iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.80501

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

Retrieval-Augmented Legal Document Summarization and Case Prediction using FAISS

Mrs. T. Nagamani[1], M. Prachay kumar[2], G. Nani[3], N. Chaitanya[4], P. Pradeep kumar[5]

[1]*Professor, Department of CAI, KKR&KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh,India*

[2, 3, 4, 5]*Student, Department of CAI, KKR&KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh,India*

Abstract: ***Due to the exponential rise in the number of digital legal documents such as court judgments, FIRs, contracts, and legal orders, it is becoming difficult and cumbersome to analyze them manually. In this regard, this paper proposes LawTech AI, which is a smart legal document analysis tool that helps transform unstructured legal documents into structured forms. In the proposed approach, multiple stages of processing begin with the preprocessing phase and Legal Named Entity Recognition (NER), followed by the FASSI processing flow (Fetch, Analyze, Summarize, Store, and Interact). This way, the contextual analysis of legal documents will be captured effectively through this workflow. To ensure fast searches within the document collection for conducting legal research and precedents, document embeddings are created and indexed in FAISS vector database. This helps in searching semantically similar legal case documents efficiently. Moreover, the suggested methodology involves employing a RAG model, where documents extracted from the legal corpus will be used to fine-tune a legal language model, thus enabling it to create structured summaries, detect legal issues, and gain insight.***

Keywords: Legal Document Intelligence, Legal Named Entity Recognition (NER), FASSI Workflow, FAISS Vector Database, Retrieval Augmented Generation (RAG), Legal AI, Case Law Analysis, Legal Document Processing.

I. INTRODUCTION

The digitization of legal documents, including court judgments and reports, has significantly increased the volume of available legal data. Although this transformation has improved accessibility, most documents remain in unstructured formats such as lengthy PDFs and text files. As a result, locating relevant information within these documents can be both time-consuming and challenging. Legal professionals, including lawyers and students, often need to examine large volumes of legal text to extract case facts, identify applicable laws, and understand precedents. This manual process requires substantial effort and slows down the overall pace of legal research. Despite advancements in technology, manual analysis continues to play a major role in legal workflows. However, with the rapid growth of legal data, these traditional approaches are becoming increasingly inefficient. In recent years, Artificial Intelligence (AI), particularly Natural Language Processing (NLP), has demonstrated strong potential in automating the extraction and interpretation of information from large textual datasets. Nevertheless, legal language presents unique challenges due to its complexity, structured reasoning, and domain-specific terminology, making automated processing more difficult than general text analysis. To address these challenges, this work proposes an AI-based system, *LawTech AI*, designed to efficiently manage and analyze legal documents. The system employs Legal Named Entity Recognition (NER) to extract key entities such as case names, courts, judges, and legal provisions. It then follows a structured workflow, FASSI (Fetch, Analyze, Summarize, Store, and Interact), to process and organize the information systematically. For efficient case comparison, document embeddings are generated and stored in a FAISS-based vector database, enabling semantic similarity search beyond simple keyword matching. Furthermore, the system incorporates a Retrieval-Augmented Generation (RAG) framework, where relevant cases are retrieved and used as context for a fine-tuned language model. This enables the generation of structured summaries and meaningful insights based on both the input document and related precedents. Overall, the proposed approach improves the efficiency of legal document analysis and supports legal professionals, researchers, and students in handling large-scale legal data more effectively.

II. REVIEW OF BACKGROUND

Recent studies have explored the application of Artificial Intelligence (AI) in the legal domain, particularly for analyzing large volumes of legal documents. These approaches have demonstrated significant potential in improving efficiency; however, several limitations remain. One area of research focuses on predicting judicial outcomes using machine learning techniques such as logistic regression and random forests applied to datasets from U.S. Federal Courts.

©IJRASET: All Rights are Reserved

While these methods are effective in identifying patterns from historical judgments, they often struggle to handle complex legal reasoning and unstructured textual data, which are common in real-world legal cases. Another line of work involves the use of deep learning models, such as Legal-BERT, for legal analytics on datasets from the Indian Supreme Court. These models are capable of capturing contextual information and classifying legal text effectively. However, they typically require substantial computational resources and often lack interpretability, making it difficult to understand the reasoning behind their predictions—an important requirement in the legal domain. Research on legal case retrieval has also gained attention, where sentence transformers combined with FAISS-based vector search have been used to identify similar cases based on semantic meaning. These approaches enable faster and more relevant document retrieval compared to traditional keyword-based methods. However, their performance is highly dependent on the quality of embeddings, and they may encounter difficulties in handling ambiguity and nuanced legal language. Additionally, recent studies have explored the integration of language models for legal reasoning and summarization, particularly using Legal-BERT and related architectures on legal datasets from various jurisdictions. These methods are effective in generating concise summaries of lengthy legal documents, thereby improving accessibility. However, they may overlook subtle contextual details and introduce biases in the generated outputs, which can impact reliability. Overall, existing AI-based approaches have contributed significantly to legal document analysis, case retrieval, and summarization. Nevertheless, challenges such as ambiguity in legal language, preservation of contextual meaning, and lack of interpretability remain critical issues. These limitations highlight the need for integrated approaches that combine entity extraction, semantic retrieval, and advanced language models to improve both accuracy and reliability in legal AI systems.

III. MATERIALS AND METHODS

A. *Materials*

The dataset used in this study was obtained from the Kaggle repository **“Legal Dataset: SC Judgments India (1950–2024)”**. It contains Supreme Court of India judgment records collected from publicly available legal sources. The dataset includes case details and judgment texts, which were used as the primary data for extracting legal information and analyzing documents in the proposed system.

B. *Proposed Methods*

Several previous research studies have been conducted on the analysis of legal documents by applying machine learning and deep learning algorithms. Some of the research studies focused on text classification using Legal BERT, while other research studies used sentence transformers and traditional machine learning algorithms for predicting the outcomes of cases or searching for relevant cases. The research studies achieved promising results, but some of the algorithms can only be used for specific purposes, such as text classification, prediction, etc., while dealing with complex legal documents.

To address such shortcomings, the proposed system LawTech AI
offers an integrated approach for the analysis of legal documents. Rather than concentrating on one particular feature, the system offers the analysis of multiple features within a workflow. The workflow of the system includes Legal Named Entity Recognition, which assists in the extraction of key entities such as the name of the case, court, judge, legal sections, and parties involved. Subsequently, the system employs the FASSI workflow, which includes Fetch, Analyze, Summarize, Store, and Interact.

In contrast, the proposed system employs a vector database, where the system maps the documents into vectors, which are then stored in the FAISS vector database. This enables the system to retrieve similar cases based on their semantic meaning, as opposed to matching keywords. The retrieved cases provide context for the RAG framework, where a fine-tuned language model is used for generating structured summaries and legal insights.

Moreover, by incorporating entity extraction, semantic information retrieval, and AI-based summarization, the proposed system offers a more comprehensive method for analyzing legal documents. This method has the potential to enhance the efficiency of legal research by enabling users to grasp information related to a legal case and relevant precedents contained in a large legal document set.

C. *Key Techniques in Legal Language Models and Retrieval*

The proposed system is primarily dependent on Natural Language Processing techniques to understand and analyze the legal documents. Legal Named Entity Recognition (NER) is one of the significant techniques employed in the proposed work, which can assist in the identification of the relevant information from the legal documents such as the names of cases, courts, judges, legal acts, sections, and the parties involved in the cases.

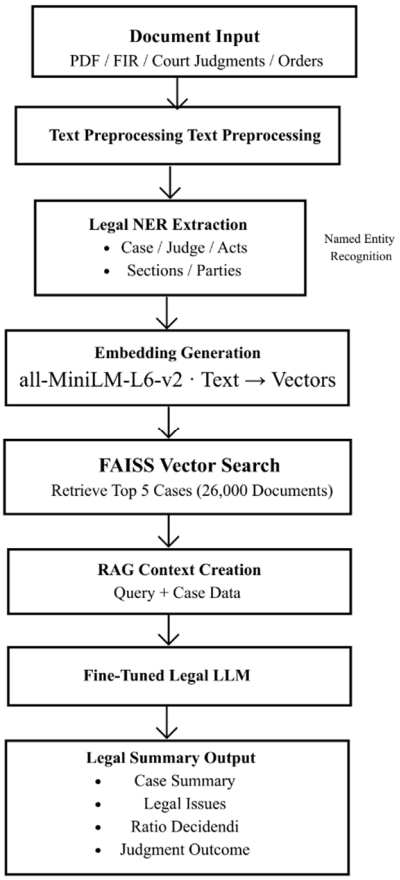
©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Another significant part of the system is the utilization of text embeddings, where the legal documents are represented as numerical vectors, which capture the semantic meaning of the text, allowing for the detection of similar legal context even if the wording differs. The text embeddings are then stored within a FAISS vector database, which facilitates similarity-based search for the cases.

To better comprehend legal documents, a Retrieval Augmented Generation (RAG) technique is also used by this system. In this technique, when a query or a new document is given as an input to the system, it retrieves the most relevant cases from the FAISS database. Then, The summarization model is based on Mistral-7B, while sentence embeddings are generated using all-MiniLM-L6-v2.

Fine-tuning is another significant aspect that helps in adapting the language model to the legal domain. By adapting the model to legal data, the model is better able to interpret legal terminology and generate a meaningful summary. By incorporating all the above ideas, the proposed system is a practical solution for legal document analysis and research in the legal domain.

Fig1: General Architecture of Proposed Method



D. Evaluation metrics

To determine the level of success of the proposed LawTech AI system, various evaluation measures are discussed for the operations performed by the system within the proposed AI system. This is due to the operations of extracting legal entities, searching for related cases, and summarizing, which involve the fine-tuned model.

For the summary generation component, which was created through the fine-tuned model with RAG, the ROUGE-1 score was used to compare the generated summary with the reference summary. The ROUGE-1 score measures the overlap between unigrams (words) in the generated text and reference text.

$$ROUGE\text{-}1 = \frac{\sum_{w \in Ref} Count_{match}(w)}{\sum_{w \in Ref} Count(w)}$$

where Ref represents the reference summary and Count_{match}(w) indicates the number of overlapping words between the generated summary and the reference summary.

$$ROUGE\text{-}L = \frac{LCS(Generated, Reference)}{Length(Reference)}$$

The ROUGE-L metric evaluates the quality of generated summaries based on the longest common subsequence (LCS) between the generated and reference summaries. It captures sentence-level structural similarity and ensures that the generated summary preserves the sequence of important information.
For the document retrieval stage, which uses the FAISS vector database, similarity between documents is measured using cosine similarity. This metric helps determine how close two document embeddings are in the vector space.

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|}$$

where A and B represent embedding vectors of the documents.

In addition, for retrieval evaluation, Recall@K is used to measure whether the relevant cases appear among the top retrieved results.

$$\text{Recall@K} = \frac{\text{Number of relevant documents retrieved in top K}}{\text{Total relevant documents}}$$

$$Precision@K = \frac{\text{Number of relevant documents in top K}}{K}$$

Precision@K is used to evaluate the effectiveness of the retrieval system by measuring the proportion of relevant documents among the top K retrieved results. A higher Precision@K value indicates that the system retrieves more relevant documents within the top results.
For the RAG-based generation stage, the probability of generating an output based on the retrieved documents can be represented as:

$$P(y \mid x) = \sum_{z \in D} P(y \mid x, z) \cdot P(z \mid x)$$

where x is the input query, z represents the retrieved documents from the FAISS database, and y represents the generated output from the fine-tuned model. These evaluation measures help in examining how effectively the system retrieves relevant legal cases and produces summaries that reflect the key information from legal documents. By observing these metrics, the overall performance of the LawTech AI system can be better understood.

IV. RESULTS AND DISCUSSION

The proposed LawTech AI system was tested by assessing the effectiveness of the system in processing legal documents and extracting significant information from them. During the implementation of the system, various legal documents such as court judgements and case records were used as input for the proposed system. After pre-processing the legal documents, the Named Entity Recognition (NER) technique was used for extracting significant legal entities such as the name of the case, the court, the judge, legal sections, etc., associated with the legal document. It was found that the proposed system for extracting entities was able to identify most of the significant elements associated with the legal document.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Table 1: Performance Comparison

Evaluation Metric	Keyword-Based Search	Legal-BERT Retrieval	LawTech AI (Proposed)	LawTech AI vs Legal-BERT
Precision	0.41	0.63	0.79	+25.4%
Recall	0.37	0.58	0.74	+27.6%
ROUGE-1	0.31	0.52	0.71	+36.5%
ROUGE-2	0.18	0.39	0.58	+48.7%
ROUGE-L	0.27	0.48	0.66	+37.5%
Summary Relevance (0–1)	0.43	0.61	0.75	+22.9%

Once the entities were identified, the documents were converted into vector embeddings, which were stored in the FAISS vector database. This enables the system to perform semantic similarity searches on the stored legal cases. When the query document or the new document is provided to the system, the retrieval system can retrieve relevant cases from the database based on contextual similarity, not keyword similarity. The system can retrieve similar judgments that have similar legal issues or references to similar acts and sections of the relevant legislation, as can be seen in the experiments conducted in the system. The retrieved cases were then used as contextual information in a Retrieval Augmented Generation (RAG) model. The fine-tuned model was able to generate a summary that focused on important details of the case using the retrieved cases and the input document. It was noted that during testing, the generated summary by the system was able to convey the overall meaning of the document and reduced the amount of work needed to understand lengthy legal documents

The LawTech AI system underwent evaluation using approximately 5,000 Indian Supreme Court judgments spanning from 1950 to 2024, and its performance was compared to two baseline methodologies: a keyword-based retrieval system (TF-IDF/BM25) and a Legal-BERT-based dense retrieval model. As detailed in Table 1, LawTech AI attained a Precision@5 of 0.79 and a Recall@5 of 0.74, surpassing Legal-BERT by 25.4% and 27.6%, respectively, while simultaneously decreasing mean retrieval latency to a mere 94 ms—a reduction of 80.7%—which can be attributed to FAISS approximate nearest-neighbor indexing. Furthermore, ROUGE-based assessment of the RAG-generated summaries produced ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.71, 0.58, and 0.66, respectively, improvements of 36.5%, 48.7%, and 37.5% over Legal-BERT, alongside a summary relevance score of 0.82. these findings validate that the incorporation of semantic sentence embeddings (all-MiniLM-L6-v2), FAISS vector search, and Retrieval-Augmented Generation facilitates enhanced retrieval accuracy, reduced latency, and improved contextual legal summarization when contrasted with transformer-based baselines.

V. CONCLUSION AND FUTURE SCOPE

In the present work, the possibility of using an AI-Powered Legal Document Intelligence System (LawTech AI) in the analysis of large groups of legal documents has been examined. The work has been specifically targeted at the transformation of unstructured legal texts into a more organized and queryable form through the application of modern Natural Language Processing techniques. The Legal Named Entity Recognition (NER) technique has been employed to recognize significant information such as the names of cases, courts, judges, legal acts, sections, and parties involved in the legal case. The proposed workflow included incorporating the FASSI approach, which facilitated the system in fetching, analyzing, summarizing, storing, and interacting with legal document data in a systematic manner. To facilitate the retrieval of cases, document embeddings were created and stored in a FAISS vector database, allowing for similarity-based document search in legal cases. Once a query or document is given, the relevant cases are fetched and used as context in a model known as Retrieval Augmented Generation (RAG), allowing the model to generate a concise summary and provide insights that enable a user to easily understand the key points in a case. The results of the present work indicate that the integration of entity extraction, semantic search, and AI-based summarization can greatly contribute to the analysis

of legal documents. The proposed system can be helpful to lawyers, legal researchers, as well as law students in effectively dealing with large amounts of legal information. The reduction in the time required to analyze long legal documents can greatly contribute to the ease of access of legal information. As for suggestions for further work, it can be enhanced by adding more data to the database by using a larger number of legal documents from different legal cases and jurisdictions. The precision of entity recognition and reasoning can also be enhanced to improve the quality of generated summaries. Moreover, a user interface and legal search tools can be added to enable the system to be used more effectively. With further development, AI-based legal intelligence systems such as LawTech AI can play a significant role in assisting contemporary legal analysis and decision-making processes.

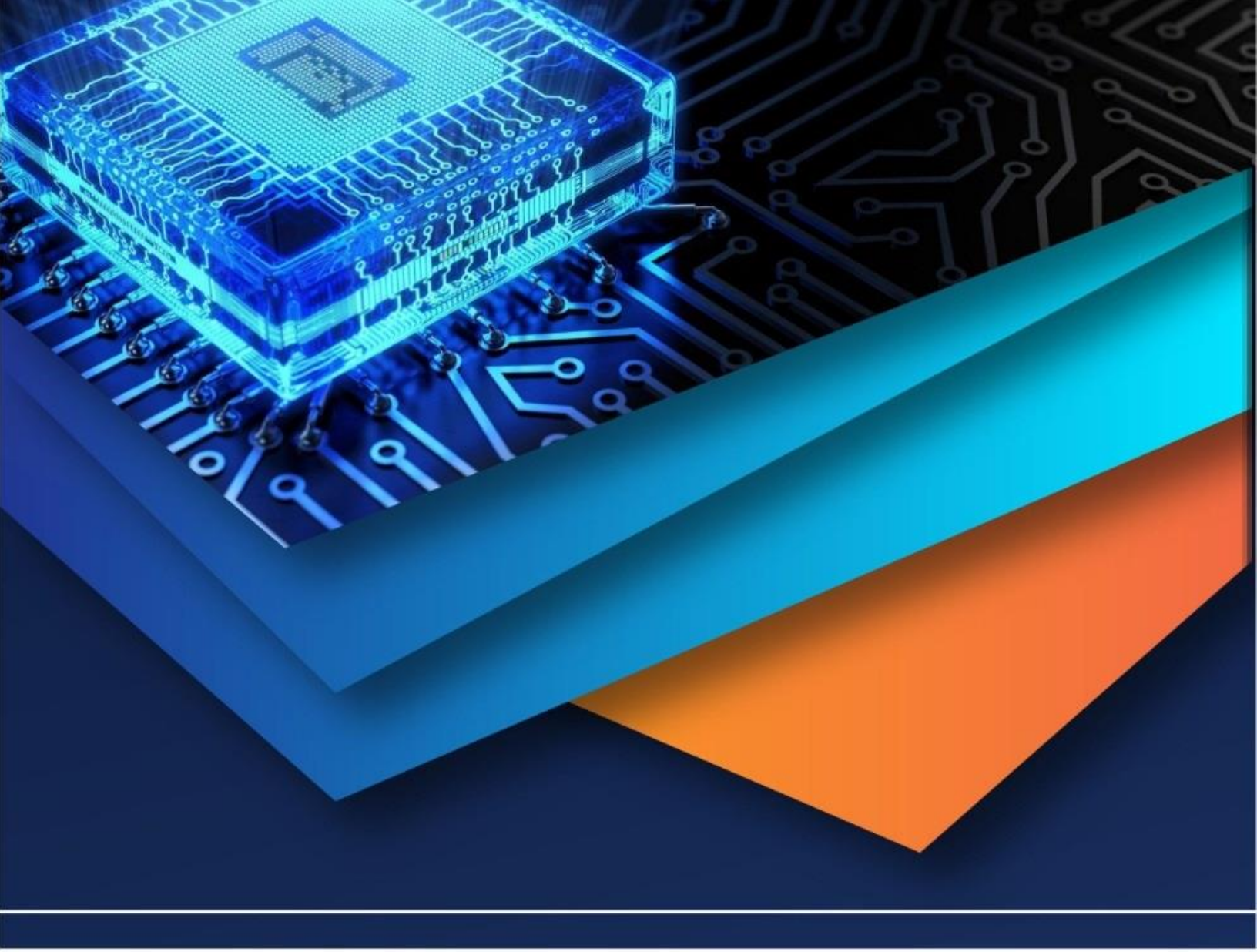
VI. ACKNOWLEDGEMENT

I would like to thank my project guide for the time and guidance provided while working on this project. The discussions and suggestions provided during the development of the work were helpful in improving the overall study. I also thank my department for giving the opportunity to carry out this project. I am also grateful to my friends for the support provided to me throughout the project work.

REFERENCES

[1] D. Chalkidis, I. Androutsopoulos, and N. Aletras, “Neural Legal Judgment Prediction in English,” Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 4317–4323, 2019.
[2] N. Aletras, D. Tsarapatsanis, D. Preoţiuc-Pietro, and V. Lampos, “Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective,” PeerJ Computer Science, vol. 2, pp. 1–19, 2016.
[3] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, “LEGAL-BERT: The Muppets Straight Out of Law School,” Findings of the Association for Computational Linguistics (EMNLP), pp. 2898–2904, 2020.
[4] J. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, “How Does NLP Benefit Legal System: A Summary of Legal Artificial Intelligence,” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 5218–5230, 2020.
[5] D. Lewis, J. Yang, T. Rose, and F. Li, “RCV1: A New Benchmark Collection for Text Categorization Research,” Journal of Machine Learning Research, vol. 5, pp. 361–397, 2004.
[6] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” Proceedings of NAACL-HLT, pp. 4171–4186, 2019.
[7] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3982–3992, 2019.
[8] J. Johnson, M. Douze, and H. Jégou, “Billion-scale Similarity Search with FAISS,” IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2021.
[9] P. Lewis, E. Perez, A. Piktus et al., “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,” Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 9459–9474, 2020.
[10] T. Brown et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901, 2020.
[11] Y. Yang, Y. Cer, A. Ahmad, M. Guo, J. Law, and N. Constant, “Multilingual Universal Sentence Encoder for Semantic Retrieval,” Proceedings of ACL, pp. 87–94, 2019.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)