LEGAL TEXT SUMMRIZER

with IPC to BNS Mapping

By Pranshu and Nilesh

LIST OF CONTENTS

Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur.

03 PROBLEM STATEMENT

04 SOLUTION PROPOSED

05 MAPPING CREATION

06 AUGUMENTATION (MAPPING)

07 BART MODEL

08 WHY ONLY BART MODEL

09 HYPERPARAMETER TUNING

10 ROUGE EVALUATION

11 FUTURE SCOPE

PROBLEM STATEMENT

Court cases often consist of extensive documentation, making it difficult for legal professionals to quickly extract relevant information. This leads to time-consuming review processes and increases the risk of overlooking critical details.

SOLUTION PROPOSED

We aim to develop a system that automates the mapping between IPC and BNS and augments legal documents with relevant BNS sections, and generates concise, accurate summaries highlighting key case details and legal updates.

Utilized the BART model for effective text summarization.

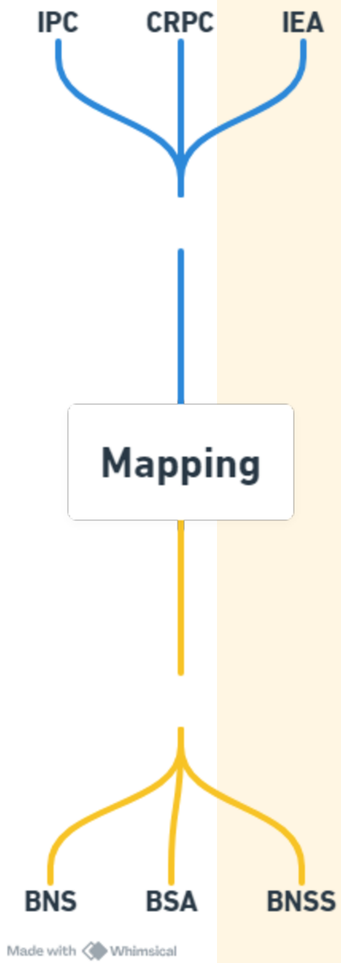
Preprocessed a diverse dataset of court case documents and with proper tagging

Fine-tuned the model to capture key information and legal context.

MAPPING CREATION

To facilitate the transition from the Indian Penal Code (IPC) to the Bhartiya Nayay Sanhita (BNS), we developed a detailed mapping between the two legal systems. This process involved:

- OCR (Optical Character Recognition)
- Nesting Object Identification
- JSON Format



8000+

Lines of mapping

AUGUMENTATION

The augmentation process focuses on updating legal documents by replacing IPC references with their corresponding BNS sections

- **Keyword Identification**

Using regex pattern matching, we identified keywords that typically define legal sections, articles, or acts in legal documents.

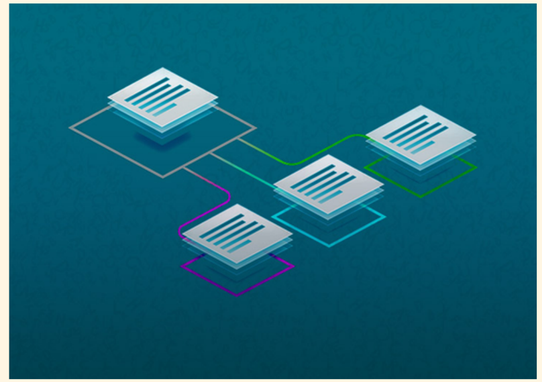
- **Section Extraction & Replacement**

Extract the IPC sections and used the JSON mapping to find the corresponding BNS provisions

Regex Pattern

Extract Sections/Act

Searching in mapping



Augmentation

BART MODEL

BART is a transformer-based model designed for natural language processing tasks, particularly for text generation and summarization.

01

Bidirectional Encoding

BART first processes the entire input text at once, capturing context from both the left and right sides of each word, which helps understand the overall meaning.

02

Auto-Regressive Decoding

For generating text, BART predicts one word at a time based on the previous words it has generated, allowing for coherent and contextually relevant output.

03

Fine-tuning

fine-tuned on specific tasks, such as summarization or translation, making it highly effective for those applications.

WHY ONLY BART?

BART uses a sequence-to-sequence (seq2seq) architecture with both an encoder and a decoder, which makes it ideal for tasks like summarization

- Versatility in Handling Noise
- Scalability
- Model Size and Efficiency

01 PEGASUS-X

Might be good choice, but give abstract summaries and in case of no context, it will use its corpus.

02 BERT

It will added, more complexity as T5 is used generating summaries.

03 Pretrained Models

InLegalBERT, CustomCaseBERT, These models are trained for classification and trained over IPC laws.

HYPERPARAMETER TUNING

Hyperparameter tuning is a crucial aspect of machine learning and model training that involves selecting the best set of hyperparameters for a model to optimize its performance

01

Epochs

Number of iterations

02

Warmup Steps

Gradually increase the learning rate

03

Batch Size

allowing model to process samples per iteration.

04

Weight Decay

applying regularization to prevent overfitting by penalizing large weights.

05

Evaluation Strategy

Determines how often to evaluate the model during training.

06

Gradient Accumulation Steps

Effective utilization of larger batch sizes by accumulating gradients over multiple iterations

ROUGE EVALUATION

ROUGE is a tool used to evaluate how good a generated summary is by comparing it to a human-written summary.

01

ROUGE-N

This measures how many words or phrases (called n-grams) match between the generated summary and a reference summary. Common types include ROUGE-1 (single words) and ROUGE-2 (pairs of words).

02

ROUGE-L

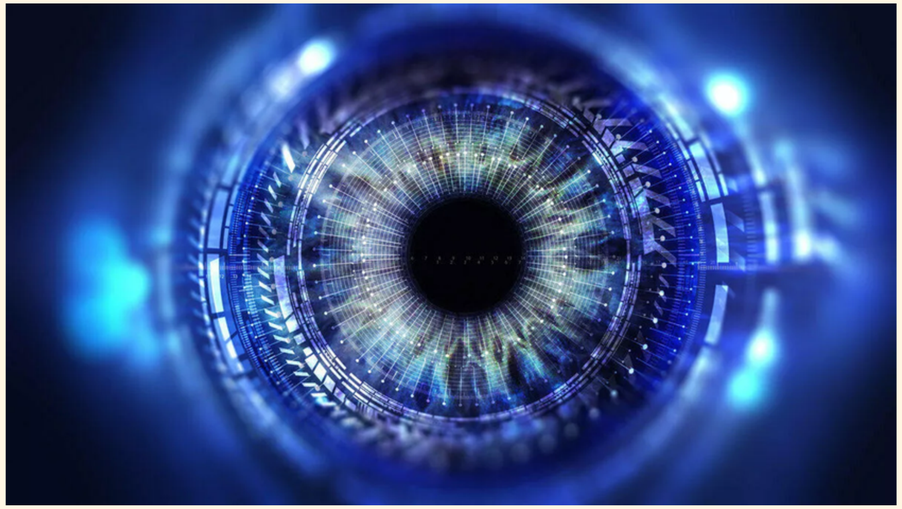
This looks at the longest common sequence of words that appear in both summaries, paying attention to the order in which they appear.

03

ROUGE-L Sum

A variant of ROUGE-L that specifically measures the LCS at the summary level, considering the entire text rather than individual sentences.

FUTURE VISION



RAG Implementation:
Expand the system to analyze various legal documents, such as Implement LightRAG

Multilingual Support:
Develop capabilities for multiple languages to serve a global audience and accommodate diverse legal systems.

Increased Input Document Size:
Allow the system to handle larger documents for more extensive case analysis and summarization.

THANKX FOR YOUR TIME

To learn, an entity must have several choices of behavior; a means of judging the success of its choice, and a way of improving its judgment.