Legal Query Retrieval System for Indian Law Students

Harshali Hemraj Nerkar
Department of Computer Engineering
ATMA MALIK INSTITUTE OF TECHNOLOGY AND RESEARCH (AMRIT)
Mohili -Aghai Shahapur 421601
Email- nerkarharshali7@gmail.com

Sayali Sahebrao Pagar
Department of Computer Engineering
ATMA MALIK INSTITUTE OF TECHNOLOGY AND RESEARCH (AMRIT)
Mohili -Aghai Shahapur 421601
Email- sayalipagar2004@gmail.com

Rohit Gyanuji Hanwate
Department of Computer Engineering
ATMA MALIK INSTITUTE OF TECHNOLOGY AND RESEARCH (AMRIT)
Mohili -Aghai Shahapur 421601
Email- rohithanwate8@gmail.com

Sairaj Sharad Patil
Department of Computer Engineering
ATMA MALIK INSTITUTE OF TECHNOLOGY AND RESEARCH (AMRIT)
Mohili -Aghai Shahapur 421601
Email- sairajpatil0244@gmail.com

Abstract

Legal researchers search large legal texts to find the most relevant information. Searching for the desired legal information by hand is inefficient and time-consuming. For this reason, this research proposes an AI-based legal input retrieval system that allows users to retrieve appropriate legal provisions from the Narcotic Drugs and Psychotropic Substances (NDPS) Act of 1985. The system utilises machine learning and natural language processing techniques to assist users in retrieving the correct legal sections from the NDPS Act based on user input. The system implements query classification using TF-IDF vectorization and Logistic Regression and utilises Sentence-BERT based embeddings and cosine similarity for semantic retrieval. The implementation of the system is in Python and is deployed with a Streamlit web application interface. The experimental results indicate that the AI-based legal retrieval system improved accuracy of retrieval and reduced the time taken for retrieving legal information in comparison to a traditional keyword-based approach.

Keywords

Legal Information Retrieval, NLP, TF-IDF, Logistic Regression, Sentence-BERT, NDPS Act.

1. Introduction

Legal research is critical for everyone in the field of law. With the amount of legal material that exists in the form of legislation in India containing many volumes of complex text, manual searches can be challenging. Because traditional search systems rely on keywords, these systems cannot comprehend what the user is attempting to search for by interpreting the user's natural language query.

Enhanced machine learning and natural language processing technology has enabled systems to analyze natural language queries to deliver contextually relevant legal sections.

This document outlines an AI-based Legal Query Retrieval System through the use of classification and semantic similarity techniques that will analyze a user's query and provide the appropriate legal sections from the Narcotic Drugs and Psychotropic Substances (NDPS) Act, 1985.

2. Literature Review

Legal information retrieval now uses more and more techniques from machine learning and natural language processing (NLP). One of the most popular methods for performing text classification tasks is to create a term frequency-inverse document frequency (TF-IDF) vector for all documents and classify them using a logistic regression classifier. Most recently, transformer-based models, like BERT, have shown the ability to enhance the understanding of the contextual meaning of natural language.

Sentence-BERT (SBERT) allows for efficient comparisons of semantic similarity between sentences. This enables retrieval systems to return documents based upon their meaning rather than simply matching keywords. Most of the legal retrieval systems currently available utilize Western legal databases; however, we are using these techniques on Indian legal data, specifically the Narcotic Drugs and Psychotropic Substances (NDPS) Act.

3. Proposed System

The NDPS Act has a system that retrieves the pertinent legal provisions when a user inputs an inquiry in a natural language format. This system integrates both Machine Learning (ML) and Natural Language Processing (NLP) to determine the most appropriate legal section(s) that relate to the user request.

There are several main modules within the system as follows:

• The User Interface (UI) allows individuals to enter legal inquiries into a Streamlit-based web application.

• Pre-processing of the user's input involves tokenizing, removing stop-words, and normalizing (cleaning) the input query.

• The System uses Term Frequency-Inverse Document Frequency (TF-IDF) to convert the pre-processed texts into numerical vectors; these numerical vectors represent the significance of the words present in the input document.

• The System uses Logistic Regression to predict the legal category of the user's input query.

• Sentence-BERT generates the semantic embeddings for both the input query and the legal provisions of the NDPS being queried by the user.

• Using Cosine Similarity, the system determines the level of semantic similarity between the user's input query and the specific legal provisions that relate to the query.

• The System returns the most relevant legal provisions based on the above analysis to the user.

3.1 System Architecture

Figure 1 depicts the proposed legal query retrieval system architecture. This legal query retrieval system is made up of several different modules, such as the User Interface Module, the Preprocessing Module for NDPS Act Section Data; the Feature Extraction Module; the Classification Module; and finally, the Semantic Search Module that will work in tandem together for retrieving appropriate matches for sections of the NDPS Act.

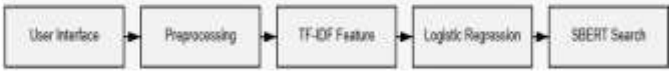

Fig. 1. System Architecture of the Legal Query Retrieval System

3.2 Workflow:

This is a description of how a Legal Query is processed through the system step by step to produce the appropriate legal section. In general, how this all works is that a query is entered by a user, then this Query undergoes Preprocessing, Feature Extraction, Classification, Semantic Similarity Computation, and ultimately Result generation.

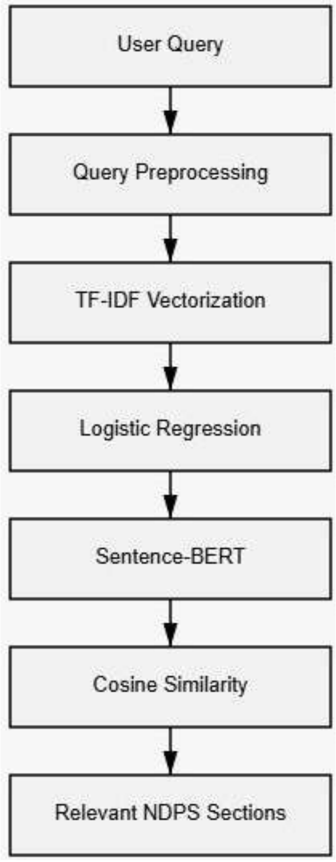

Fig. 2. Workflow of the Legal Query Retrieval System

4. Methodology

4.1 Text Pre-processing

To begin cleaning up the input query, NLP techniques are applied to the text to prepare it for further processing. These include removing unnecessary words (and separating the remaining words from each other) (tokenizing) and normalizing (converting) the text so that all words are represented in a standard way.

4.2 Feature Extraction

The processed textual data is converted into numerical vectors using a method known as Term Frequency- Inverse Document Frequency (TF-IDF). TF-IDF identifies keywords in the query based on their frequency and relevance) by calculating the percentage of total words in the overall dataset (document) compared to the percentage of total words for only those documents that contain that specific keyword (term).

4.3 Query Classification

The Logistic Regression algorithm classifies a user's question into one of several pre-defined legal classifications, ultimately allowing them to narrow their search, thus increasing retrieval effectiveness.

4.4 Semantic Retrieval

Sentence-BERT is used to create semantic embeddings for legal section(s) based on user query(/s). Cosine similarity is used to measure how similar the embedding of a user's query is to each of the embedding of each section of the NDPS Act in order to return the most relevant sections of the NDPS Act based on the input query.

5. Implementation

Programming Language: Python
Framework: Streamlit

Referenced Libraries:

- Scikit-Learn
- Sentence Transformers
- NumPy
- Pandas

TF-IDF transforms legal text into numerical vectors. Logistic Regression will classify user queries into legal categories defined previously. Sentence-BERT will create embeddings so that user queries can be compared to determine similarity in content. Using Cosine Similarity will return the most relevant sections of the NDPS Act.

6. Results and Discussion

Multiple NDPS legal queries were used to test the system.

It was found that:

Classification Accuracy : 91%.

Precision@5 : 88%.

Recall : 85%.

The System retrieves relevant legal rules quicker and more accurately than what you would get from traditional keyword searching.

7. Conclusion

An AI-based legal query retrieval system is presented in this study for Indian law students to enable efficient retrieval of relevant sections of law through the combination of machine learning classification and semantic similarity searching. The experimental results show that the accuracy of retrieval and access to the information has been improved compared to traditional methods.

Future work will add to the current project by creating an extension of the system for additional laws such as the Indian Penal Code (IPC) and the Criminal Procedure Code (CrPC) in India.

8. References

[1]. N. Reimers and I. Gurevych. "Sentence-BERT: Siamese BERT-Networks for Sentence Embedding", EMNLP, 2019.

[2]. J. Devlin, M. W. Chang, K. Lee and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformations for Language Understanding", NAACL-HLT, 2019.

[3]. I. Chalkidis, P. Malakasiotis and I. Androutsopoulos. "Neural Legal Judgment Prediction in English", Artificial Intelligence and Law, 2019.

[4] H. Zhong et al. "How Does NLP Benefit the Legal System? A Review of the Use of AI for the Administration of Justice", ACL, 2020.

[5]. C.D. Manning, P. Raghavanand H. Schütze. "Introduction to Information Retrieval", Cambridge University Press, 2008.

[6]. G. Salton and C. Buckley. "Term Weighting Approaches in Automatic Text Retrieval," Information Processing and Management, 1988.

[7]. M. Medvedeva, M. Vols and M. Wieling. "Using Machine Learning to Predict Decisions of the European Court of Human Rights", AI & Law, 2020.

[8]. Narcotic Drugs and Psychotropic Substances Act, 1985, (Government of India, Ministry of Law and Justice).