International Journal of Electrical and Computer Engineering (IJECE)
Vol. 16, No. 1, February 2026, pp. 246~255
ISSN: 2088-8708, DOI: 10.11591/ijece.v16i1.pp246-255

An information retrieval system for Indian legal documents

Rasmi Rani Dhala[1], Akuleti Vijay S. Pavan Kumar[1], Soumya Priyadarsini Panda[2]
[1]Department of Computer Science and Engineering, Gandhi Institute of Engineering and Technology University, Gunupur, India
[2]Department of Computer Science and Engineering, Silicon University, Odisha, India

Article Info

Article history:

Received Mar 13, 2025
Revised Sep 19, 2025
Accepted Nov 23, 2025

Keywords:

Deep neural network
Domain classification
Legal document retrieval
Natural language processing
Query processing

ABSTRACT

In this work, a legal document retrieval system is presented that estimates the significance of the user queries to appropriate legal sub-domains and extracts the key documents containing required information quickly. In order to develop such a system, a document repository is prepared comprising the documents and case study reports of different Indian legal matters of last five years. A legal sub-domain classification technique using deep neural network (DNN) model is used to obtain the relevance of the user queries with respective legal sub-domains for quick information retrieval. A query-document relevance (QDR) score-based technique is presented to rank the output documents in relation to the query terms. The presented model is evaluated by performing several experiments under different context and the performance of the presented model is analyzed. The presented model achieves an average precision score of 0.98 and recall score of 0.97 in the experiments performed. The retrieval model is assessed with other retrieval models and the presented model achieves 13% and 12% increase average accuracy with respect to precision scores and recall measures respectively compared to the traditional models showing the strength of the presented model.

Corresponding Author:

Soumya Priyadarsini Panda
Department of Computer Science and Engineering, Silicon University
Odisha, India
Email: soumya.panda@silicon.ac.in

1. INTRODUCTION

The field of document information retrieval (IR) concerns on retrieving relevant documents out of a varied document collection based on some user entered query [1]. It has its applications in designing web search engines [2], question answering systems [3], digital libraries [4], recommendation systems [5], organizational data retrieval systems [6]. The users in need of some information may present a query to the system comprising related keywords, and the retrieval model returns a list of output documents as per their content match with the words in the query [7]. With the availability of massive number of digital repositories available online and their rapid growth [8], it is challenging to fetch the required information quickly [9], [10]. Also, user queries may be the partial specification of their needs, handling different ambiguities in the query words and processing the actual intent of the users in different domains is still a challenging task [11]. This makes the IR technology an active area of research with plenty of future research opportunities for the researchers to design specialized methods that may address the context level processing in different domains [12].

There are a number of domain specific retrieval models have been documented in the domains of agriculture, medical, digital libraries, legal information retrieval, and different organizational data retrieval [13], [14]. Over the past few years, the field of legal information retrieval has received significant importance

among legal practitioners and technologists due to its potentiality to bring significant innovation to the legal industry [15], [16]. Also, recently, the IR technology has undergone significant advancements due to the adaptation of different machine learning models to achieve better performance [11]. The use of documents related to different legal matters has the capacity to expand access to justice by providing accessible and high-quality legal support at reduced cost to the users [17]. The IR systems on legal domain may allow extracting useful information from previous cases to support current analysis and quick decision making in similar legal matters [18]. Also, accurate retrieval of legal information is vital to provide access to the law to laymen and legal professionals. Due to the rapid increase in the legal documents available in electronic form, legal IR systems are becoming important these days with a huge demand from varied user communities [19]. This work focuses on using the IR technology in the legal domain to design an IR system for the Indian legal documents.

A legal document retrieval (LDR) system is a variant of the IR technology that focuses on efficiently locating relevant legal documents from collection of documents and case study reports related to different case matters [20]. The LDR systems are very helpful for the legal professionals and other users in need of some legal information to fetch the required information related to some legal matter quickly from a vast landscape of legal documents. As digital information sources on different legal matters are increasing day by day, it is becoming difficult for the legal practitioners to fetch the required information quickly. Also, manual searching of legal information from vast and substantial length of legal documents which includes the reports on different legal case studies, statutes, documents or text related to legal matters, contracts, rules and regulations, legal opinions, is a tedious task [21]. It is crucial in facilitating legal researchers and practitioners to quickly provide access to relevant information from large collection of digital legal information sources [22]. Therefore, there is a need for the development of effective methods to work on the varied legal data sources and may process the vast collection of legal documents and provide the needed information quickly to the legal practitioner to refer to similar case studies in the previously reported cases to be considered to reach to some conclusions.

Legal document retrieval (LDR) is a challenging task and has received interest from both researchers and industry recently to support the law practitioners to minimize the heavy manual work they carried out to perform different case studies [23]. Also, an LDR system may also be useful for the common people in need of some basic legal information to retrieve the appropriate legal information quickly as per their needs. However, the major challenge in designing a legal IR system is due to the variety of legal cases and the vast collection of digital sources, which requires expertise in understanding both the case and the associated law related to that [24]. Also, the frequent legislative changes may render prior case law obsolete or inapplicable in drafting legal acts [25]. Legal texts may also include statutes, case law, regulations, and legal opinions, different quotations from other judgments, and legislative references that require advanced text analysis techniques to be incorporated to identify the exact intent as per the need [19]. The presence of common terminologies with domain-specific terms added more challenges to the development of such system. Legal documents may contain information in the forms of abstract, formal or a judicial language that may contain large narrative parts which are difficult to analyze by simple word level text processing methods. Therefore, retrieval of legal matters from large collection of documents is still a matter of considerable difficulty and specialized methods and approaches are needed due to the distinctive characteristics of legal documents [23].

There is fewer research documented in the LDR domain. The earlier LDR research focuses on extracting relevant legal information using keyword-based match score techniques [17]. Those methods are found to be less effective in terms of fetching the relevant information quickly from huge collections of legal documents with the complex legal text formats [18]. With the shift in research focus towards deep learning architectures, attention-based models emerged as a means to obtain improved data representations in legal domains [19]. In legal document retrieval process, relevant information goes beyond simple keyword matching to process the context and meaning of legal terminologies and concepts. [20] discusses about a legal knowledge graphs technique where, legal notions, cases, statutes, and their linkages are represented graph-wise, allowing semantic searches that take into account the links between different entities to be performed more easily.

In the past decade, LDR technology has been investigated by using different information processing technologies. However, as legal documents frequently employ ritualistic language and rhetorical structures, citations from other norms [21], processing those dependencies and identifying the exact information need by focusing on simple keyword match-based techniques are not found to be effective in legal scenarios [22]. Instead, the semantic aspect of text processing plays vital role in analysis and retrieval of legal documents. In this regard, applying deep learning models may be more beneficial over other traditional models [23]. Designing efficient models to process the complex legal text and extracting the appropriate information quickly from them is still an intricate task. In this regard, retrieval models or databases are available for retrieval of European, New Zealand, UK/US laws and case studies. However, the retrieval of relevant

information from legal documents and case studies related to different Indian case matters is not yet supported by significant research findings [25] and is the focus of the work presented in this paper.

The major contributions of this work include design of a document repository considering the legal case reports and related documents belonging to different Indian legal matters. We present legal document information retrieval system for the Indian legal matters that determines the significance of the user entered queries with respective legal sub-domains and extracts the most suitable documents quickly. The novelty of the work is applying a legal sub-domain classification technique to categorize the queries into respective groups which minimizes the search space and provides the required results quickly. To start with the development of such a model, the document repository is prepared first by collecting the documents and case study reports of Indian legal matters of last five years. As the legal documents may be categorized into some broader legal sub domains, the traditional IR models may not provide the required performance. Instead, the domain-based methods may be more suitable to fulfill the user's needs with respect to the domain-specific requirements. This may reduce the search space significantly and enhance the performance with respect to response time. Therefore, a deep neural network (DNN)-based legal sub-domain categorization method is considered in this work for finding the appropriateness of the queries with respect to the considered 8 legal sub-domains for quick retrieval. A query-document relevance (QDR) algorithm is also presented in this work to further rank the output documents as per their relevance.

The performance of the presented technique is analyzed under several experiments. Tests were conducted on varied number of user queries belonging to the sub domains considered and the accuracy and the time required for classification are analyzed. The model achieves an average accuracy of 98% with average classification time of 0.3 seconds. The DNN-based legal sub-domain classification technique is also compared with other classifiers such as logistic regression (LR), random forest (RF), k-nearest neighbors (KNN), and XGBoost (XGB) classifiers trained in the same environment. The results indicate the strength of the presented technique in relevant legal document retrieval. The presented IR system's performance for Indian legal domain is also assessed over the two retrieval models: Boolean retrieval model (BM) [8], and the fuzzy clustering-based semantic retrieval (FCSR) model [10]. The presented model achieves 13% and 12% average increased results for precision and recall scores respectively over the two models showing the effectiveness of the presented techniques. The reminder of the paper is organized as follows. The proposed model for legal document IR is discussed in detail in section 2. The details of the performance analysis and results are explained in section 3. Section 4 summarizes the research findings with a discussion on the scope of future research in this work.

2. LEGAL DOCUMENT RETRIEVAL MODEL

This section presents a detailed explanation on the domain classification-based legal information retrieval system (DCLIRS) for Indian case documents. As like development of any other standard IR model for document retrieval, our model also considers the 3 major phases in the document retrieval process. Those phases include: document repository creation, query processing, and document retrieval and ranking. The document repository creation phase focuses on collecting the documents which are used in the retrieval process followed by an indexed mechanism. In this work, the legal documents and reports are collected to create a legal document repository and the documents are categorized with respect to their legal sub-domains. The methodology used for repository creation for the Indian legal documents is discussed in section 2.1. The user query processing phase processes the queries entered by the user to normalize the text and obtain the keywords which are used in the retrieval process. The details of the query processing techniques used in this work as discussed in section 2.2.

Based on the matching of the user query key terms with the document key terms, a list of documents are retrieved through the document retrieval phase and are ranked with respect to their relevance to the user queries. There are various ranking algorithms available for the same. However, we have presented a query document relevance score-based algorithm that may better function in the considered legal document retrieval model. The details of the document retrieval approach and the ranking algorithm used are presented in section 2.4. In addition to those, we have included a query domain classification phase to work after the query processing phase. The novelty of this work is to classify the queries to appropriate legal sub domains. This resulted in searching for only those documents that belong to the same domain related to the query instead of searching for the entire document repository. The details of the legal query sub-domain classification phase used are discussed in section 2.3. The input provided to the model is the user queries on varied legal topics and after processing through various phases it presents a ranked list of related documents as the output. The proposed legal document retrieval system for the Indian documents is shown in Figure 1.

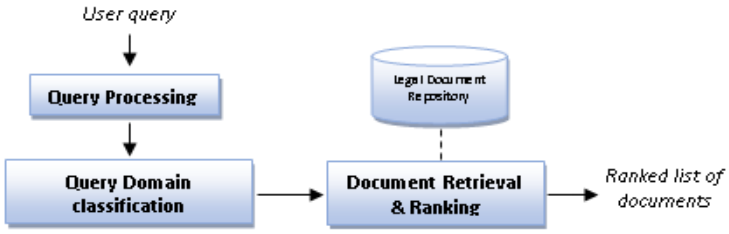

Figure1. Overview of the legal IR model

2.1. Repository creation

At the initial phase of the implementation of the model, a set of 800 legal documents were considered covering 8 legal sub domains such as: criminal, family, civil, corporate, intellectual property, tax, environmental and labor. Those documents were collected from different internet sources available over the web covering different publicly available legal reports and documents in the year range of 2020-2024. The collected documents were grouped into the considered sub-domains by referring to the words present in the document title and the domain of the documents. For this manual grouping process, legal expert's consultation has been taken. An inverted document list is then created for the same. The contents of the file include: the document IDs, the year of publication of the article, the set of associated keywords, and the respective legal sub-domain labels. The associated keywords set for each document is prepared by considering the keywords present in the title of the document along with the related other domain specific information associated with the documents. The designed inverted file is considered for retrieving the documents on user entered queries.

2.2. Query processing

In this phase, the input queries are cleaned to detect the key terms to be considered for document retrieval. The stop words removal and lemmatization techniques were used for this purpose to remove the unwanted words and to obtain the root words of the words respectively. This helps in matching the key term with all morphological variants of the root word to address all possible usage of a word in varied context. A keyword expansion process is then applied to add a set of similar terms for preferable retrieval of the documents. For this purpose, the senses of the words in the dictionary are considered, and the synonyms are considered by using the WordNet [8]. The use of WordNet allows expansion of the input query terms into all possible related words helping in retrieval of more documents from the collection. For each query q_i, a query key term set is prepared that includes the final key terms $\{t_1, t_2, \ldots . t_p\}$ including the root words of the query words. The prepared keyword set is provided to the legal sub-domain classification phase for further processing. The detail of the legal sub-domain classification phase is presented in section 2.3.

2.3. DNN-based legal query domain classification

For every user entered word sequences, the domain classification model considers the keywords from the query processing step and categorizes the query to appropriate legal sub-domains. For this purpose, a DNN based model is considered [24]. To train the model to identify different legal sub-domains, a dataset is created collecting possible user queries from different internet sources available over the web. A collection of 4358 queries with associated domain labels are prepared covering the considered 8 legal sub domains. The queries considered for the domain classification task cover the 8 domains almost in equal proportion. The domain-wise distribution of the prepared data set and the training and testing proportions considered are shown in Figure 2. During preparation of the sample data for the model, it is assumed that any user entered query may be related to at most 3 domain classes. Therefore, the considered problem is a multi-class label classification problem with maximum 3 class labels for each query. Approximately 52% of the prepared dataset are related to a single domain and 48% are to the multiple domain groups. The deep neural network (DNN) based classifier is trained on 80% of the prepared data set on the considered 8 domains. The model is then applied to predict the category of any new query. For this purpose, the remaining 20 % of the data set is considered and the results are analyzed.

The TF-IDF (term frequencies (TF) and inverse document frequencies (IDF)) scores [25] is considered for feature vector creation of the model. For any query q, the TF-IDF of any term t in q is estimated by using the formulas given in (1), (2) and (3). The number of keywords considered for the word representation ranges between 1 to 10 and a zero-filling approach is adopted for the feature value with less than 10 key terms.

$$TF - IDF(t) = TF(t,q) * IDF \quad (1)$$

$$TF(t,q) = \frac{Number\ of\ times\ key\ term\ t\ appears\ in\ the\ query\ q}{Total\ number\ of\ key\ terms\ in\ the\ query} \quad (2)$$

$$IDF(t) = \log_2\left(\frac{Total\ number\ of\ queries}{Number\ of\ queries\ with\ term\ t}\right) \quad (3)$$

The training example (X_i, Y_i) for i = 1 to 3488 is provided as the input to the model. The vector X_i is the feature value of maximum size 10 and Y_i vector represents the considered 8 legal sub-domains. The domain mapping of all the keywords is performed by binary value filling technique, where 1 indicates belongingness and 0 indicates not relevant. If any query i is related to the first and fifth label, the corresponding representation for the same considered is: Y_i= [1, 0, 0, 0, 1, 0, 0, 0]. The number of output classes is indicated as nodes in the output layer. The *"Sigmoid"* activation function is considered and the trained model is fitted with the binary cross-entropy loss function. The input layer uses the *ReLU* activation function. The detailed network architecture is shown in Figure 3.

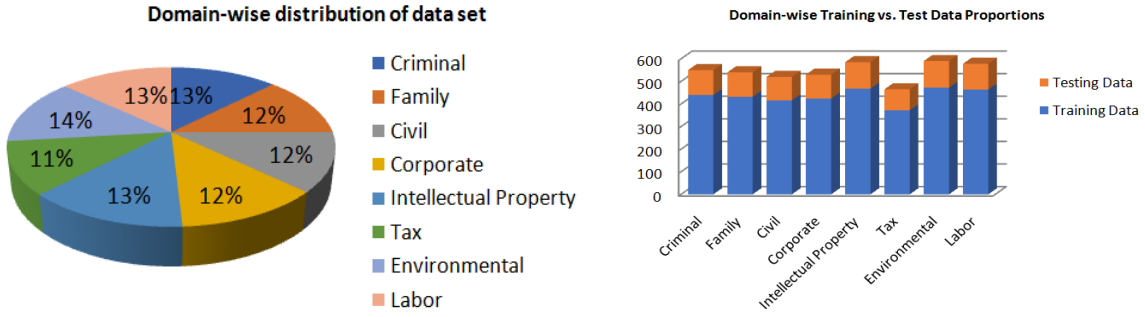

Figure 2. Domain-wise distribution of data set

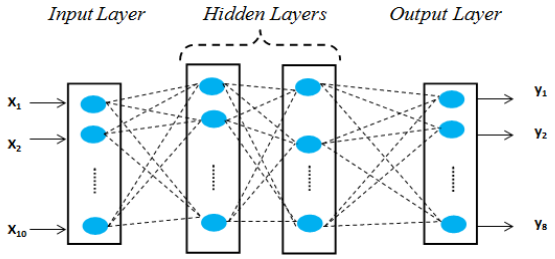

Figure 3. Network architecture of DNN

2.4. Document retrieval and ranking

As per the domain classes obtained in the DNN-based domain classification technique, the words in the query are matched with the document inverted file contents belonging to same domains. Let Q denote the possible queries $\{q_1, q_2, \ldots, q_n\}$ and D represents a document corpus containing legal articles (*i.e.*, $D=\{d_1, d_2, \ldots, d_r\}$). The objective is to fetch a subset $D' \subseteq D$, where each $d_i \in D'$ is relevant to the any query $q_{jk} \in Q$. The reports fetched by the model are then ranked using a query-document relevance (QDR) score-based technique as presented in Algorithm (1). In the algorithm, $\mu(d_i)$ is the function which estimates the significance of d_i to q_i for $i=1, 2, \ldots, m$, where m is the total number of documents retrieved. The document d_m with maximum relevance score (*MRS)* is obtained as per query terms and document term matching results and assigned with a membership value of 1 indicating the most relevant document as per the considered query. All other documents in D' are assigned with a value between *[0-1]* by the membership function $\mu(d_i)$ considering the number of keyword match scores. After obtaining the scores for all documents in D', the relevance score set S is used and the retrieved documents are reordered and presented as the output of the model.

Algorithm 1. QDR score-based ranking

Step-1: Estimate the relevant score set S= {(d_i, $\mu(d_i)$} for D' on query q_j
Step-2: Obtain d_m with *MRS* for each document in D'
Step-3: Assign a value of 1 to the d_m and a value between *[0-1]* to other documents in D' as per the number of keywords match scores by using the function $\mu(d_i)$.
Step-4: Use the values obtained in S for ranking the of output documents.
Step-5: If the score obtained on any document d_i in D' is same as any other document d_j in D', apply reordering of the list considering the recent reports first mechanism.

3. RESULTS AND DISCUSSION

The DCLIRS presented in this paper is implemented under python environment with the use of natural language toolkit (NLTK) tool for text preprocessing. The performance of the presented legal document retrieval system is assessed through four phases of experiments. In the first phase of result analysis, the legal sub-domain classification model is tested to assess its performance in appropriate domain classification. For this purpose, 20% of the remaining data set from the prepared data set is considered. The assessment metrics considered are model accuracy, precision scores and F1 score. The presented legal sub-domain classification model successfully classifies the domains of the new samples with average accuracy of 98.37% and precision score and F1 score of 0.98. Due to the distinctive characteristics of the legal documents, presence of information in abstract, formal or in a judicial language, availability of different quotations from other judgments, legislative references and presence of large narrative parts, the legal data are considered to be very complex patterns in designing any text processing applications. As the DNN-based models can automatically learn complex features from raw data and can deal with large and complex datasets also, therefore, the presented legal domain classification technique is considered to be more effective in the legal documents addressing a variety of forms.

In the second experimental phase, the legal domain classification model is evaluated with other classifiers. For this purpose, the logistic regression (LR), random forest (RF), k-nearest neighbors (KNN), and XGBoost (XGB) classifiers [16] are used. The results obtained in all the experiments are shown in Table 1 it may be observed from the results shown in Figure 4 that the DNN-based query classification model achieves the highest accuracy scores, approximately 98% in all the considered evaluation parameters compared to the other classifiers. The DNN model achieves an increase in accuracy of 3.58% over the LR model, 1.95% over the RF model, 6.44% over KNN and 3.48% over XGB model. In terms of the precision measures, the DNN model achieves an increase of 3% over the LR model, 1% over the RF, 4% over the KNN model, and 2% over the XGB model. While in terms of the F1-score measure, the presented DNN–based domain classification model achieves an increase of 3% over LR, 2% over RF, 6% over KNN, and 3% over XGB classifiers. This shows the presented model is best fitted in the considered domain and the data set over the other classifiers. The confusion matrix (CM) for the considered classifiers is shown in Figure 5. This indicates the actual label and the predicted labels by the respective models in the tests conducted. It may be observed that for all 8 considered domain labels, the DNN-based model outperforms the other models in terms of accurately predicting the domains of the user queries.

Table 1. Result analysis of DNN-based classifier with other classifiers

Text classifier	Accuracy %	Precision Score	F1 Score
Logistic regression	94.79	0.95	0.95
Random forest	96.42	0.97	0.96
K-nearest neighbors	91.93	0.94	0.92
XGBoost	94.89	0.96	0.95
Deep neural network	98.37	0.98	0.98

In the third phase of the experiments, the BM, and FCSR models [7] were considered to compare the results of the presented DCLIRS. Precision evaluation measures and recall metrics are considered for this purpose. A total of 40 random legal queries were collected from different users covering the 8 considered legal domains. Table 2 shows the average accuracy percentage of the tests. While the BM model uses the technique of presence or absence of the keywords in the documents, the FCSR focuses on calculating the relevance score for the documents to belong to some domain groups. However, both models considered a direct match of the key terms and are unable to address the synonym or antonym concepts. The model presented addresses those issues resulting in achieving better results in all the tests conducted. The DCLIRS attain a 96% precision and 95% recall average accuracy respectively. As compared to BM, there has been an increase of 15% and 13% in precision and recall measure respectively. In comparison with the FCSR model, the DCLIRS achieves an increased precision and recall measure of 11% in both measures. Overall, the model

achieves an average 13% increase in precision measure and 12% increase accuracy in recall measure compared to the two considered models.

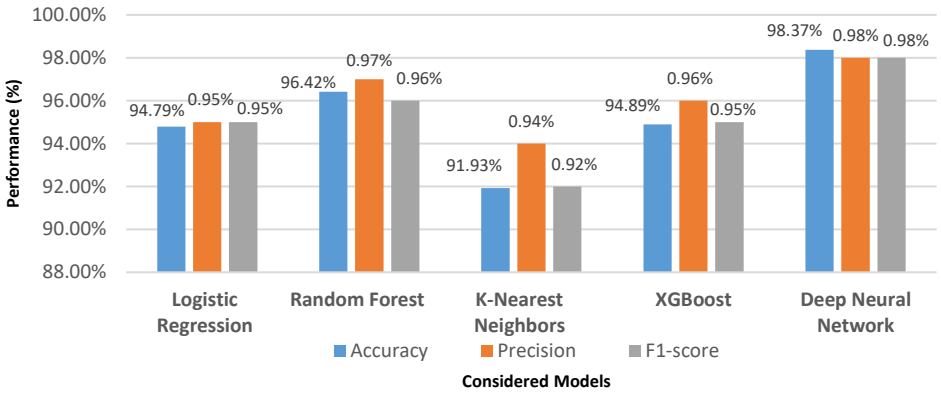

Figure 4. Performance comparison of the query domain classification techniques

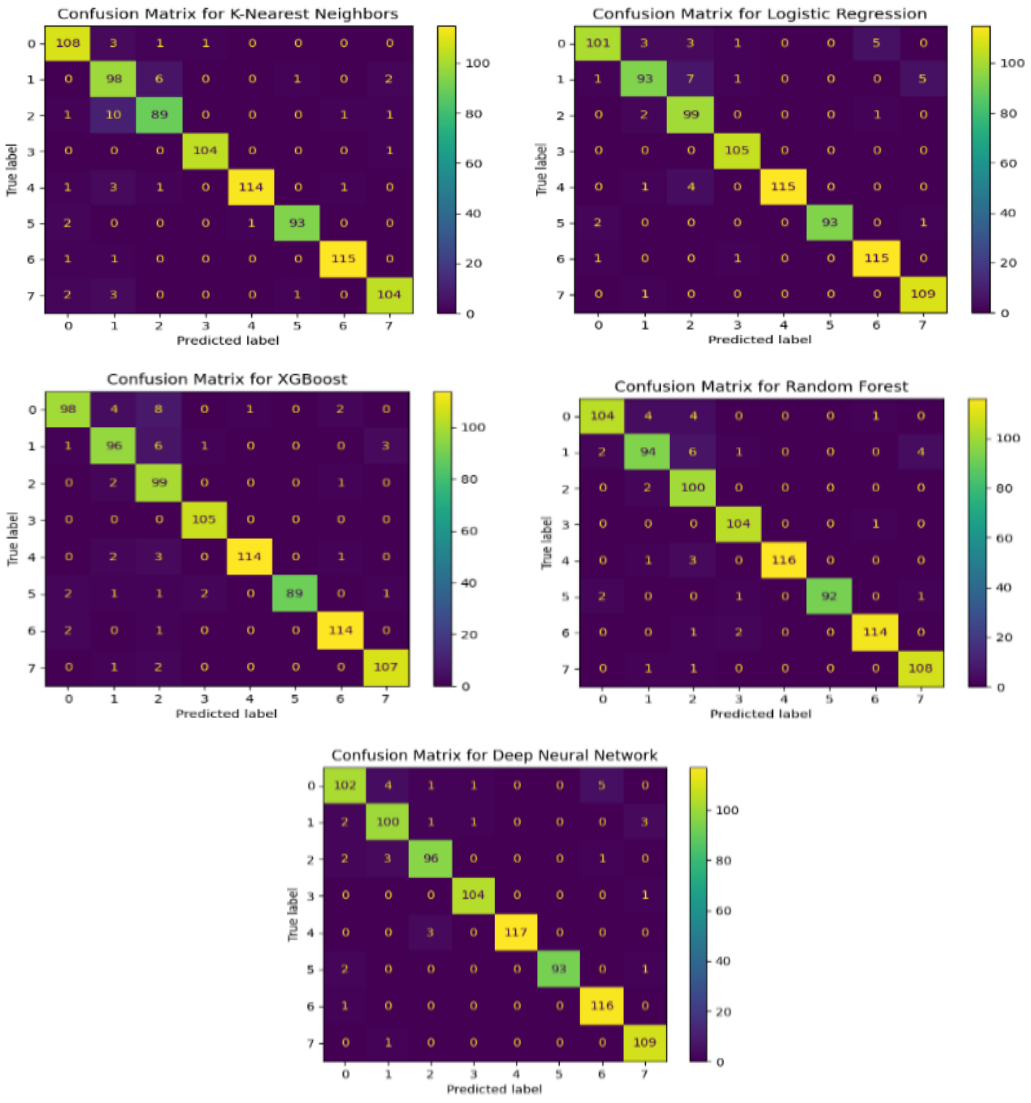

Figure 5. CM for the domain classification techniques

Table 2. Result analysis of IR models

IR model	Precision %	Recall %
BM	81	82
FCSR	85	84
DCLIRS	96	95

In the fourth experimental phase, 40 queries were collected randomly and availability of appropriate related documents in the repository are checked manually. The DCLIRS is evaluated on those samples and the relevant documents were retrieved by the model. The time of submission of the queries and retrieval of appropriate documents are noted and evaluated for all experiments. The domain-wise time in getting the results is shown in Figure 6. The domain-wise time estimations are represented in Figure 6(a) and the average time with respect to key terms in queries is shown in Figure 6(b). It may be noticed that for the considered 8 legal sub domains, the model maintains an average classification time of approximately 0.3 seconds. This may be observed form the graph that there is a linear growth rate of time in all observations.

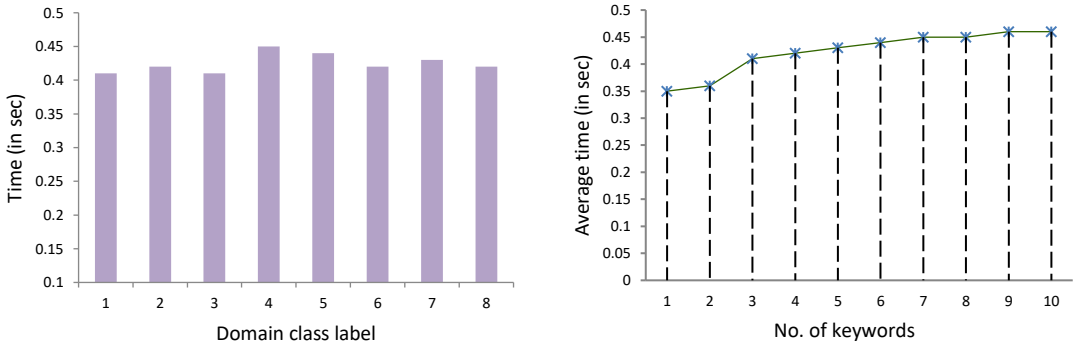

Figure 6. Comparison of average time for (a) legal sub-domains and (b) number of keywords

4. CONCLUSION

In this work, a legal document retrieval system is presented for the Indian case reports. Identification of the appropriateness of the user queries to respective legal sub-domains and retrieval of the most relevant documents quickly are the main objectives addressed in the presented work. A document repository is created for this purpose including 800 legal documents in the year range of 2020-2024 covering 8 legal sub domains. A deep learning based legal sub-domain classification approach is then applied to classify the user queries to appropriate legal domains. A QDR-score algorithm is presented to rank the fetched documents on any user query. Different evaluation metrics were considered to analyze model performance and a number of experiments were conducted. The presented legal sub-domain classification technique achieves an average precision accuracy of 98.37% and F1 score of 0.98 in accurately classifying the queries to respective legal sub domains. Also, the technique takes approximately 0.3 seconds on average to classify the queries containing key terms in the range of 1 to 10 words. The presented DCLIRS model achieves 13% and 12% increased accuracy results in average for precision and recall respectively compared to the BM and the FCSR models. This ensures the credibility of the proposed methodology in Indian legal document retrieval process.

There are diverse areas where work in this research may further be carried out. The semantic level processing of the words present in the user queries may be an important aspect to be included in the work to process the context of the words and achieve better results. This work uses a manual domain labeling technique to label the documents as per their relevance to different sub domains. However, with increase in number of documents and the availability of documents of new domains, development of an automatic labeling technique may make the model more dynamic and adoptive where new domains and documents may easily be included. This may result in development of a more dynamic model which may work for other domains also improving the usability of the system over time.

FUNDING INFORMATION

There is no funding agencies associated with this research.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Rasmi Rani Dhala	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓			
Akuleti Vijay S. Pavan Kumar	✓	✓			✓	✓	✓			✓	✓	✓		
Soumya Priyadarsini Panda	✓	✓		✓	✓	✓	✓		✓	✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

There are no conflicts of interest regarding the publication of this paper.

DATA AVAILABILITY

The datasets used for this research work are available from the corresponding author on reasonable requests.

REFERENCES

[1] R. Bansal and S. Chawla, "Design and development of semantic web-based system for computer science domain-specific information retrieval," *Perspectives in Science*, vol. 8, pp. 330–333, 2016, doi: 10.1016/j.pisc.2016.04.067.

[2] H. K. Azad and A. Deepak, "Query expansion techniques for information retrieval: A survey," *Information Processing and Management*, vol. 56, no. 5, pp. 1698–1735, 2019, doi: 10.1016/j.ipm.2019.05.009.

[3] K. A. Hambarde and H. Proenca, "Information retrieval: recent advances and beyond," *IEEE Access*, vol. 11, pp. 76581–76604, 2023, doi: 10.1109/ACCESS.2023.3295776.

[4] W. Chen *et al.*, "Deep learning for instance retrieval: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7270–7292, 2023, doi: 10.1109/TPAMI.2022.3218591.

[5] Y. Zhu, E. Yan, and I. Y. Song, "A natural language interface to a graph-based bibliographic information retrieval system," *Data and Knowledge Engineering*, vol. 111, pp. 73–89, 2017, doi: 10.1016/j.datak.2017.06.006.

[6] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 721–735, 2009, doi: 10.1109/TPAMI.2008.110.

[7] S. P. Panda and J. P. Mohanty, "An institutional student project report retrieval system using deep neural network-based domain classification technique," *Progress in Artificial Intelligence*, vol. 14, no. 3, pp. 371–385, 2025, doi: 10.1007/s13748-025-00371-2.

[8] N. Girdhar, M. Coustaty, and A. Doucet, "Digitizing history: transitioning historical paper documents to digital content for information retrieval and mining-a comprehensive survey," *IEEE Transactions on Computational Social Systems*, vol. 11, no. 5, pp. 6151–6180, 2024, doi: 10.1109/TCSS.2024.3378419.

[9] H. Wu *et al.*, "Result diversification in search and recommendation: a survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 10, pp. 5354–5373, 2024, doi: 10.1109/TKDE.2024.3382262.

[10] Q. H. Ngo, T. Kechadi, and N. A. Le-Khac, "Domain specific entity recognition with semantic-based deep learning approach," *IEEE Access*, vol. 9, pp. 152892–152902, 2021, doi: 10.1109/ACCESS.2021.3128178.

[11] H. A. Almuzaini and A. M. Azmi, "Impact of stemming and word embedding on deep learning-based Arabic text categorization," *IEEE Access*, vol. 8, pp. 127913–127928, 2020, doi: 10.1109/ACCESS.2020.3009217.

[12] S. Chang, G. J. Ahn, and S. Park, "Improving performance of neural IR models by using a keyword-extraction-based weak-supervision method," *IEEE Access*, vol. 12, pp. 46851–46863, 2024, doi: 10.1109/ACCESS.2024.3382190.

[13] Y. Zhu *et al.*, "Large language models for information retrieval: A survey," *ACM Transactions on Information Systems*, vol. arXiv:2308, 2025, doi: 10.1145/3748304.

[14] W. Song, J. Z. Liang, X. L. Cao, and S. C. Park, "An effective query recommendation approach using semantic strategies for intelligent information retrieval," *Expert Systems with Applications*, vol. 41, no. 2, pp. 366–372, 2014, doi: 10.1016/j.eswa.2013.07.052.

[15] M. Y. Chen, H. C. Chu, and Y. M. Chen, "Developing a semantic-enable information retrieval mechanism," *Expert Systems with Applications*, vol. 37, no. 1, pp. 322–340, 2010, doi: 10.1016/j.eswa.2009.05.055.

[16] C. Sansone and G. Sperlí, "Legal information retrieval systems: state-of-the-art and open issues," *Information Systems*, vol. 106, p. 101967, 2022, doi: 10.1016/j.is.2021.101967.

[17] T. Bench-Capon *et al.*, "A history of AI and law in 50 papers: 25 Years of the international conference on AI and law," *Artificial Intelligence and Law*, vol. 20, no. 3, pp. 215–319, 2012, doi: 10.1007/s10506-012-9131-x.

[18] K. D. Ashley, *Artificial intelligence and legal analytics: New tools for law practice in the digital age*. Cambridge University Press, 2017.

[19] M. Palmirani and G. Governatori, "Modelling legal knowledge for GDPR compliance checking," in *Frontiers in Artificial Intelligence and Applications*, 2018, vol. 313, pp. 101–110, doi: 10.3233/978-1-61499-935-5-101.
[20] M. Medvedeva, M. Vols, and M. Wieling, "Using machine learning to predict decisions of the European Court of Human Rights," *Artificial Intelligence and Law*, vol. 28, no. 2, pp. 237–266, 2020, doi: 10.1007/s10506-019-09255-y.
[21] S. Brüninghaus and K. D. Ashley, "Improving the representation of legal case texts with information extraction methods," in *Proceedings of the International Conference on Artificial Intelligence and Law*, 2001, pp. 42–51, doi: 10.1145/383535.383540.
[22] M. van Opijnen and C. Santos, "On the concept of relevance in legal information retrieval," *Artificial Intelligence and Law*, vol. 25, no. 1, pp. 65–87, 2017, doi: 10.1007/s10506-017-9195-8.
[23] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria, "Automatic semantics extraction in law documents," in *Proceedings of the International Conference on Artificial Intelligence and Law*, 2005, pp. 133–140, doi: 10.1145/1165485.1165506.
[24] D. Song, A. Vold, K. Madan, and F. Schilder, "Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training," *Information Systems*, vol. 106, p. 101718, 2022, doi: 10.1016/j.is.2021.101718.
[25] S. Sharma, S. Srivastava, P. Verma, A. Verma, and S. N. Chaurasia, "A comprehensive analysis of Indian legal documents summarization techniques," *SN Computer Science*, vol. 4, no. 5, 2023, doi: 10.1007/s42979-023-01983-y.

BIOGRAPHIES OF AUTHORS

Rasmi Rani Dhala has received a M.Tech. degree in computer science and engineering and is currently pursuing Ph.D. in computer science and engineering at GIET University, Gunupur, India. Her research interests include artificial intelligence, natural language processing, machine learning, and information retrieval. She can be contacted at email: rasmi.ranidhala@giet.edu.

Akuleti Vijay S. Pavan Kumar is currently working as an associate professor at Gandhi Institute of Engineering and Technology University, Gunupur. He has a M.Tech. and Ph.D. degree in computer science and engineering. His research interest includes data mining, machine learning, and natural language processing. He has more than 18 years of teaching experience and has published various research articles in reputed journals and conferences. He can be contacted at email: avspavankumar@giet.edu.

Soumya Priyadarsini Panda is currently working as a Sr. assistant professor in the Department of Computer Science and Engineering, Silicon University, Odisha, India. She has M.Tech. and Ph.D. degree in computer science and engineering and has published more than 30 research papers in reputed journals and conferences. Her research interest includes natural language processing, speech processing, artificial intelligence, information retrieval, and machine learning. She can be contacted at email: soumya.panda@silicon.ac.in.