Journals Pub

International Journal of
Software Computing and Testing

ISSN: 2456-2351
Volume 11, Issue 1, 2025
January–June
DOI (Journal): 10.37628/IJSCT

https://journalspub.com/journal/ijsct

Review IJSCT

Legal Ease: An End-to-End Automated Legal Document Processing System Using LLMs and OCR

Bhasha Sinha[1], Shrishti Saxena[2,*], Anuva Vashishtha[3], Aradhna Aggarwal[4], Rajesh Yadav[5]

Abstract

This paper describes an automated legal document processing system that uses modern technologies, such as Large Language Models (LLMs) and Optical Character Recognition (OCR). Designed to streamline legal procedures, the system allows users to submit legal documents in PDF format and uses Tesseract-powered OCR to extract and digitize text while keeping the document's layout and structure. The digitized text is processed by an LLM that has been fine-tuned using legal datasets to ensure that key legal nuances are retained. The system provides a variety of functionalities, including full-document summarization, clause-specific summarization, and targeted clause extraction, to meet a wide range of legal purposes. With its user-centric design, the React is interface offers an interactive and customizable experience, allowing users to select specific areas for summarization, highlight critical bits, and fine tune results to meet their needs. The results are presented in a structured way to ensure clarity and practical insights. The backend uses FastAPI for easy processing and PyTorch for model implementation, which ensures robustness and scalability. Furthermore, tools like NLTK help with natural language processing tasks like text analysis and structure. To improve user ease, the system incorporates Google Drive for automatic session management, allowing users to review or edit previously processed documents. This system revolutionizes legal document automation by merging cutting-edge technology, such as Tesseract OCR, PyTorch, and finely designed LLMs. It benefits legal professionals by reducing time, increasing efficiency, and ensuring the correctness and context of legal documents. This system is a vital tool for modern legal offices, meeting the growing demand for intelligent and efficient document processing solutions.

Keywords: Artificial Intelligence (AI) in Law, Document Summarization, Large Language Models (LLMs), Legal Document Automation, Legal Technology, Natural Language Processing (NLP), Optical Character Recognition (OCR)

INTRODUCTION

Dealing with lengthy and intricate documents, like contracts, case summaries, and regulatory texts, presents enormous hurdles for legal professionals. The legal industry, which is notorious for its need for accuracy and deadline-driven processes, frequently faces inefficiencies brought on by manual document processing. Summarizing legal case judgments is a complex task in Legal Natural Language Processing (NLP), with a gap in understanding how various summarization models include extractive and abstractive approaches and analyzing the performance within the domain of legal documents [1]. New approaches to tackling these issues have been made possible by developments in artificial intelligence (AI). Text can be extracted from scanned or image-based legal documents using optical character recognition (OCR) technology like Tesseract OCR.

***Author for Correspondence**
Shrishti Saxena.
E-mail: shrishtisaxena22.set@modyuniversity.ac.in

[1–4]Student, Department of Computer Science & Engineering, Mody University of Science & Technology, Laxmangarh, Rajasthan, India
[5]Associate Professor, Department of Computer Science & Engineering, Mody University of Science & Technology, Laxmangarh, Rajasthan, India

Received Date: November 25, 2024
Accepted Date: November 29, 2024
Published Date: April 15, 2025

Citation: Bhasha Sinha, Shrishti Saxena, Anuva Vashishtha, Aradhna Aggarwal, Rajesh Yadav. Legal Ease: An End-to-End Automated Legal Document Processing System Using LLMs and OCR. International Journal of Software Computing and Testing. 2025; 11(1): 29–32p.

However, when trained on domain-specific datasets, Large Language Models (LLMs) can interpret and summarize legal writings with unprecedented precision. Additionally, legal texts are often extensive, further increasing the summarization task's complexity [2].

These state-of-the-art technologies are seamlessly integrated into the suggested system. A web-based frontend interface created with CSS for styling and Reactjs for dynamic rendering allows users to submit legal documents in PDF format. The text is digitized using OCR technology, which maintains the formatting and legal context. To provide summaries or extract clauses as needed, this text is processed by a backend that is driven by an LLM, optimized with PyTorch, and trained on carefully selected legal datasets. The front end displays the condensed content, allowing users to interact with it or download it for later use.

Scalability and robustness are guaranteed by the tools utilized in this project, which include Google Colab, NLTK for natural language processing, FastAPI for backend support, Pandas for effective data handling, and Google Drive for data storage [3]. This technology improves accuracy and compliance while freeing up legal experts to concentrate on higher-value work by automating legal document processing. This integrated approach is a crucial innovation for modernizing legal procedures since it fills a significant vacuum in legal technology.

AI-based approaches can extract useful information from unstructured documents automatically [4]. Legal teams may accomplish these goals with this technology while making sure that important legal details are maintained in the automated process.

RELATED WORK

- *Text Extraction Using OCR:* OCR technologies are frequently used in many different industries to digitize documents. Particularly, Tesseract OCR has proven to be highly accurate at extracting text from complicated document formats and low-resolution scanned images. However, there are several situations where the data is not digitized, and it might become essential to extract text from those to store in digitized form [5].
- *Legal Summarization Using LLMs:* In recent years, there has been a notable increase in the use of large language models in legal technology. As LLM platforms, such as ChatGPT, Gemini and LAMDA become accessible to the public, everyday citizens may come to rely on these platforms for legal advice [6]. Nonetheless, the study pointed out that attaining superior outcomes in the legal field requires domain-specific fine-tuning.
- *Using AI in Legal Technology Integration:* Fully integrated solutions for legal workflows are still uncommon, despite the widespread use of standalone OCR and NLP tools. Despite decades of research and the existence of established commercial products, the output from optical character recognition (OCR) processes often contains errors [7]. The suggested approach fills these gaps by providing a user-friendly GUI.
- *Legal Tools Focused on the User:* The significance of user-friendly interfaces in legal technology was emphasized by research in [8]. According to the study, because they reduce the learning curve for non-technical users, tools with intuitive designs are more likely to be adopted. This idea served as a guide for creating the frontend of the suggested system, guaranteeing that legal experts could easily use it.

CHALLENGES

There are various obstacles to overcome while putting in place an automated system for processing legal documents.

Legal Document OCR Accuracy

Multiple fonts, handwritten annotations, and poor text quality are common features of scanned legal documents. To overcome these obstacles, preprocessing methods, such as noise reduction and image enhancement must be used, however they can increase computing overhead [9].

Adjusting LLMs

To preserve domain-specific subtleties, LLMs must be trained on legal datasets. However, it takes a lot of resources to gather high-quality annotated legal data, and fine-tuning calls for a large amount of computational power [10].

SUMMARY

Processing legal documents could be completely transformed by combining OCR and LLM technologies. The suggested solution increases workflow efficiency and decreases manual labor by automating the extraction, summary, and analysis of legal documents. Modern tools like PyTorch, FastAPI, React.js, and Tesseract OCR are used in the system to provide scalability, robustness, and interaction.

This initiative adds value to the legal technology landscape by bridging the gap between document digitization and actionable insights. Its interactive features and React.js-based interface further improve accessibility, making it easy for legal professionals to use.

All things considered, this approach can greatly cut down on the time and effort needed to process legal documents, freeing up legal practitioners to concentrate on more important work. The suggested method represents a major advancement in legal technology going forward, with the potential to improve document accuracy and expedite legal procedures.

FUTURE SCOPE

Despite its great effectiveness, the suggested method can be improved and expanded in several ways to meet a wider range of legal document processing requirements:

Support for Multiple Languages

Adding multilingual capability will allow the system to process legal papers in multiple languages, which is advantageous given the worldwide nature of legal documents. This might be accomplished by teaching the LLM to comprehend and summarize documents in numerous languages and optimizing OCR models to recognize various language scripts. International law firms or legal teams operating across jurisdictions would especially benefit from this improvement.

Collaboration in Real Time

Working as a team is crucial in the legal profession. Several legal experts might collaborate on a single document at once if real-time collaboration elements were integrated. By enabling users to annotate, highlight, and comment on specific document portions, this feature may improve communication and expedite document review. Furthermore, regardless matter where they are, all users would always have access to the most recent versions of the document thanks to the integration of cloud storage systems.

Risk Analysis Driven by AI

Including AI-powered risk analysis is another possible improvement. Technology might find possible dangers or conflicts in contracts, like unclear phrases or provisions that could result in legal problems, by examining patterns in the legal papers. By addressing problems before they become more serious, this proactive strategy may assist legal teams in lowering the likelihood of litigation or compliance problems.

Adaptive Education and Ongoing Development

The system may eventually include adaptive learning features that enhance its functionality in response to user input and fresh legal developments. To ensure that the system stays current with evolving legal vocabulary or document formats, for example, the LLM might be adjusted on a regular basis based on user feedback. Additionally, by incorporating user-specific learning, the system might enhance its extraction and summarization capabilities according to each user's unique requirements and preferences.

Combining Other Legal Resources

A complete solution for legal professionals would be offered by integrating the system with other popular legal tools, such as case management software, e-discovery tools, or legal research databases. Users will be able to handle all their legal responsibilities on a single platform, significantly improving workflow efficiency, by facilitating smooth data movement between systems.

Improved Clause Extraction and Identification

The system might be improved to better recognize and categorize legal clauses by expanding on the current clause extraction capabilities. The technology might give customers more precise control by categorizing phrases by kind (such as termination, secrecy, and indemnity) using sophisticated NLP approaches.

CONCLUSIONS

The integration of OCR and LLMs revolutionizes legal document processing, automating tedious tasks and enhancing efficiency. The user-friendly React.js interface and robust backend infrastructure, powered by PyTorch and FastAPI, ensure seamless operation and scalability.

While this system offers significant advancements, future developments can further optimize its capabilities. Potential areas of improvement include multilingual support, real-time collaboration, AI-driven risk analysis, blockchain integration, and adaptive learning. By addressing these areas, the system can become an indispensable tool for legal professionals, streamlining workflows and improving accuracy.

REFERENCES

1. Kumar H, Jayanth P. Large Language Models for Indian Legal Text Summarisation. In 2024 IEEE Int Conf Electron, Compute Commune Technol. 2024 Jul 12:1–5.
2. Preti D, Giannone C, Favalli A, Romagnoli R. Automatic Summarization of Legal Texts, Extractive Summarization using LLMs. Ital-IA 2024: 4th National Conference on Artificial Intelligence, organized by CINI, May 29–30, 2024, Naples, Italy. Available from https://ceur-ws.org/Vol-3762/501.pdf.
3. Millstein F. Natural language processing with python: Natural language processing using NLTK. Frank Millstein; 2020 Jul 6.
4. Kolandaisamy R, Rajagopal H, Kolandaisamy I, Sinnappan GS. The Smart Document Processing with Artificial Intelligence. The 2024 Int Conf Artif Life and Robot J: Com HorutoHall, Oita, Japan; 2024. 534–540p.
5. Mittal R, Garg A. Text extraction using OCR: A systematic review. In 2020, The Second Int Conf Inven Res Comput Appl. IEEE. 2020 Jul 15:357–362.
6. Krook J, Schneiders E, Seabrooke T, Leesakul N, Clos J. Large Language Models (LLMs) for Legal Advice: A Scoping Review. 2024 Oct 4. Available at SSRN 4976189.
7. Lopresti D. Optical character recognition errors and their effects on natural language processing. In Proceedings of the second workshop on Analytics for Noisy Unstructured Text Data; 2008 Jul 24. pp. 9–16.
8. Schweighofer E, Merkl D. A learning technique for legal document analysis. In Proceedings of the 7th Int Conf Artif Intell Law. 1999 Jun14:156–163.
9. Croft WB, Harding SM, Taghva K, Borsack J. An evaluation of information retrieval accuracy with simulated OCR output. In Symposium on Document Analysis and Information Retrieval; 1994 Apr. p. 115–126.
10. Lin X, Wang W, Li Y, Yang S, Feng F, Wei Y, et al. Data-efficient Fine-tuning for LLM-based Recommendation. In Proceedings of the 47th Int ACM SIGIR Conf Res Develo Information Retr; 2024 Jul 10. pp. 365–374.

