



International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

LEGAL INTELLIGENT SUMMARIZATION AND ANNOTATION (LISA)

Dr. Vijayant Verma¹, Saswat Rath², Nomita³, Abhinav Sahu⁴

¹Professor, Bhilai Institute of Technology, Raipur, Chhattisgarh, India

^{2,3,4}Students, Bhilai Institute of Technology, Raipur, Chhattisgarh, India

ABSTRACT :

An AI-powered platform called LISA (Legal Intelligent Summarization and Annotation) was created to automate important processes like text extraction, entity recognition, and document summarization in order to expedite the examination of legal documents. LISA uses Optical Character Recognition (OCR) to scan legal documents or process PDF files and extract important text using sophisticated Natural Language Processing (NLP) algorithms. Key legal entities, including individuals, organizations, and laws, are then identified and highlighted by the system using Named Entity Recognition (NER). Additionally, LISA reduces the amount of manual labor needed to evaluate large and complex legal texts by producing succinct summaries of legal papers. By offering rapid insights into case laws, contracts, and legislative documents, the platform seeks to increase the productivity of legal practitioners while saving them a significant amount of time and money. In addition to providing succinct summaries and enabling accurate annotation of crucial components such as case facts, judicial reasoning, legal principles, and outcomes, this system is intended to automate the extraction of important insights from legal documents. To guarantee context-aware comprehension of legal materials, the model integrates knowledge graph, text categorization, and semantic analysis approaches. LISA is an effective tool for attorneys, paralegals, and legal scholars because it can comprehend legal language and hierarchies by including domain-specific linguistic models.

Keywords- LISA, pytesract, pdfplumber, BERT, BART

Introduction of Legal Intelligent Summarization and Annotation

Legal documents play a pivotal role in defining, protecting, and enforcing the rights and obligations of individuals and entities. Contracts, for instance, are formal agreements between two or more parties, detailing their mutual rights and responsibilities in various contexts, such as business dealings, employment relationships, and service provisions. Wills and testaments, on the other hand, outline a person's wishes regarding the distribution of their estate after their death, ensuring that their assets are allocated as intended. Deeds are specialized legal instruments used to transfer property ownership, commonly in real estate transactions. Court pleadings are critical documents filed during legal proceedings, providing a formal framework for defining, protecting, and enforcing rights and obligations. Contracts are among the most widely used legal documents, serving as agreements between two or more parties that clearly outline the rights, responsibilities, and expectations of all involved. These can range from business contracts that govern partnerships or transactions to employment contracts that establish the duties of employers and employees, and service agreements that ensure clarity in professional engagements. Wills and testaments are another vital type of legal document, designed to communicate an individual's wishes regarding the distribution of their estate after death. These documents help ensure that personal assets are allocated in accordance with the deceased's preferences, reducing the likelihood of disputes among heirs. Deeds, meanwhile, are used specifically to transfer ownership of property, such as in real estate transactions. They legally formalize the process of transferring rights from one party to another, ensuring the rightful owner is documented. Court pleadings are essential for initiating or responding to lawsuits. These include complaints that outline the grievances of the plaintiff, motions requesting court orders, answers providing the defendant's responses, and subpoenas compelling parties to present evidence or testify. They serve as the foundational documents for legal proceedings, guiding the course of litigation. The primary purposes of legal documents are multifaceted and critical. They provide official evidence that can be relied upon in court to prove agreements, transactions, or facts. They also formalize agreements to make them legally binding, ensuring that parties are accountable for their commitments. By explicitly outlining terms, conditions, and responsibilities, legal documents ensure clarity and help prevent misunderstandings. Importantly, they also serve as tools for dispute resolution, offering a documented basis for resolving disagreements fairly and systematically. Overall, legal documents uphold fairness, transparency, and order, making them indispensable tools in personal, business, and societal interactions.

Purpose of this project

LISA (Legal Intelligent Summarization and Annotation) is an advanced AI-powered platform designed to enhance the efficiency of legal professionals by automating and streamlining the analysis of legal documents. Here's a detailed explanation of the key features and purposes of LISA:

1. Text Extraction and OCR (Optical Character Recognition) • Purpose: Legal documents are often stored in various formats, including scanned images, PDFs, or handwritten notes. These documents might contain valuable information, but manually extracting text from them is time-consuming and prone to error. LISA addresses this by leveraging OCR technology, which scans and converts scanned images or PDFs into machine-readable text.
2. Entity Recognition (Named Entity Recognition - NER) • Purpose: Legal documents often contain complex language, including references to key entities such as people, organizations, laws, dates, and legal terms. Identifying these entities manually is labor-intensive. LISA uses Named Entity Recognition (NER) to automatically extract and highlight these critical pieces of information.
3. Document Summarization • Purpose: Legal documents are often lengthy and detailed, making it difficult for legal professionals to quickly grasp the essential points. Summarization helps by generating concise versions of the document, providing an overview of the most important information.
4. Improved Legal Research and Case Review • Purpose: Legal professionals spend significant time reviewing case laws, contracts, and legislative documents. LISA enhances the efficiency of legal research by providing quick, AI-generated insights, reducing the need for exhaustive manual review.

3. Method for Legal Intelligent Summarization and Annotation :

It is a method of legal summarization and annotation which is used to summarize the pdf, images etc. Following libraries are important to the requirement of legal intelligence summarization and annotation they are as follows :-

3.1 Libraries :-

- **CV2:** The Open Source Computer Vision Library, or OpenCV for short, is a robust and open-source software library made to meet the complex requirements of machine learning and computer vision applications. It provides tools for real-time image and video processing, computer vision, and machine learning tasks.
- **Pytesseract :** It is a Python library enabling developers to extract text from images and scanned documents programmatically. Pytesseract is commonly used in applications like digitizing printed or handwritten documents, automating data entry, and building text analysis systems from scanned records.
- **PDFPlumber:** PDFPlumber is a Python library used for extracting structured data from PDFs, especially those containing tables, images, and text. It is often used for extracting tabular data and analyzing PDF documents in a more programmatically accessible form.
- **BART (Bidirectional and Auto-Regressive Transformers):** BART is a sequence-to-sequence model developed by Facebook AI, designed for both understanding and generating text. It combines the benefits of both BERT (bidirectional context) and GPT (autoregressive generation), making it highly versatile for various NLP tasks.
- **BERT (Bidirectional Encoder Representations from Transformers):** BERT is a pre-trained language model that helps with understanding the context of words in a sentence. It involves a structured process to leverage its capabilities for natural language processing tasks.
- **Tesseract (OCR):** Tesseract is used for Optical Character Recognition (OCR) to convert images (such as scanned documents or pictures) into machine-readable text. After preprocessing, the image input step involves providing the image, in formats such as PNG, TIFF, or JPG, to Tesseract.

The following steps include the - Legal Intelligent Summarization and Annotation they are as follows :-

3.2 Steps of Methodology:-

The **methodology of Legal Intelligent Summarization and Annotation** involves a structured approach that combines advanced technologies, natural language processing (NLP), and machine learning to process, analyze, and extract meaningful insights from legal documents. They are as follows .

The following steps are required to **legal intelligence summarization and annotation** are as:

1. **First step Collect raw data** - Legal documents are provided as inputs in formats like scanned images, PDFs, Word files, or plain text. Scanned images are digital representations of physical documents captured via a scanner or camera.
2. **Second step Preprocessing image** - After data collection, it is time to process the data,
 - **OCR (Optical Character Recognition):** Converts scanned or image-based documents into machine-readable text.
 - **PDF Parsing:** Extracts text, metadata, and structure from digital PDFs (e.g., contracts, judgments, or legal notices).
 - **Cleaning and Normalization:** Removes noise (e.g., special characters, page numbers) and ensures text consistency.

3. Third step User input : The user provides input to the system. This input can take two forms:

- **Scan a Legal Document:** If the user has a physical legal document, it is scanned and sent to the system for processing.
- **Upload a PDF File of Legal Document:** If the legal document is already in digital form, it is uploaded as a PDF file.

4. Fourth step Scan a Legal Document:

If the document is scanned, it goes through OCR (Optical Character Recognition).

- **OCR:** The OCR technology converts the scanned image into machine-readable text, enabling further processing.

5. Fifth step Upload a PDF File of Legal Document:

If the user uploads a PDF file, it is processed by a **PDF Parser**

- **PDF Parser :** it is a tool or library designed to read PDF files and extract data, including text, metadata, and the structural layout. It can handle both **text-based PDFs** (machine-readable text) and **image-based PDFs** (scanned documents with text stored as images).

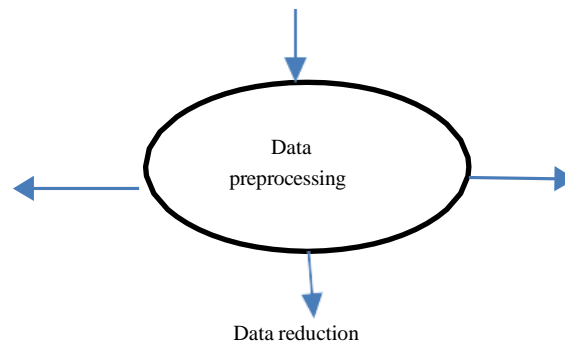
6. Sixth step Extracted Text:

Once the document is processed (either through OCR or PDF parsing), the text is extracted. This is the raw textual data derived from the input document.

7. Seventh step Named Entity Recognition (NER):

The extracted text is passed to a Named Entity Recognition (NER) system. This step helps in structuring and highlighting critical information within the text.

- **NER:** Identifies and classifies specific entities within the text, such as:
 - Names of individuals, organizations, or places.
 - Dates, legal clauses, and financial figures.
 - Key legal terms or provisions.



8. Eighth step Text Summarization:

After the entities are recognized, the system performs **text summarization**.

- **Purpose:** To create a concise version of the document, retaining only the most relevant and important points.
- This step helps users understand the document quickly without reading it in its entirety.

BART (Bidirectional and Auto-Regressive Transformer):

- A transformer-based language model developed by Facebook AI. Designed for sequence-to-sequence tasks such as text generation, summarization, and translation.
- Generating summaries for articles, paraphrasing text, or building dialogue systems.

9. Ninth step User Action :

The summarized text is presented to the user, and they are given options to proceed:

- **Copy the Content:** The user can copy the processed content for use elsewhere.
- **Download the Content:** The user can download the summarized and annotated content as a file.

10. Tenth step End:

The process concludes once the user has either copied or downloaded the content.

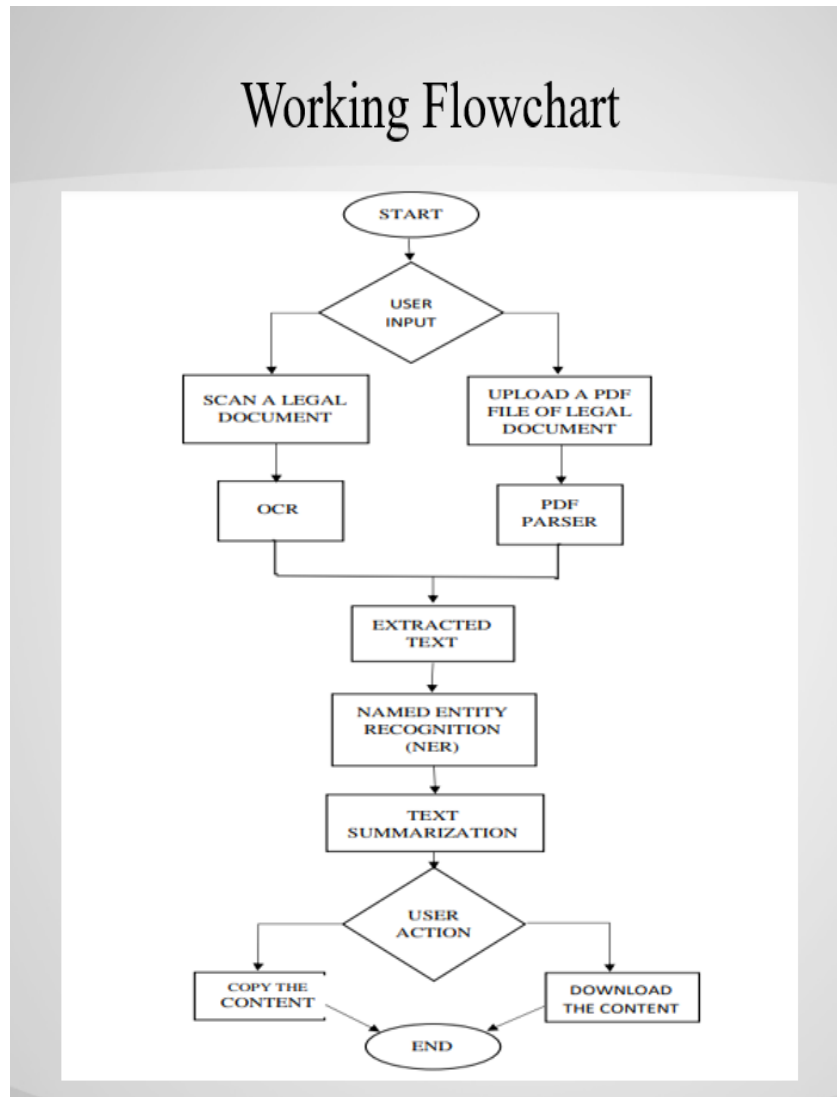


Fig-1 Flow chart of LISA

4. Benefit of Legal Intelligent Summarization and Annotation :

LISA's technology offers several benefits across different industries and applications. Following are the benefits, they are as follows:

Benefits:-

- Time and Cost Savings
- Enhanced Document Management:
- Improved Decision-Making:
- Increased Accessibility
- Support for Legal Research and Compliance
- AI-Powered Legal Insights
- Collaboration and Communication
- Scalability and Growth
- Competitive Advantage

Applications of Legal Intelligent Summarization and Annotation:

The applications of Legal Intelligence Summarization and Annotation span various areas of law and legal practice, offering significant benefits in efficiency, accuracy, and decision-making. Some key applications include:

Some key applications include:

- **Document Digitization** :Converts physical documents (e.g., invoices, receipts, contracts) into searchable and editable text formats.Helps in creating digital archives of printed materials.
- **Preventing the spread of COVID-19:** Facial recognition technologies had a broader application during the pandemic outbreak and are currently being used to track individuals who are COVID-positive and need to stay at home. A smartphone app with facial recognition capabilities requests a selfie from the individual being quarantined and verifies their identification to ensure they follow the guidelines for self-isolation.
- **assisting with mental health treatment:** Face recognition tracks the patterns and behaviors of patients' mental health. It interprets the wearer's facial emotions and presents appropriate indications, such as "anxious" or "happy."
- **Education:** In order to verify the identities of anyone entering or leaving the campus, a facial recognition system scans the faces of those people and compares them to a database of people who have been granted permission, which includes parents, existing staff members, and students.
- **Attendance monitoring:** In this sense, facial recognition software provides a quicker and less disruptive method of identifying individuals in the room. It not only saves valuable instructional time, but it also enables curriculum designers to plan classes perfectly and create more accommodating learning environments..
- **Banking and finance:** It makes sense that customer verification processes would follow, given that financial services have virtually completely moved online. Enhancing eKYC, an electronic variant of the "know your customer" protocol that manages the validation and authenticity of a customer's information, financial establishments can now completely transition the customer onboarding procedure on the internet.
- **Increasing learning engagemen:** Building upon eKYC, the electronic equivalent of the "know your customer" standard that controls the verification and authentication of a client's data, financial institutions can now move the full customer onboarding procedure online.\

Results of LISA :

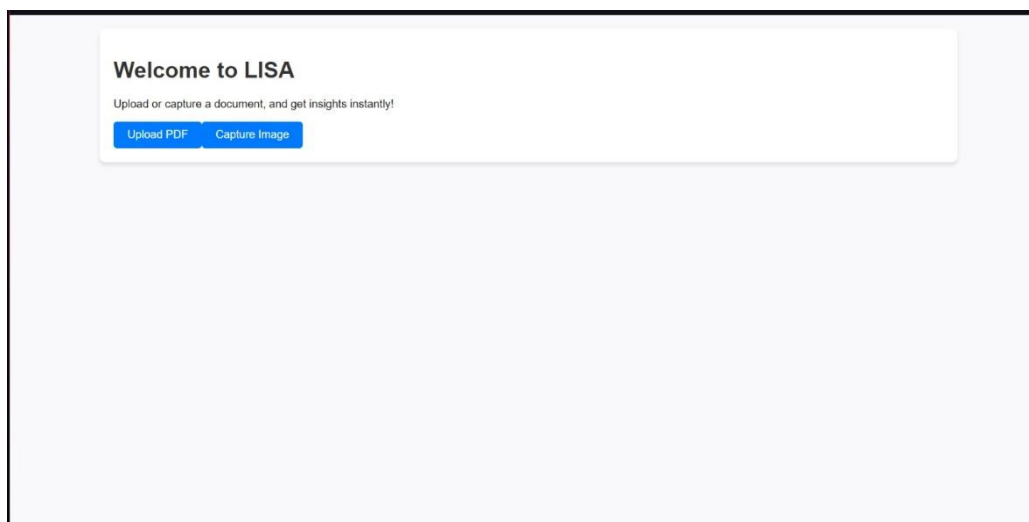
The result of legal intelligence summarization and annotation is a well-structured, concise, and enriched representation of the original legal document. Summarization condenses lengthy, complex legal texts into shorter, easily understandable formats, highlighting critical information such as key clauses, deadlines, obligations, and risks. Annotation adds further value by identifying and marking important entities (e.g., parties, dates, monetary values, legal terms) and linking them to relevant metadata or definitions, making the document more navigable and context-rich. These processes transform unstructured legal content into machine-readable and searchable formats like JSON, XML, or annotated PDFs, enabling advanced search, integration with legal tech tools, and streamlined workflows. Additionally, summarization and annotation help flag potential risks, ambiguities, or compliance issues, empowering legal teams to make informed decisions quickly. The final output may include summarized text, highlighted key points, and structured data, ensuring enhanced accessibility, accuracy, and efficiency in handling legal documents.

Key Results in Points

1. **Concise Summarization:** Legal documents are condensed into brief, easily understandable formats without losing critical information.
2. **Entity Highlighting:** Important entities like names, dates, monetary amounts, and legal terms are identified and tagged.
3. **Enhanced Navigability:** Key sections, clauses, and obligations are highlighted for easier document navigation.
4. **Risk Identification:** Potential risks, ambiguities, or compliance issues are flagged for quick review.
5. **Structured Data Output:** Transforms unstructured legal text into machine-readable formats (e.g., JSON, XML) for integration with legal tools.

6.1 LISA Front page

fig -2 Capture Image page



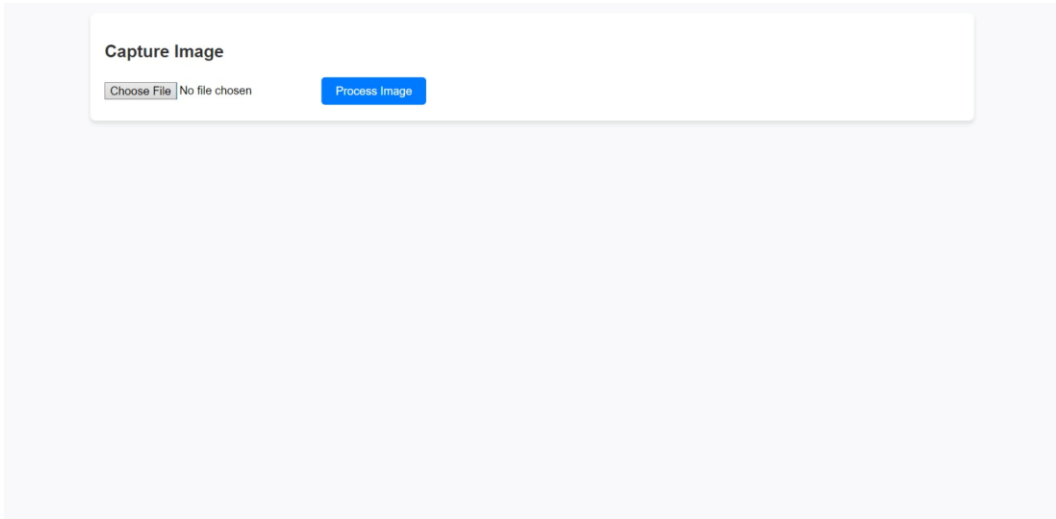


fig 3- An image ready for processing

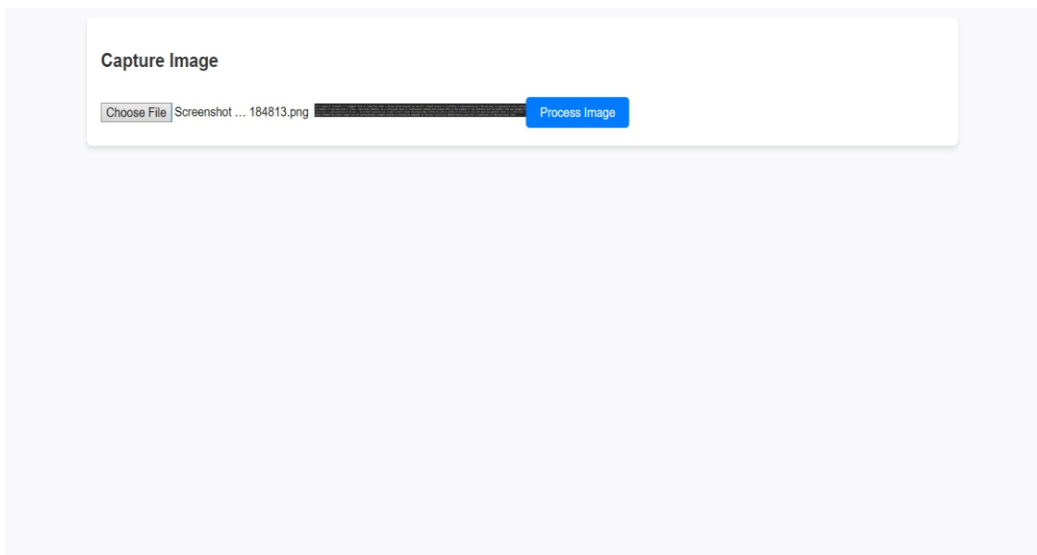


fig 4 - Summary of input text

Processed Results

Extracted Text:

Since Australian Softwood Forest Pty Ltd v Attorney-General (NSW) [1981] HCA 49 ; (1981) 148 CLR 121, esp at 125, the courts have recognised that the grant of declaratory relief on the application of a statutory body such as ASIC may serve important law enforcement purposes: see Corporate Affairs Commission (NSW) v Transphere Pty Ltd (1988) NSWLR 596 at 603; Australian Securities and Investments Commission v Sweeney [2001] NSWSC 114 at [30] - [31]; and Pegasus at 571. ASIC is charged with the administration and enforcement of the Act , and there will be many cases where it is in the public interest for the courts to make a declaration on ASIC's application that the Act has been contravened in specified respects. The making of such a declaration does not simply record the outcome of enforcement proceedings; it may also be an appropriate way of marking the court's disapproval of the contravening conduct: see Tobacco Institute of Australia Ltd v Australian Federation of Consumer Organisations Inc (No 2) (1993) 41 FCR 3@ at 97-99, 106 and 1ie.]

Summary:

Since Australian Softwood Forest Pty Ltd v Attorney - General (NSW) [1981] HCA 49 ; (1981) 148 CLR 121, esp at 125, the courts have recognised that the grant of declaratory relief on the application of a statutory body such as ASIC may serve important law enforcement purposes. The making of such a declaration does not simply record the outcome of enforcement proceedings; it may also be an appropriate way of marking the court's disapproval of the contravening conduct.

6.2 Following are the steps of result of Legal Intelligent Summarization and Annotation

- Upload document or capture an image
- Use ocr (optical character recognition) for image and pdf plumber for document
- Ocr extract the text of image and pdf plumber extract structure data from document
- After process document the data is extracted
- The extracted text is passed to a Named Entity Recognition (NER) system
- This step helps in structuring and highlighting critical information within the text.
- After the entities are recognized, the system performs text summarization

Future scope of LISA :

The field of legal intelligent summarization and annotation has significant potential to evolve and make legal processes more efficient. Here are the detailed aspects of its future scope:

- Enhanced AI and Machine Learning Models
- Multilingual Support
- Integration with Legal Tech Ecosystems
- Predictive Analysis and Risk Assessment
- Real-Time Legal Assistance

Conclusion :

The paragraph begins by describing **LISA**, which is likely an automated system or software tool designed to simplify and enhance the handling of legal documents. "Streamlining" refers to making a process more efficient, usually by eliminating unnecessary steps or automating tasks. In the context of **legal document processing**, LISA can automate many of the manual, repetitive tasks that are traditionally involved in dealing with legal documents.

For example, legal professionals often spend a significant amount of time reviewing contracts, legal briefs, case law, or regulatory documents. These tasks often require extracting key pieces of information, ensuring the document meets certain standards, and ensuring compliance with laws or regulations. By automating these processes, LISA helps **reduce manual effort**, which means lawyers and legal teams can spend less time on mundane tasks and focus more on complex legal analysis or decision-making.

Moreover, because LISA automates the extraction of information, it can **increase accuracy**. Manual data entry or document review is prone to human error, but LISA, powered by sophisticated algorithms, can process documents more consistently and reliably, leading to more accurate results.

7. REFERENCES :

- [1] Harshil Darji, Jelena Mitrovic, Michael Granitzer, " German BERT Model for Legal Named Entity Recognition", 2023, from arXiv:2303.05388 [cs.CL]
- [2] Santosh T.Y.S.S, Mahmoud Aly, Matthias Grabmair, " LexAbSumm: Aspect-based Summarization of Legal Decisions", 2024 from arXiv:2404.00594 [cs.CL]
- [3] van Heusden, R., de Ruijter, A., Majoor, R., Marx, M. (2023). Detection of Redacted Text in Legal Documents. In: Alonso, O., Cousijn, H., Silvello, G., Marrero, M., Teixeira Lopes, C., Marchesin, S. (eds) Linking Theory and Practice of Digital Libraries. TPD L 2023. Lecture Notes in Computer Science, vol 14241. Springer, Cham. https://doi.org/10.1007/978-3-031-43849-3_28
- [4] Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, Vivek Raghavan, " Named Entity Recognition in Indian court judgments", 2022 from arXiv:2211.03442 [cs.CL]
- [5] Abhay Shukla, Paheli Bhattacharya, Soham Poddar, Rajdeep Mukherjee, Kripabandhu Ghosh, Pawan Goyal, Saptarshi Ghosh, Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation from arXiv:2210.07544 [cs.CL]
- [6] Deepa Anand, Rupali Wagh, Effective deep learning approaches for summarization of legal texts, Journal of King Saud University - Computer and Information Sciences, Volume 34, Issue 5, 2022, Pages 2141-2150, ISSN 1319-1578, <https://doi.org/10.1016/j.jksuci.2019.11.015>. (<https://www.sciencedirect.com/science/article/pii/S1319157819301259>)
- [7] R. Giri, Y. Porwal, V. Shukla, P. Chadha and R. Kaushal, "Approaches for information retrieval in legal documents," 2017 Tenth International Conference on Contemporary Computing (IC3), Noida, India, 2017, pp. 1-6, doi: 10.1109/IC3.2017.8284324. keywords: {Semantics;Law;Task analysis;Tools;Information retrieval;Data mining},
- [8] arXiv:2210.07544 [cs.CL]
- [9] arXiv:2112.14731 [cs.CL]