Journal of Electronics and Informatics (ISSN: 2582-3825)
www.irojournals.com/iroei/

‘LAWGIC’: An Application for Automated Legal Document Classification and Retrieval

Adarsh S Thambi[1], Arpitha Devangavi[2], Pankaja R[3], Aditi Joshi[4]

Department of AI&ML, BNMIT, Visvesvaraya Technological University, Bangalore, India

E-mail: [1]adarshsthambi@gmail.com, [2]arpithadevangavi@bnmit.in, [3]pankajar@bnmit.in, [4]aditi.v2702@gmail.com

Abstract

This research proposes LAWGIC (where law meets logic), an automated legal document classification and retrieval system leveraging topic modelling using Latent Dirichlet Allocation (LDA). LAWGIC utilizes the Indian Kanoon website as its data source, focusing on Supreme Court of India documents. LDA is used to extract meaningful themes from legal documents and assign them to the most relevant topic. This frees legal professionals from manual categorization, improves accuracy, and empowers them to work with greater efficiency. The system also provides a user interface for efficient retrieval and exploration of documents and topics.

Keywords: Automated Document Categorization, Information Retrieval, Legal, Latent Dirichlet Allocation (LDA), Topic Modelling

1. Introduction

The legal profession is drowning in a sea of documents. Contracts, court decisions, legal briefs, and case law accumulate, resulting in an information overload that impedes efficient and accurate legal research. Manual sorting and retrieval of these papers takes time and is prone to human error, which could jeopardize the success of judicial actions. This

DOI: https://doi.org/10.36548/jei.2024.2.007
Received: 08.04.2024, received in revised form: 04.05.2024, accepted: 25.05.2024, published: 07.06.2024

research provides a new solution: an automated legal document categorization and retrieval system based on topic modelling with Latent Dirichlet Allocation (LDA). LDA, a powerful statistical technique, works by assuming each document is a mixture of underlying topics. It analyses the words and their frequency within a document and across the entire document collection. Based on this analysis, LDA identifies a set of latent topics that best explain the distribution of words across documents. For each topic, LDA also reveals the most relevant keywords, providing valuable insight into the thematic content of the document. This approach allows us to categorize each document based on the topic it most closely resembles. This not only relieves legal professionals of the burden of manual categorization, but it also provides a standardized and accurate process, eliminating the risk of human error inherent in manual methods.

2. Problem Statement

The present manual procedures for categorizing and retrieving legal documents pose significant challenges within the legal industry. These challenges include time-consuming processes, susceptibility to errors, and overall inefficiency, resulting in a range of issues such as:

- Information Overload: The manual sorting of documents becomes a daunting task, hindering the ability to efficiently identify and extract relevant information.
- Inefficient research and delay in analysis: The time invested in sifting through extensive databases delays the analysis process, hindering prompt decision-making and strategic planning.
- Lack of standardized organization: Without a systematic categorization framework, locating specific documents becomes challenging, impeding the overall organization of legal information.

The aim is to provide an automated legal document categorizer and retrieval system that can help in addressing these challenges. We intend to specifically address two major issues:

- Automated document categorization
- Enhanced search capabilities

3. Dataset Used

The LAWGIC system leverages the "India Kanoon" website as its primary data source. This comprehensive dataset offers a rich repository of legal documents and judgments procured from various Indian courts, encompassing a diverse range of cases and legal issues. Data scraping techniques are employed to systematically extract and download legal documents from the "India Kanoon" website. Each document is downloaded and stored as a raw text file. We followed a similar way of storing the downloaded documents as they were stored in the website, i.e., categorising the documents based on the year. For this project, we have used a subset of '5341' samples from the Indian Supreme Court ranging from the years 2014 to 2024. This approach, utilizing Python libraries like 'BeautifulSoup' and requests, ensures the availability of fresh and relevant documents for analysis. Before applying topic modelling techniques, the extracted legal documents undergo a preprocessing stage to improve the quality and consistency of the data. This preprocessing step utilizes a custom Python function defined as preprocess(text). Below is the breakdown of the preprocess function:

- **Tokenization:** The function first tokenizes the text, breaking it down into individual words or meaningful units. This is achieved using the nltk.word_tokenize function from the Natural Language Toolkit (NLTK) library.
- **Stop Word Removal:** Legal documents often contain common words like "the," "a," and "of" that hold little meaning for topic modelling. The function removes these stop words using a pre-defined list from the NLTK library's stopwords.words ('English') function.
- **Lemmatization:** To capture the root meaning of words and reduce variations, the function performs lemmatization. This process converts words to their base form, for example, "running" becomes "run". The 'WordNetLemmatizer' class from NLTK is employed for this purpose, specifying the part-of-speech ("v" for verb) to ensure proper lemmatization.

- **Lowercasing and Punctuation Removal:** All words are converted to lowercase for consistency. Additionally, punctuation marks are removed as they don't contribute to the thematic content of the documents.

- **Minimum Word Length:** The function excludes words with a length of one character, as these are unlikely to be meaningful in the legal context.

By applying these preprocessing steps, the preprocess(text) function ensures cleaner and more standardized text data, ready for further analysis through topic modelling. Extracted documents and related data are stored securely following a document content storage wherein, the original content of each document is preserved, ensuring access to the complete legal text.

To facilitate efficient querying and analysis of document properties, metadata specific to documents (e.g., topics identified through LDA, keywords extracted from content, and unique document IDs) is stored in CSV format. This structured format allows for easy search and manipulation of the data using common data analysis tools. By combining data scraping techniques with secure storage solutions and a structured CSV format for document properties, LAWGIC ensures a robust and efficient foundation for legal document analysis and retrieval.

4. Related Work

Research by Mariana Y. Noguti, Eduardo Vellasques, Luiz S. Oliveira in the 2020 International Joint Conference on Neural Networks (IJCNN) employed a combination of Word2Vec trained on a legal domain corpus and a Recurrent Neural Network (RNN) architecture for this purpose. Their dataset included 17,740 legal documents from various fields obtained from the PRO-MP system [1].

Research by Sonam Gupta, Arun Kumar Yadav, Divankar Yadav, Utkarsh Dixit in the 2022 SSRN Electronic Journal performs comparison of test classification for legal document with machine learning and deep learning algorithms. They have performed text classification of legal documents with machine learning algorithms using the dataset provided by the Govt. of USA and text classification with deep learning algorithm using Brazil's supreme court dataset. Features used by them in their work include the words present in the document, labels assigned to the document and the classes the documents are assigned to [2].

Purbid Bambroo and Aditi Awasthi in their work published in the IEEE International Conference on Big data have proposed a Long DistilBERT for Legal Document Classification. The dataset they have used includes 5 types of documents publicly available at CourtListener. They have obtained an accuracy of 0.881 [3].

Rachayita Gupta, Yosha Porwal, Vaibhavi Shukla and Palak Chadha in their work published in the 2017 tenth International Conference on Contemporary Computing (IC3) have proposed approaches for information retrieval in legal documents. They have used legal documents from the website of Supreme Court of India. Methodology involved is semantic search and IPC based search. The authors state that the proposed approach achieves promising results for legal document retrieval [4].

A classification retrieval approach for English legal texts proposed by Zhonghao Li proposes an improved version of TF-IDF and SVM for text classification. His work was published in the 2019 International conference on Intelligent Transportation, Big Data and Smart City (ICITBS). An accuracy of 88% was obtained [5].

Robert Keeling, Rishi Chhatwal, Nathaniel Huber, Jianping Zhang, Fusheng Wei, Haozhen Zhao, Shi Ye, Han Qin have proposed an Empirical Comparisons of CNN with Other Learning Algorithms for Text Classification in Legal Document Review. Their work was published in the 2019 IEEE International Conference on Big Data (Big Data). Comparison of CNN, SVM, Logistic Regression(LR), Random Forest(RF) has been performed. Four data sets, from confidential, non-public, real legal matters across various industries such as social media, communications, education, and security. Their findings indicate that CNN achieved the highest precision among the four algorithms: nine times out of 16 experiments [6].

Olha Kovalchuk, Serhiy Banakh, Mariia Masonkova, Kateryna Berezka, Serhii Mokhun, Olha Fedchyshyn proposed a Text mining system for the analysis of legal texts in the 2022 12th International Conference on Advanced Computer Information Technologies (ACIT). Dataset includes 1800 text documents used in court proceedings in Ukraine from 2000 to 2020. A decision tree model using the Chi-square Automatic Interaction Detector (CHAID) growing method for classifying legal texts has been used. Classification accuracy of 92.3% was obtained [7].

Kannika Wiratchawa, Tanutcha Khunthong, Thanapong Intharah have proposed a LegalBERT-th : development of legal Q&A dataset and automated question tagging system. Their work was published in the proceedings of the 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON). They propose a law topic classification model based on BERT. Accuracy of 92% was obtained [8].

A CNN Application in detection of privileged documents in legal documents review was proposed by Rishi Chhatwal, Robert Keeling, Peter Gronvall, Nathaniel Huber, Jianping Zhang, Haozhen Zhao in the 2020 IEEE International Conference on Big Data (Big Data). Convolutional Network methodology for Text Classification has been used. The features include 1-max pooling, dropout rate and epochs parameters. Precision at different recall levels were calculated [9].

Christian Mahoney, Peter Gronvall, Nathaniel Huber, Jianping Zhang have proposed explainable text classification techniques in legal document review : Locating rationales without using human annotated training text snippets in the 2022 IEEE International Conference on Big Data (Big Data). A document-Level model Snippet model and an iterative snippet model has been used [10].

A Long-length Legal Document Classification has been proposed by Lulu Wan, Michael Seddon, George Papageorgiou, Mirko Bernardoni in the IEEE Access in 2023. This work makes use of Audio segmentation, Word2Vec, BiLSTM, Softmax classifier and SVM. An accuracy of 91.4% on the 20 Newsgroups dataset was obtained [11].

5. Proposed Work

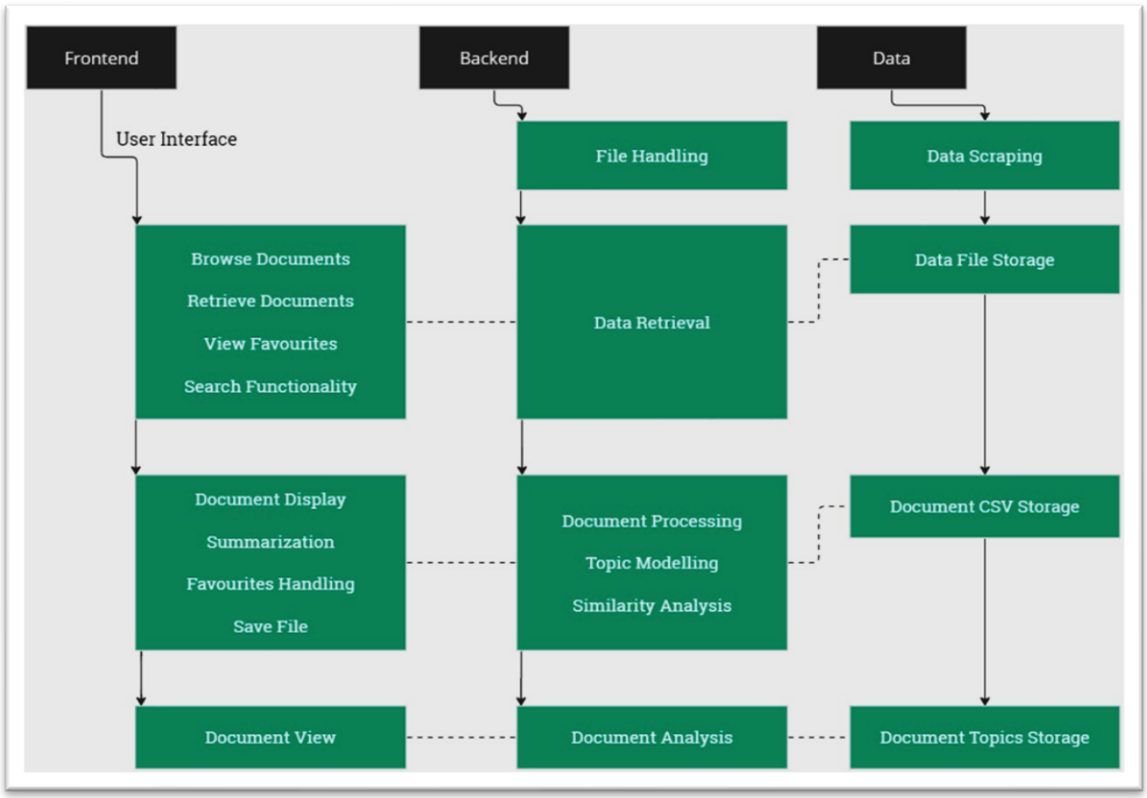

Figure 1. Architectural design of the System

Frontend Description: The frontend of the user interface provides a user-friendly way for users to interact with the document management system. It includes features such as browsing documents, retrieving specific documents, viewing favourites, and searching for documents based on keywords.

1. User Interface Features

- **Browse Documents:** Users can explore available documents, organized by categories or tags.
- **Retrieve Documents:** Users can search for specific documents based on keywords or other criteria.
- **View Favourites:** Users can access a personalized list of favourite documents for quick access.

- **Search Functionality:** Users can perform keyword-based searches across the entire document collection, and search results display relevant documents matching the query.

2. **Document Display Features**

- **Summarization:** The system generates concise summaries for each document to provide an overview of content.
- **Favourites Handling:** Users can mark specific documents as favourites for easy access.
- **Save File:** Users can download the full document for offline access and archiving.

3. **Document View:** When a user selects a specific document, the full content is displayed, allowing users to read the entire document, view attachments, and explore related sections.

Backend Description: The backend of the system handles file management, document processing, and data analysis tasks to ensure efficient document retrieval, topic modelling, and similarity analysis.

1. **File Handling**

- **Data Retrieval:** Documents are retrieved from storage efficiently when requested by users.

2. **Document Processing**

- **Topic Modelling with LDA:** The system applies Latent Dirichlet Allocation (LDA), a topic modelling technique, to analyse the content of all documents in the collection. LDA identifies underlying topics and categorizes documents accordingly.

3. **Latent Dirichlet Allocation (LDA) Usage**: LDA is a probabilistic topic modelling technique used in the backend to analyse the content of documents and identify underlying topics. Here's how LDA is applied to achieve the described tasks:

- **Building the Model:** using the pre-processed documents, we train an LDA (Latent Dirichlet Allocation) model. This model assumes that each document is a mix of various topics, and each topic is a mix of different words. Then we, decide how many topics we want to find in the documents.

- **Extracting Topics and Keywords:** after training, the LDA model will give a list of topics as shown in Figure 2. Each topic will have a set of keywords that are most likely to appear in documents related to that topic.

- **Finding Top Documents for Each Topic:** for each document, the LDA model will show how likely it is that the document belongs to each topic.

```
[(0,
  [('court', 0.020678664),
   ('case', 0.0090584345),
   ('state', 0.008164642),
   ('high', 0.00642785),
   ('appellant', 0.0062507647)])
 (1,
  [('court', 0.016069362),
   ('section', 0.009566174),
   ('act', 0.0069557154),
   ('state', 0.006858333),
   ('high', 0.0062090103)]),
 (2,
  [('court', 0.016495222),
   ('section', 0.008861837),
   ('act', 0.00858332),
   ('order', 0.0076159104),
   ('state', 0.0069846534)]),
 (3,
  [('court', 0.009772983),
   ('act', 0.008561384),
   ('state', 0.0074788267),
   ('section', 0.0072722905),
   ('order', 0.0065895864)]),
```

Figure 2. Extracted Topics and the Keywords Corresponding to Each Topic

- **Finding Similar Documents:** When we apply the LDA model to our document collection, it assigns a set of probabilities to each document, indicating the contribution of each topic. This allows us to identify which topics are more or less significant for each document. We then use the MatrixSimilarity function with these topic distributions. This function evaluates the similarity between every pair of documents in the collection based on their topic distributions using cosine similarity.

The result is a similarity matrix, a square grid where each cell represents the similarity score between two documents. Higher scores indicate greater similarity, while lower scores suggest less similarity.

$$\text{cosine\_similarity}(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Overall, LDA plays a central role in the backend processing of documents, enabling the extraction of topics, identification of relevant documents, and discovery of similar content within the document collection.

Tools and Libraries Used

- **Frontend Development**
    - Toolkit: Tkinter (Python's standard GUI library)
    - Custom Libraries: `customtkinter` for enhanced GUI elements
- **Backend Development**
    - Language: Python
    - Libraries: Sumy (for text summarization), JSON (for data storage)

6. Results and Discussion

In our proposed study, unit testing involves testing individual components or modules of the system, such as the document retrieval module, document preprocessing module, and topic modelling module. For the document preprocessing module, unit tests verify that the text preprocessing functions (e.g., tokenization, stop word removal, lemmatization) work correctly and produce the expected output for different types of legal documents. Integration tests verify that documents are retrieved correctly from the database, pre-processed appropriately, and then passed to the topic modelling module for analysis. Integration testing ensures that different parts of the system work together seamlessly, detecting any issues related to data flow, communication between components, or integration points.

As a sample, a draft document was uploaded and the fetched documents were manually checked for relevance. The similarities between the query document and the fetch documents are tabulated in Table 1. The results fetched by the application is shown in Figures 3 and 4. Accuracy testing in the research involves validating the accuracy and effectiveness of the topic modelling and document similarity analysis functionalities. Coherence: A coherence score of 0.381 suggests that the topics generated by the model are moderately coherent. This means that the words within each topic are related to each other but might not be as tightly related or distinct as desired. Perplexity: Perplexity measures how well a probability model predicts a sample. In standard usage, lower perplexity values indicate better performance, suggesting that the model is better at predicting the test dataset.

Table 1. Test Case

Test Cases	Fetched Document 1	Fetched Document 2	Results (Similarity Score)
Test document 1 : Sajjad Quraishi vs Anwarul Haq Quraishi and Others on 30 January, 2020	Title : State Of Punjab vs Gurmit Singh on 2 July, 2014: Nature of Offense: In both cases, the allegations involve criminal offenses. Summoning Orders: In both cases, the trial courts issued summoning orders against the accused/respondent based on the evidence presented during the proceedings. High Court Intervention: Both cases involve intervention by a High Court. Appeals: Appeals were filed in both cases to challenge the decisions made by the High Courts. Abuse of Process: In both cases, there are allegations of the abuse of the legal process.	Madhya Pradesh vs Deepak on 13 March, 2019: Legal Context: Both documents pertain to legal cases involving criminal proceedings. Court Analysis: In both documents, the judges provide detailed analysis and interpretation of relevant laws, citing previous judgments and legal principles. Appellate Proceedings: Both documents involve appeals filed against decisions made at lower court levels. Final Verdict: In both cases, the appellate court overturns the decisions of the lower courts.	Document 1 : 0.999 Document 2: 0.998

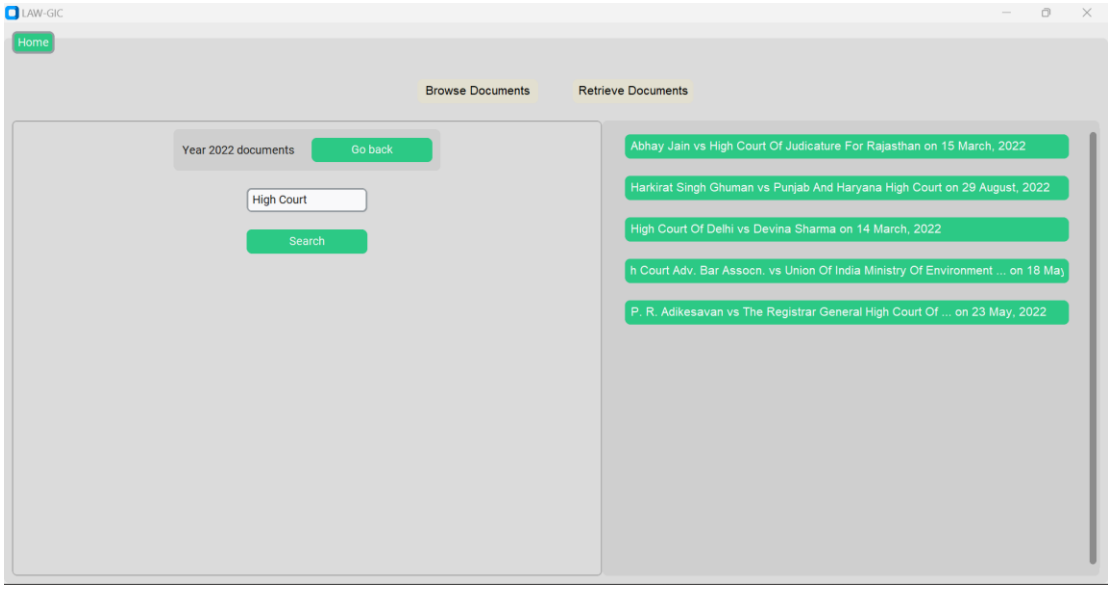

Figure 3. Browse Documents with Search Functionality

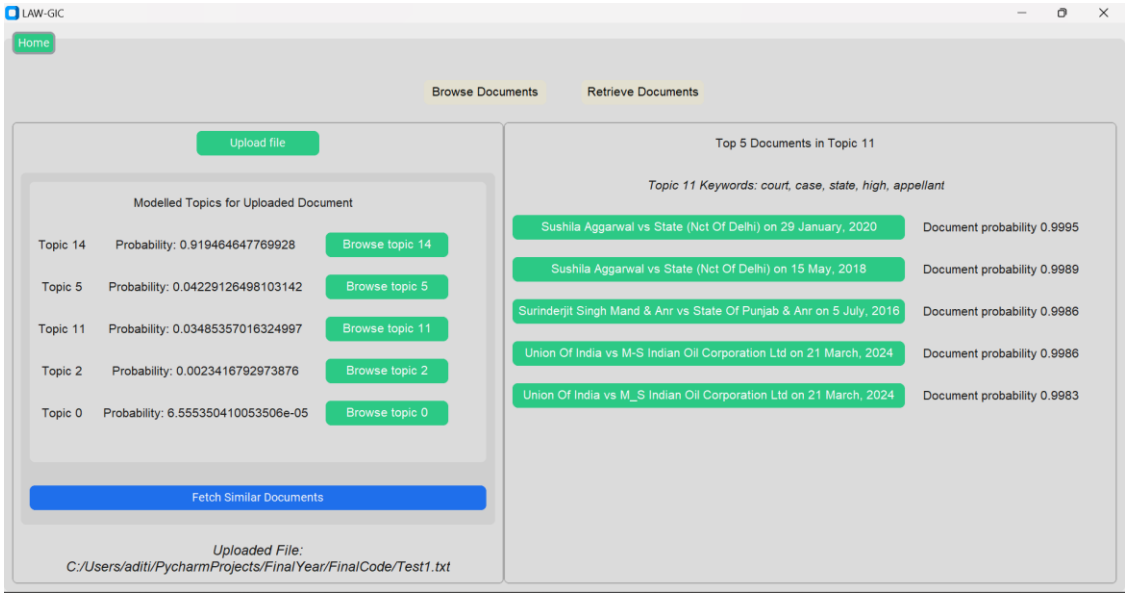

Figure 4. Similar Documents and Topic Retrieval

7. Conclusion

This research has successfully developed an automated system for legal document categorization and retrieval using Latent Dirichlet Allocation (LDA) for topic modelling. The

system addresses the challenge of managing large volumes of legal documents by automating the categorization and retrieval processes, thereby enhancing the efficiency and accuracy of legal research. the evaluations indicate that the system performs robustly with a moderate topic coherence and a low perplexity score, demonstrating its capability to organize and process legal texts effectively. The implementation of this system offers significant advantages over traditional manual methods, including reduced time for document processing and improved accuracy in document retrieval, which mitigates the risk of human error. The system's ability to quickly and accurately retrieve relevant documents has the potential to transform legal research, making it more efficient and less labour-intensive. .

References

[1] Noguti, Mariana Y., Eduardo Vellasques, and Luiz S. Oliveira. "Legal document classification: An application to law area prediction of petitions to public prosecution service." In 2020 International joint conference on neural networks (IJCNN), pp. 1-8. IEEE, 2020.

[2] Gupta, Sonam, Arun Yadav, Divakar Yadav, and Utkarsh Dixit. "Analysis of Automatic Text Classification of Legal Documents." In Proceedings of the International Conference on Innovative Computing & Communication (ICICC). 2022.

[3] Bambroo, Purbid, and Aditi Awasthi. "Legaldb: Long distilbert for legal document classification." In 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1-4. IEEE, 2021.

[4] Giri, Rachayita, Yosha Porwal, Vaibhavi Shukla, Palak Chadha, and Rishabh Kaushal. "Approaches for information retrieval in legal documents." In 2017 Tenth International Conference on Contemporary Computing (IC3), pp. 1-6. IEEE, 2017.

[5] Li, Zhonghao. "A classification retrieval approach for English legal texts." In 2019 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), pp. 220-223. IEEE, 2019.

[6] Huber-Fliflet, Nathaniel, Jianping Zhang, Fusheng Wei, Q. Han, Shi Ye, and H. Zhao. "Empirical Comparisons of CNN with Other Learning Algorithms for Text

Classification in Legal Document Review." In 2019 IEEE International Big Data Conference. 2019.

[7] Kovalchuk, Olha, Serhiy Banakh, Mariia Masonkova, Kateryna Berezka, Serhii Mokhun, and Olha Fedchyshyn. "Text mining for the analysis of legal texts." In 2022 12th International Conference on Advanced Computer Information Technologies (ACIT), pp. 502-505. IEEE, 2022.

[8] Wiratchawa, Kannika, Tanutcha Khunthong, and Thanapong Intharah. "LegalBERT-th: Development of Legal Q&A Dataset and Automatic Question Tagging." In 2021 18th International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), pp. 1159-1162. IEEE, 2021.

[9] Chhatwal, Rishi, Robert Keeling, Peter Gronvall, Nathaniel Huber-Fliflet, Jianping Zhang, and Haozhen Zhao. "CNN application in detection of privileged documents in legal document review." In 2020 IEEE international conference on big data (big data), pp. 1485-1492. IEEE, 2020.

[10] Mahoney, Christian, Peter Gronvall, Nathaniel Huber-Fliflet, and Jianping Zhang. "Explainable Text Classification Techniques in Legal Document Review: Locating Rationales without Using Human Annotated Training Text Snippets." In 2022 IEEE International Conference on Big Data (Big Data), pp. 2044-2051. IEEE, 2022.

[11] Wan, Lulu, George Papageorgiou, Michael Seddon, and Mirko Bernardoni. "Long-length legal document classification." arXiv preprint arXiv:1912.06905 (2019).