 2.1.3 Evaluation Metrics	
	2.2. Tool Requirements and Specifications	**Page- 6&7**
	2.2.1 Crime Data Types 2.2.2 Crime Pattern Detection 2.2.3 Crime Hotspot Analysis	
	2.3 Predictive Modelling	**Page-7**
	2.4 External Application Tools Requirement	**Page- 7**
Chapter 3	**Literature Review**	
	3.1. Evolution of Legal Text Classification	**Page- 8**
	3.2. Semantic Embedding Models in Information Retrieval	
	3.3. High-Dimensional Indexing and Vector Search	
Chapter 4	**Problem Formulation**	
	4.1. Inadequacy of Lexical Search in Legal Contexts	**Page- 9**
	4.2. Limitations of Rule-Based Systems	

Chapters	Topics	Page no.
Chapter 5	**Proposed Work**	
	5.1. Data Flow and Processing Pipeline 5.1.1 Flowchart	**Page- 10&11**
	5.2. Core Components	
Chapter 6	**Experiment and Analysis**	
	6.1. Performance Metrics (Top-K Recall)	**Page- 12**
	6.2. Confidence Thresholding and Error Analysis	
Chapter 7	**Future Scopes**	**Page- 13**
	Conclusions	**Page- 13**
	References	**Page- 13&14**

Chapter 1. INTRODUCTION

The project addresses the intersection of Crime Analytics, Data Science, Machine Learning, and Legal Informatics. Modern crime patterns, growing data volumes, and limited law enforcement resources necessitate intelligent, data-driven solutions. Traditional policing methods relying on manual processes are inadequate for handling increasing crime volumes, delays in FIR registration and inconsistencies in legal mapping.

The system integrates:
• Crime prediction using Machine Learning and Deep Learning.
• NLP-based Complaint-to-FIR automation.
• Automatic legal mapping under IPC and BNS.
This approach enhances efficiency, transparency, and consistency in the criminal justice system.

1.1. Purpose/Objective

The primary objective is the engineering of a robust, automated system for accurately mapping narrative crime descriptions (FIRs) to the corresponding statutory provisions of the **Bharatiya Nyaya Sanhita (BNS), 2023** [4]. This system ensures an error-minimised legal classification during the legislative transition from the IPC [5] to the BNS. The methodology relies on deep **semantic analysis** over keyword matching to establish conceptual fidelity. The required performance mandates that the correct BNS section be present within the **Top 5** suggested results, thereby achieving substantial efficiency gains for legal professionals and mitigating misclassification errors.

1.2. Domain Definition

This research is situated at the nexus of **Legal Technology** [17] and **Information Retrieval**, automating foundational legal processes within the **Indian criminal justice system**. The system leverages sophisticated components, including NLP and high-speed Vector Databases (FAISS). The operational principle is rapid semantic search: the FIR narrative is converted into a high-dimensional vector, which is used to locate the nearest corresponding BNS statutory vectors. This integration simplifies BNS implementation by ensuring that legal classification is derived from the contextual facts of the case, providing a scalable solution for statutory mapping during the legislative transition.

1.3. Motivation

The application is motivated by three interconnected needs:
1. **Mitigating BNS Transition Complexity:** The BNS enactment necessitates the comprehensive reassignment of criminal acts (e.g., IPC 302 to BNS 101). This large-scale statutory displacement risks **administrative confusion** and clerical error. The automated system serves as an essential transitional

utility [12], enabling the immediate and correct utilisation of the new code without relying on the end-user's memorisation of revised statutory numbers.

2. Enhancing Procedural Accuracy in FIR Filing: Manual classification is susceptible to error, often resulting in incorrect initial charges. Further challenges include the protracted **delay in FIR registration** (up to seven days in some states) due to cumbersome manual verification and the **unstructured nature of complaints** (over 70% from free-text reports), contributing directly to classification inconsistencies. The AI-driven system standardises initial classification, promoting consistency across enforcement bodies and enhancing overall criminal justice efficiency.

3. Addressing the Limitations of Lexical Search: Conventional information retrieval (keyword or Boolean search) is inadequate for complex legal documentation. Lexical systems fail when colloquial terminology in the FIR (e.g., "drunken driving") differs from formal statutory language (e.g., "culpable homicide"). **Lexical search cannot interpret the underlying conceptual similarity or intent**, a critical requirement for accurate legal classification where context and meaning are paramount.

Chapter 2. PRELIMINARIES

2.1. Basic Natural Language Processing (NLP)

NLP constitutes the methodological core, enabling the system to process and interpret human-generated textual input and accurately extract the essential *contextual* elements of the crime from raw narrative data[11].

2.1.1. Pre-processing and Text Normalisation

Before feature extraction, textual data requires rigorous cleaning and normalisation. This stage involves the systematic removal of extraneous linguistic elements such as punctuation, stop words, and redundant boilerplate legal phrases that contribute minimally to semantic uniqueness. The process also includes the standardisation of domain-specific anomalies, such as abbreviations or vernacular expressions commonly found in police reports, ensuring the input data exhibits optimal consistency for all subsequent analytical steps. Supervised learning involves training ML models using labelled datasets. Algorithms learn the mapping between input variables and output labels. In crime prediction, supervised models classify crime types or predict crime intensity.

Common algorithms:

- Logistic Regression
- Support Vector Machines
- Random Forest
- Gradient Boosting
- Neural Networks

2.1.2. Vector Representation Generation (Embeddings)

The most pivotal operational step is the transformation of all text inputs (BNS definitions and FIR narratives) into numerical representations known as **embeddings** or **vectors**. These vectors encode the semantic information of the text into a high-dimensional space (384 dimensions). The underlying principle is that the conceptual similarity between the original texts is directly proportional to the spatial proximity of their corresponding numerical vectors, providing a mathematically quantifiable metric for rapid comparison. Unsupervised learning identifies hidden patterns from unlabelled data. In crime analytics, clustering is used to identify crime hotspots or group similar crime activities.

Algorithms include:

- K-means Clustering
- DBSCAN
- Hierarchical Clustering

2.1.3. Evaluation Metrics

Model performance is measured using:

- Accuracy
- Precision & Recall
- F1-Score
- Confusion Matrix
- ROC-AUC

These metrics help determine the effectiveness of crime prediction models.

2.2. Tool Requirements and Specifications

The system's reliable and high-speed performance is contingent upon the integration of specific, highly optimised software resources:

Types	Tools Name	Function and Application
Software	Python 3.x	Primary programming language for implementing all core logic, data pipeline construction, and workflow orchestration.
Software	Pandas & NumPy	Essential libraries for data manipulation, large-scale cleaning of the BNS dataset, and optimised handling of high-dimensional vector arrays.
Software	sentence-transformers	The software framework hosting the pre-trained **all-MiniLM-L6-v2** model, responsible for generating high-fidelity semantic vector embeddings[2].
Software	faiss-cpu	High-performance indexing library specifically dedicated to executing instantaneous nearest-neighbour searches across the large collection of BNS section vectors[1].
Hardware	Google Colab (CPU/GPU)	Computational environment used during development, testing, and resource-intensive, large-scale generation of the BNS vector index.

Crime analytics refers to the systematic analysis of crime data to support law enforcement, crime prevention, and strategic planning. It involves pattern recognition, spatial analysis, and statistical modelling.

2.2.1. Crime Data Types

Crime datasets may include:

- Crime category (e.g., theft, assault, cybercrime)
- Location (GPS coordinates, streets)
- Time & date
- Demographic attributes

- Severity levels
  These variables help build predictive features.

2.2.2. Crime Pattern Detection

Patterns include:

- Temporal patterns (daily, weekly, seasonal trends)
- Spatial clustering (crime hotspots)
- Demographic influence
  ML algorithms help detect these patterns automatically.

2.2.3. Crime Hotspot Analysis

Hotspots are geographic areas with high crime concentration. GIS, heatmaps, and clustering algorithms identify these regions. Hotspot prediction assists in patrol optimization and resource deployment.

2.3. Concept Basics 3 — Predictive Modelling

Predictive modelling combines data preprocessing, feature engineering, algorithm selection, training, testing, and evaluation.

Steps include:

1. Data Cleaning: removing missing values, duplicates.
2. Feature Engineering: generating useful variables such as time-of-day or crime density.
3. Model Training: selecting ML algorithms.
4. Model Testing: validation using test data.
5. Visualization: interactive dashboards or maps.

Predictive modelling forms the core of the crime prediction system.

2.4. External Application Tools Requirement

Hardware Tools

- Standard PC or laptop with minimum:
  - 8 GB RAM
  - Intel i5 processor or equivalent
  - 250 GB storage

Software Tools

- Python 3.x
- Jupyter Notebook / Google Colab
- Scikit-Learn
- Pandas & NumPy
- Matplotlib & Seaborn for visualization
- GIS libraries (optional): Folium, GeoPandas
- Dataset sources (public crime datasets)

Chapter 3. LITERATURE REVIEW

3.1. Evolution of Legal Text Classification

Automated legal text classification has transitioned from early, brittle **Rule-Based Systems** to highly adaptive machine learning approaches[18]. Initial methods, relying on predefined rules, proved inherently **fragile** and susceptible to failure upon changes in legal terminology or structural reform, such as the IPC to BNS transition. This necessitated the adoption of adaptable statistical and computational linguistic methods, which marked the pivot toward modelling the probabilistic and conceptual relationships between legal texts[3, 15]. Datasets, such as the Chicago Crime Data[6, 7], often serve as benchmarks for evaluating general crime classification models.

3.2. Semantic Embedding Models in Information Retrieval

The advent of the Transformer architecture[8]and its evolution into models like BERT[9] revolutionised Information Retrieval by generating context-aware word representations, or **embeddings**, capable of capturing deeper, contextual meaning. The Sentence-BERT (SBERT) framework[2] optimised this by adapting the base architecture to generate fixed-size vectors for entire sentences or paragraphs. This design significantly reduced computational overhead while preserving high semantic fidelity[19]. The selection of the **all-MiniLM-L6-v2** model is a direct application of this proven SBERT methodology, building upon foundational work like Word2Vec[10].

3.3. High-Dimensional Indexing and Vector Search

The efficient retrieval of information from massive datasets of high-dimensional vectors requires specialised indexing structures, comprehensively addressed by the literature on Approximate Nearest Neighbour (ANN) search. Traditional, exhaustive Euclidean distance calculation becomes computationally prohibitive for large datasets. FAISS[1], demonstrated by Johnson, Douze, and Jégou, utilises advanced techniques like **Product Quantisation (PQ)**[20] or **Hierarchical Navigable Small Worlds (HNSW)** for rapid search. The integration of FAISS is crucial for ensuring the solution's future readiness and scalability to a comprehensive legal knowledge base incorporating vast numbers of judicial judgments. The principles of deep learning[14] underpin the vector generation capability of the system.

Chapter 4. PROBLEM FORMATION

4.1. Inadequacy of Lexical Search in Legal Contexts

The central problem is the critical linguistic disparity between the formal statutory language and the colloquial language of FIRs. Traditional keyword or Boolean search logic fails when the crime narrative's terminology differs from the statute's definition. Lexical search cannot interpret the underlying conceptual similarity or intent of the crime, a necessity for accurate legal classification. This inadequacy is severely exacerbated during the BNS transition, where established mental maps between terms and statutory numbers are rendered obsolete.

4.2. Limitations of Rule-Based Systems

Historically, legal classification utilised **Rule-Based Paradigms** (strict IF-THEN-ELSE logic). While precise in constrained settings, they exhibit critical fragility in dynamic domains like criminal law. The BNS transition illustrates that a complete statutory overhaul necessitates the manual re-coding and rigorous verification of every rule in the system. Rule-based systems demonstrably lack the **adaptability and resilience** required to manage large-scale legislative change, proving them to be an unsustainable foundation for modern legal-tech solutions.

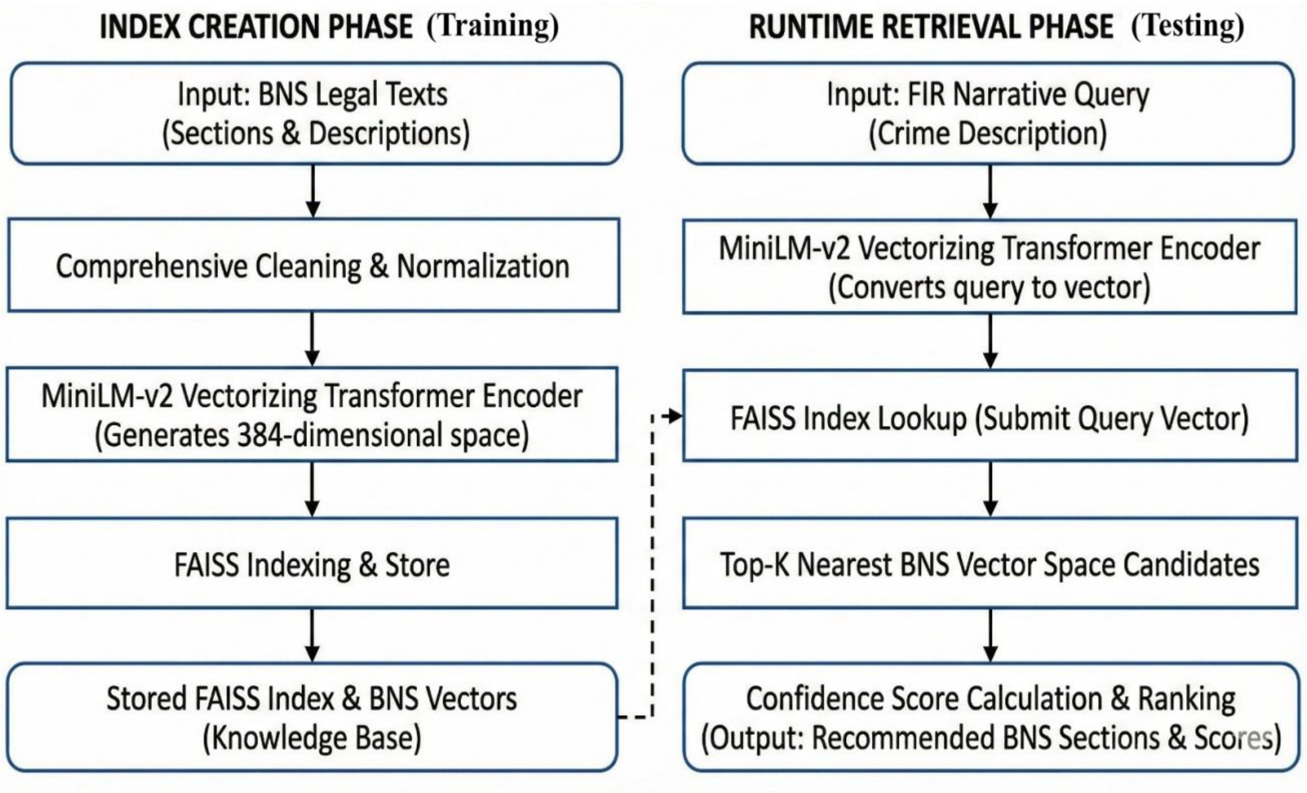
Chapter 5. PROPOSED WORK

5.1. Data Flow and Processing Pipeline

The proposed architecture is founded upon a meticulously structured **Retrieval-Augmented Classification Pipeline**. This pipeline is bifurcated into two distinct phases: the **Index Creation Phase** and the **Runtime Retrieval Phase**.

Index Creation Phase: This offline process involves the ingestion of the entire BNS statutory text, followed by comprehensive cleaning and normalisation (Chapter 2.1.1). The pre-trained **all-MiniLM-L6-v2** model is then applied to generate a dense, 384-dimensional vector for every single BNS section. These vectors are inserted into the FAISS indexing structure, creating the searchable BNS vector database.

Runtime Retrieval Phase: This is the live, user-facing process. A user inputs an FIR narrative (Query), which is subjected to the same pre-processing and vector generation steps. The resulting Query Vector is submitted to the FAISS index, which executes a high-speed L_2 distance search across the BNS vector space. The index retrieves the **Top 5** closest statutory vectors, which are then presented to the end-user as the most probable BNS section classifications, alongside their corresponding confidence scores.


INDEX CREATION PHASE (Training)
Input: BNS Legal Texts (Sections & Descriptions)
Comprehensive Cleaning & Normalization
MiniLM-v2 Vectorizing Transformer Encoder (Generates 384-dimensional space)
FAISS Indexing & Store
Stored FAISS Index & BNS Vectors (Knowledge Base)
RUNTIME RETRIEVAL PHASE (Testing)
Input: FIR Narrative Query (Crime Description)
MiniLM-v2 Vectorizing Transformer Encoder (Converts query to vector)
FAISS Index Lookup (Submit Query Vector)
Top-K Nearest BNS Vector Space Candidates
Confidence Score Calculation & Ranking (Output: Recommended BNS Sections & Scores)

5.2. Core Components

- **Semantic Vector Space Construction:** The operational core relies on mapping all textual data into a **384-dimensional vector space**. This transformation enables the comparison of the FIR narrative against BNS statutes based on comprehensive *semantic content*, where vector proximity directly measures semantic similarity.
- **Embedding Model:** The selection of the **all-MiniLM-L6-v2** model provides an optimal trade-off between computational efficiency and output semantic fidelity. As a compact Sentence Transformer variant, it ensures rapid processing speeds and low memory consumption while providing the necessary analytical depth for statistical prediction.

- **Similarity Measurement:** The metric for quantifying vector similarity is the **Euclidean Distance** (***L_2*** Norm). A minimal distance value signifies maximum semantic similarity:

$$\textbf{Distance(A, B)} = \sqrt{\sum_{i=1}^{n} \left(A_i - B_i\right)^2}$$

- **High-Performance Indexing:** The FAISS library, utilising the **IndexFlatL2** strategy, is employed for its ability to execute exhaustive, brute-force search. This guarantees the mathematically *exact* retrieval of the nearest neighbours, ensuring high-accuracy results without approximation errors for the current scale of the BNS.

Chapter 6. EXPERIMENTS AND ANALYSIS

6.1. Performance Metrics (Top-K Recall)

System effectiveness will be rigorously evaluated using **Top-K Recall** on a substantial test set of factual crime scenarios with pre-verified BNS classifications. The **Top-5 Score** (where the correct section is within the **Top 5** suggestions) is the primary benchmark for confirming the system's practical utility. Achieving a score of **> *85%*** is the target for validating reduced manual verification effort.

6.2. Confidence Thresholding and Error Analysis

The resulting L_2 distance score will be utilised for dynamic **confidence thresholding**. If the minimum distance for the top result exceeds a predefined threshold (e.g., 0.8), a **Low Confidence Warning** is generated, indicating high ambiguity in the input narrative. Error analysis will systematically classify failure modes to facilitate the iterative refinement of the system parameters.

Chapter 7. FUTURE SCOPES, CONCLUSIONS AND REFERENCES

FUTURE SCOPES

The immediate future scope includes:

- **Expansion to Regional Languages:** Integration of multilingual Sentence Transformer models to enable the processing of FIRs filed in major Indian regional languages (e.g., Hindi, Tamil, Bengali).
- **Dynamic Threshold Optimisation via Machine Learning:** Implementing a dynamic confidence threshold system using a separate classifier (e.g., logistic regression trained on search results and known labels) to predict the correctness of the top-ranked result. This mitigates prediction errors in highly ambiguous cases by dynamically adjusting the warning sensitivity, moving beyond a static $\boldsymbol{L_2}$ distance cutoff.
- **Continuous Self-Refinement through Adversarial Training:** Introducing an adversarial training component where synthetic, ambiguous or challenging crime scenarios are generated. The model is then trained to correctly classify these edge cases, actively improving its resilience and fine-tuning its ability to differentiate between highly similar statutory sections.

CONCLUSIONS

The proposed system will integrate crime prediction, complaint-to-FIR automation, and IPC/BNS legal mapping into a unified platform. By using machine learning, deep learning, and NLP, it will quickly convert unstructured complaints into accurate FIR drafts, improving efficiency and transparency.

Predictive analytics will help identify emerging crime trends, while multilingual NLP and automated legal mapping will ensure accurate interpretation and consistent application of IPC/BNS sections.

Overall, the system will streamline policing workflows and support India’s shift toward automated, data-driven, and citizen-centric law enforcement.

REFERENCES

1 .Johnson, J., Douze, M., & Jégou, H. (2019). "Billion-scale similarity search with GPUs." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *38*(5), 978-989.

2 . Reimers, N., & Gurevych, I. (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 390-399.

3 . Surana, S., & Singhal, G. (2020). "Automated Classification of Legal Documents using Deep Learning." *International Journal of Advanced Research in Computer Engineering & Technology*, *9*(12), 11-16.

4 . The Government of India. (2023). "The Bharatiya Nyaya Sanhita, 2023." *The Gazette of India*. Govt. of India. Retrieved 2025-12-01.

5 . The Government of India. (1860). "The Indian Penal Code, 1860." *India Code*. Govt. of India. Retrieved 2025-12-01.

6 . City of Chicago Data Portal. (n.d.). "Crimes — 2001 to Present dataset." *City of Chicago Open Data*. Retrieved 2025-12-01.

7 .Chicago Police Department. (n.d.). "Illinois Uniform Crime Reporting (IUCR) Codes." *City of Chicago Data Portal*. Retrieved 2025-12-01.

8 . Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). "Attention Is All You Need." *Advances in Neural Information Processing Systems* (NIPS 2017), 5998–6008.

9 . Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171-4186.

10 . Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). "Efficient Estimation of Word Representations in Vector Space." *Proceedings of International Conference on Learning Representations* (ICLR 2013).

11 .Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* MIT Press.

12 . Ministry of Home Affairs, India. (2023). "Review of Criminal Laws: The Rationale for New Codes." *Official Policy Document*. Govt. of India. Retrieved 2025-12-01.

13 . Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems*, *33*, 1877-1901.

14 .LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep learning." *Nature*, *521*(7553), 436–444.

15 .Chen, W., Sun, R., Ding, K., & Lin, C. (2021). "Embedding-based Legal Information Retrieval: A Survey." *ACM Computing Surveys (CSUR)*, *54*(6), 1-37.

16 . National Law Institute University, Bhopal. (2024). "Analysis of the Bharatiya Nyaya Sanhita: Challenges in Data Annotation and Implementation." *Academic Journal*.

17 . World Bank Group. (2021). "The Digital Transformation of Justice: Using AI to Improve Efficiency and Access." *Official Report*.

18 . Alani, H., Rayson, P., & Jones, S. (2018). "Automatic Classification of Legal Documents: A Survey." *Artificial Intelligence and Law*, *26*(4), 387-425.

19 . Gao, T., Yao, X., & Chen, D. (2021). "SimCSE: Simple Contrastive Learning of Sentence Embeddings." *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (EMNLP).

20 .Jegou, H., Perronnin, F., Douze, M., & Schmid, C. (2011). "Product Quantisation for Nearest Neighbour Search." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(1), 117-128.

21 . Committee on Criminal Law Reforms. (2023). "Report on the Proposed New Criminal Codes." *Ministry of Law and Justice*. Govt. of India. Retrieved 2025-12-01.