AI-POWERED CONTRACT ANALYSIS SYSTEM USING OCR, NLP, AND LARGE LANGUAGE MODELS

1st Ouku Bhulakshmi
Assistant Professor,
Department of CSE,
SanthiramEngineeringCollege,
Nandyal, Andhra Pradesh
oukubhulakshmi@gmail.com

2nd B.Harika
Department of Computer Science and Engineering,
SanthiramEngineering College,
Nandyal, Andhra Pradesh
23x51a0532@srecnandyal.edu.in

3rd Kumar Devapogu
Computer Science and Engineering,
Assistant Professor,
Vignan's Foundation For Science,Technology&Research-Guntur
Kumar.mtech1@gmail.com

4th G.Swapna
Department of Computer Science and Engineering,
Santhiram Engineering College,
Nandyal, Andhra Pradesh
23x51a0576@srecnandyal.edu.in

5th E.Anuja
Department of Computer Science and Engineering,
Santhiram Engineering College,
Nandyal, Andhra Pradesh
23x51a0556@srecnandyal.edu.in

6th C.Lakshmi Keerthana
Department of Computer Science and Engineering,
Santhiram Engineering College,
Nandyal, Andhra Pradesh
23x51a0537@srecnandyal.edu.in

ABSTRACT: ***The process of manually analyzing contracts often takes long amounts of time, requires legal training and knowledge, and is not typically affordable to individuals and small businesses. In addition, current forms of analysis offer only moderately compliant output due to the fact that they typically only provide basic functionality of reviewing documents and the ability to search for key words, and therefore do not provide the ability to intelligently assess and identify potential risk, nor do they provide multi-lingual support.***

In an effort to improve the process of contract analysis, this paper introduces an AI-based Contract Analysis System that has been developed to facilitate and automate the process of interpreting legal documents. Included in this system is Optical Character Recognition (OCR) for the extraction of text from PDF and image formats, Natural Language Processing (NLP) for the identification of clauses and semantic analysis of those clauses, and Large Language Models (LLM) for creating short summaries and contextual explanations of clauses.

This planned framework is intended to provide a clause-by clause analysis of important contract clauses as well as to identify any possible risks, which include the penalty of termination, termination conditions, and compliance obligations. A compliance scoring mechanism is also created to allow for the evaluation of the user's level of compliance with the terms of these contracts. Additionally, the system offers services in various Indian languages to allow users to interpret their legal documents in their chosen language.

Tests and evaluations conducted show that the system greatly enhances the effectiveness, accuracy, and usability of legal document analysis when compared to the prior art of manually conducting the same types of analyses. The combination of OCR, NLP, and generative AI technologies enable the solution to support an intelligent and scalable method for performing legal document analyses, and thereby provide users with a method for making decisions with much less complexity and risk.

KEYWORDS: Contract Analysis, Legal Document Processing, Optical Character Recognition (OCR), Natural Language Processing (NLP), Large Language Models (LLM), Risk Detection, Clause Extraction, Generative AI, Multilingual Systems, Legal AI Assistants

I.INTRODUCTION

As businesses become more reliant on both legal contracts in their commercial dealings and also personal arrangements, understanding contracts has become critically important. Because of the complex language, long clauses, and ambiguous terms contained in most legal documents, many individuals and small companies run the risk of misinterpreting the content unless they have the assistance of a qualified attorney. As a result, the risk of misinterpreting legal documents increases as does the exposure of individuals / small companies to financial and/or legal liability.

The digitisation of legal documents has spurred the urgent requirement for automated and intelligent solutions to quickly analyse and interpret contracts. Contract analysis has traditionally been a manual, time-consuming process requiring expert knowledge of the field which makes it difficult for non-legal personnel to access [7]. Additionally, the existing tools that support the digital nature of documents have primarily been focused on storing documents or providing keyword search capabilities, there are limited or no advanced features available for semantic understanding, clause-level analysis, or identifying risks within a document [8].

The development of Artificial Intelligence (AI) technologies like Natural Language Processing (NLP) have made it easier to work with unstructured text data [1], [9]. Named entity recognition, classification of documents, and semantic analysis have all been utilized within Legal Informatics to create actionable results from complicated documents [10]. OCR has been another advancement that allows text to be extracted from scanned documents and images so it can be processed end-to-end in a document processing pipeline [11], [23].

Recent advancements in machine learning for natural language processing (NLP) have greatly changed our ability to comprehend and create written text by providing a level of contextual information that is comparable to

those produced by human beings. These models have the ability to generate written content with near-human quality, both in explaining concepts and summarizing whole pieces of content (e.g., text) [12]. Furthermore, applying these types of model when interpreting clauses enables one to identify risks, obligations, and compliance requirements with greater accuracy than has been achieved previously within legal frameworks.

Although there has been progress in developing tools for contract analysis, most existing systems have not fully incorporated optical character recognition (OCR), natural language processing (NLP), and generative AI into one integrated solution. In addition, their limited ability to interpret multiple languages and assess risk in real-time has limited their overall effectiveness in fast-moving and constantly changing environments [13].

The purpose of this paper is to develop an AI powered contract analysis tool which facilitates and automates the review of contract clauses to help users understand the obligations set forth within contracts between parties. Using OCR, NLP and large LLMs; the system will identify risks, provide clause summaries and compute a compliance score for each clause in order for end users to have an easier time understanding their contractual obligations with other parties. Additionally, this system will provide users whom speak multiple languages with access to the same level of information in their native tongue.

The paper will provide the following contributions

- The development of a unified AI framework that integrates OCR, NLP, and LLM technologies to provide a comprehensive approach to analyzing contracts;
- The implementation of the detection of risk associated with contract clauses and compliance scoring;
- Providing a means to interpret languages to increase access
- An empirical evaluation of the AI contract analysis tool; comparing its usability and effectiveness versus traditional methods of contract analysis.

The intention of designing this tool is to provide an intelligent, scalable and accessible method for users to easily understand the complexities of contracting process.

II. LITERATURE REVIEW

In recent years, the rising amount of digital contracts and the increasing requirement for automated interpreting systems have greatly increased the focus of legal document analysis. Historically, these systems of analyzing legal documents have been based on two primary methods: rule-based systems that depend on predefined sets of rules that can be applied to analyze legal documents; and keyword matching systems, which rely solely on searching for specific keywords to identify and analyze legal documents. However, keyword-based systems generally lack the capability to present the contextual meaning or to interpret the complicated relationships found within legal documents [14]. Because of this limitation, the most accurate interpretations of the contents of legal documents have often been missed due to the dependency of the keyword-based systems on the predefined rules and the absence of semantic understanding.

With ongoing developments in the advancement of Natural Language Processing (NLP), machine learning based techniques are now being used to conduct legal text analysis [2],[15]. Methods of legal text analysis that use machine learning technologies such as text classification, named entity recognition (NER), and syntactic parsing have become widely adopted to accurately extract key pieces of information from legal documents [15]. Nevertheless, these systems continue to experience difficulties in conducting legal text analysis because of the length of legal sentences, the ambiguity in the structure of clauses, and the specialized terminologies used within legal disciplines.

Optical Character Recognition (OCR) has also been used extensively in the digitization of scanned documents of the legal document type and to extract text from images and PDF files containing legal text. Modern OCR technologies (including deep learning-based OCR technologies) have improved the accuracy of extracting text; however significant challenges still remain in extracting text from images due to the noisy inputs generated by poorly scanned documents, the formation of handwritten text within a legal document, and the complexity of the layout of legal documents [16].

Utilizing natural language processing (NLP) and deep learning models like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Transformer based structures are becoming popular in recent times to enhance how well we understand legal documents. All of these models have been able to achieve success at determining how the clauses relate to one another in legal documents and help with determining what specific information should be extracted from those clauses [17].

Legal artificial intelligence is being advanced by large language models (LLMs). LLMs can provide contextual reasoning, summarize large amounts of text, and give human-level explanations of those texts, thus making them an extremely attractive option for contract analysis [18]. Problems such as high computational cost, difficulty in generating output that is easy to read (interpretable), and inability to easily perform domain-specific fine tuning present challenges to their development and will be the subject of continued research.

Another area where multiple studies have been completed includes identifying penalties, obligations and termination clauses in contract documents for purposes of identifying

risk. These types of systems usually rely on predefined patterns of clauses or supervised learning models and will not be able to detect anything that has not already been seen or is more complex [19].

Yet another emerging opportunity for legal document analysis is to have the ability to analyze legal documents in multiple languages, which is particularly relevant in countries with multiple language groups. Almost all systems that are currently being developed for legal documents do not have solid capabilities in this area, and many potential users of these systems, who cannot read English, will be unable to access these systems [20].

The existing contract analysis systems suffer from some shortcomings despite the capabilities of today's technology. Many of these methods are still focused on only one of the five areas discussed above (OCR or NLP) and do not have a combined approach to resolve the contract item. Also, most current systems are unable to conduct a clause-level analysis, provide real-time risk indicator detection, or have an easy-to-use interface for non-expert individuals [21].

While, The proposed system brings together the OCR technology, NLP Technology and LLM Technology into one complete system so that the user will have a fully functional end-to-end contract analysis solution. The system also improves upon the accuracy, usability and accessibility of existing solutions by providing clause level information, real-time risk detection, compliance scores and multilingual capabilities thereby overcoming many of the issues created by the previous methods discussed earlier. [22]

III. METHODOLOGY

The proposed Contract Analysis System consists of several layers of artificial intelligence technologies, which will combine Optical Character Recognition (OCR), Natural Language Processing (NLP), and Large Language Models (LLM) to create an easy-to-use automated system that interprets legal documents.

The System Methodology focuses on: efficiently extracting relevant text, semantic analysis of the text for underlying meaning, identifying risks from the text, and producing a user-friendly Output (to end-users). A. Data Input And Pre-processing

The Contract Analysis System will accept different documents; however, only PDFs, JPEGs and scanned documents will be inputted in As-is condition due to requiring OCR (Optical Character Recognition). The OCR technology will read the text and extract the text from the scanned documents/images. After the text has been extracted, it will go through pre-processing stage to remove noise (incorrect formatting)[3], [6] correct mistakes made[5] (to create readable output) as well as classify extracted material into useful groupings (e.g sentence and clause) [24].

B. Clause Extraction/NLP Processing Stage

Extracted text will have undergone Natural Language Processing (NLP) analysis. Operations that will be performed include: the identification of clause (using sentence boundary detection and syntactic parsing); performing key NLP tasks on extracted clauses (i.e. tokenize, part of speech tags, named entity recognition); this will allow us to identify significant entities, obligations and conditions contained in the contract [25].

In addition to the clause extraction and analysis steps, each respective clause will also be categorized into categories like payment terms, termination clauses, penalties and compliance. This will help provide the user with clarity regarding complex legal documents.

C. Mechanism for Detecting Risk

The company also uses a set of rules to identify possible areas of risk in contracts by applying Natural Language Processing (NLP). The risks identified will include those indicated by patterns or through use of the same word, as well as by reference to previously defined contract elements, such as those having high monetary penalties or being subject to restrictive termination clauses or having clauses that are not clear or unambiguous. This will enable the user to identify key areas of attention [25].

D. Summarization and Explanation using a LLM

Language models serve as the foundation of generating short summaries and contextual explanations for each clause contained within the contract by automatically processing the data that has been extracted from the contract and condensing it into a simpler version for nonlegal users to interpret complex contractual content. As such, this will provide users with an improved ability to interpret and use the contract.

E. Compliance Score

A compliance score will be used to provide an assessment of a contract's degree of conformity to standard legal parameters (i.e., satisfactory). The calculated score will be based on a number (e.g., certain essential clauses contained in a contract or by examining various types of risk associated with a contract or by determining the completeness of a contract). The compliance score will provide an approximate assessment of the reliability of a given contract to the user.

F. Multilingual support

The program is designed so that users can access a range of Indian languages. In order to increase accessibility, translation models will help convert extracted text and

produced summaries into the language of the user, thereby widening the potential for use among all users

G. System workflow

The flow of the system includes the following steps:

- upload document (PDF/Image),
- extract text from document via OCR,
- prepare/segment the text,
- extract clauses from prepared/segmented text and analyze with NLP,
- detect and classify risks from clauses,
- summarize with LLM, and
- score compliance and provide translated output to user.

By having the system structured this way, users can receive accurate, efficient and scalable analysis of contracts, therefore allowing them to make better decisions and have actionable information.

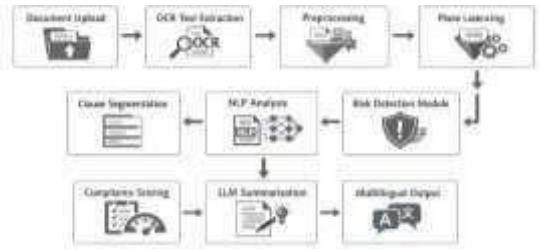

Fig. 1. Methodology Workflow

IV. SYSTEM ARCHITECTURE

To Design an AI-based Contract Analysis System, we have architected a modularized layered framework that will allow us to achieve scalability, flexibility and efficiency for processing legal documents. Each architectural layer will accommodate integrated components from Optical Character Recognition (OCR) and Natural Language Processing (NLP) to Large Language Models (LLM) [4],[8], thus giving us a complete end-to-end contract analysis solution. A. Overall Architecture Overview

The overall system has six distinct functional layers: User Interface, Input Data, OCR Processing, NLP/Course Analysis, LLM Processing and Output/Visualisation. Each Functional Layer serves a unique purpose to support the continuous processing of data throughout the pipeline.

B. User Interface Layer

The User Interface (UI) Layer provides the user an interactive interface where they can upload their legal documents (PDF or Image) for analysis. The UI also allows the user to choose what language they wish to have the documents analysed in. The UI provides the user with the analysis results (Summary, Risk Alerts, Compliance Scores).

C. Input Data Layer

The Input Data Layer receives the uploaded documents and supports the various file types that will be uploaded. This is done in order to ensure that the uploaded data is properly sent to the OCR Module for further processing.

D. Optical Character Recognition (OCR) Processing Layer

The document processing system's Optical Character Recognition (OCR) layer will extract the textual information from the scanned documents and PDFs that were uploaded. Using OCR technology to determine how best to extract the text by interpreting complex document structures means that the extracted text will be available for further analysis in the future. [6], [11]

E. Natural Language Processing (NLP) Clause Analysis Layer

Following the initial extraction of the text from the documents, the second step is the application of Natural Language Processing (NLP) techniques to the extracted text to segment the clauses and recognize entities within the clauses. It also detects the contract sections (i.e., payment, penalties, termination, etc.) and analyzes the structure and meaning of the contracts to identify key elements.

F. Large Language Model (LLM) Processing Layer

The LLM processing layer processes the extracted clauses and provides the clauses' summary and explanation in a simple manner. It uses reasoning, in context, to create a more detailed explanation (e.g., of electronic contracts) as well as to estimate the probability of appropriate reasons supporting the clause.

G. Output and Visualization Layer

The output and visualization layer presents the final results of the processing described above. This includes:

- Clause-level summaries
- Highlighted risk factors
- Compliance scores
- Translations into other languages

The users can access this information through a userfriendly dashboard. The information can be reviewed and acted on without delay.

H. Flow of Data

The system processes data in a linear fashion as follows:

1. A user uploads a legal contract document
2. The OCR functionality extracts textual data from the legal contract document
3. The NLP component segments and processes identified clauses from the legal contract document

4. Risk Identification identifies potentially risky attributes within the identified clauses
5. The LLM will summarize and explain the legal summary of the contract
6. Acompliance score for the contract will be generated
7. The user will then see the results

I. Scalability and Adaptability

The architecture of the system is designed to be cloudbased and scalable. The system's modularity allows for easy integration with other components, such as more complex NLP models or legal datasets for specific areas of law. The system can also adapt to changes in contract formats and changes in the law; therefore, the system can continuously incorporate new contract formats and new laws.

This multi-tiered architecture will allow for efficient processing of information, accurate analysis of the information processed, and enhanced usability of the system; thus, the system will be applicable in the real world within the legal profession.

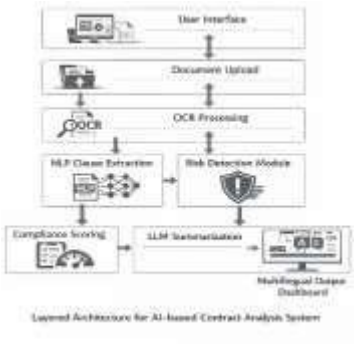

Fig. 2. System Architecture

V. RESULTS & DISCUSSIONS

Evaluating the performance of the new AI-based Contract Analysis System includes activities such as extracting text, understanding clauses, identifying risks, and summarising key findings. The incorporation of OCR, NLP & LLM technologies into one system allows for quick and accurate processing of legal documents.

In terms of extracting text from PDF and image documents, the Contract Analysis System is successful at this task regardless of whether the layout is moderately complex or not; the use of OCR provides accurate results.Identifying key clauses (for example: payment terms, conditions of termination, penalties) using NLP provides users with a better way to structure their understanding of contract content. Detecting high-risk clauses such as those related to high penalties, ambiguous obligations and strict termination clauses allows the Contract Analysis System to help users identify and mitigate the risks associated with their contract content

The LLM technology-based summary tool will assist you to easily comprehend complex legal jargon and give you a simplified summary of important clauses (i.e., a description of the clause written in understandable language).

The compliance scoring tool will provide you with a quantitative score of the reliability of your contracts and assist you in determining your contractual obligations. Additionally, by providing multilingual capability, you will have the option to view contracts in your chosen language.

In summary, this aforementioned methodology offers a scalable, intelligent means for the analysis of legal documents and allows users to make informed decisions with less complexity and more confidence.

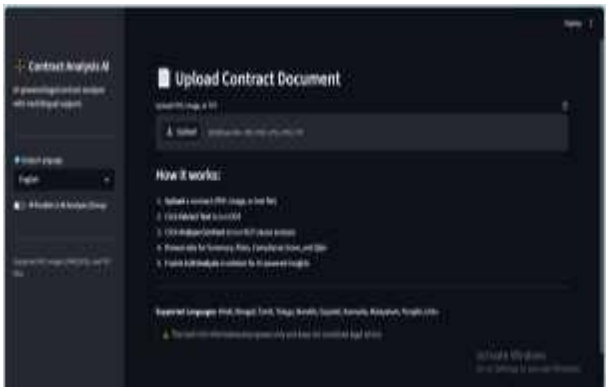

Fig. 3. Document Upload Interface

Fig. 4. Contract Summary Generated by LLM

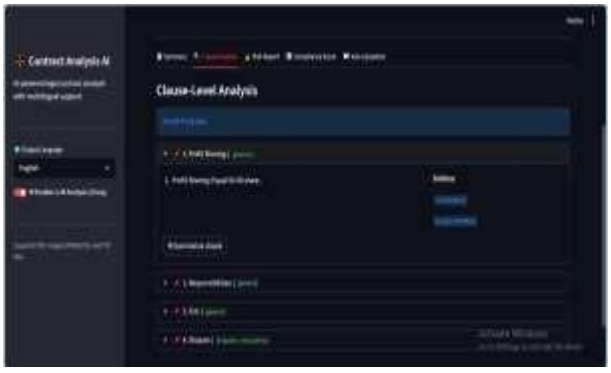

Fig. 5. Clause Classification and Analysis Dashboard

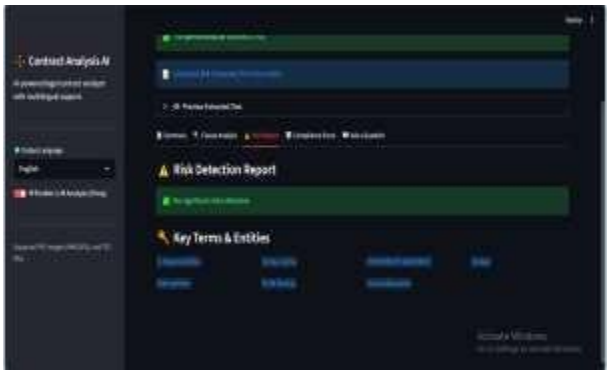

Fig. 6. Risk Detection Output

Fig. 7. Compliance Score and Multilingual Output

Fig. 8. AI Chat

VI. CONCLUSION & FUTURE WORK

An analysis of a Contract system that will use Artificial Intelligence technologies to simplify reading legal documents. An example of this System would use 3 different types of technology (Optical Character Recognition - OCR; Natural Language Processing - NLP; and Large Language Models - LLM). The result will be a completely automated approach to extracting text from contracts, analysing clauses of the contract, identifying high risk items and providing a summary of the contract. This means that when you analyse at the clause level and can identify high risk items, users will be able to understand what they are obliged to do under that contract, and also determine if they are going to have to pay for any loss or damage resulting from a breach of that contract.

Based on the results of a number of tests, the system is capable of processing larger volumes of legal documents at a much faster rate, with greater accuracy and reliability than traditional methods of processing these documents manually. The inclusion of a compliance score gives users a way to evaluate how reliable the contract is, and multilingual capability gives users access to the system across a wider range of people who may speak languages other than English, thus improving user experience. As a result of the addition of generative AI, the system also improves the interpretability of legal documents by converting complex wordings into simple terms.

There are also some limitations of this system. The OCR performance can be negatively impacted by the quality of the documents and how well the documents are formatted. The accuracy of the NLP models used in the system may also be less predictive when dealing with very specific legal terms that are unique to different industries (such as insurance). Finally, LLM-based documents may add additional processing time for applications that need realtime responses.

In the next stage of the project, we will focus on improving our ability to fine-tune model performance across the legal domain. We will develop more sophisticated design architectures for our transformer models to increase the accuracy of classifying clauses. Finally, we will build out the ability to process data in near real-time and increase the scale at which we can deploy these models.

Further, we will explore using XAI methodologies to allow for greater transparency within our data science analytics, and integrate our models with the various legal knowledge bases that exist today to complete the analysis of contracts.

The overall goal of the proposed system is to provide a scalable, intelligent, and easy-to-use solution for automated contract analysis, allowing users to make informed legal decisions with greater ease and confidence.

REFERENCES

[1] J. V. Suman et al., “Leveraging Natural Language Processing in Conversational AI Agents to Improve Healthcare Security,” in Conversational Artificial Intelligence, 2024, pp. 699–711.

[2] M. Sharmila Devi et al., “Extracting and Analyzing Features in Natural Language Processing for Deep Learning with English Language,” Journal of Research Publication and Reviews, vol. 4, no. 4, pp. 497–502, 2023.

[3] S. N. Chaudhri, A. Mishra, N. S. Rajput, Y. Mallikarjuna Rao, and M. V. Subramanyam, “Vision Transformer-Based LULC Classification Using Remotely Sensed Hyperspectral Image,” in Lecture Notes in Electrical Engineering, vol. 1157, 2024, pp. 127–136.

[4] M. Amareswara Kumar et al., “An Artificial Intelligent (AI) Automated Compliance Framework for Securing Critical Infrastructure,” in 2026 6th International Conference on Image Processing and Capsule Networks (ICIPCN), IEEE, 2026.

[5] K. Malllikarjuna, G. R. C. K. Sarma, M. V. Subramanyam, and K. S. Prasad, “EBP based GKLM method for neural network training,” in 2011 3rd International Conference on Computer Research and Development (ICCRD), vol. 2, 2011, pp. 504–507.

[6] M. Farooq Sunar, K. Balasubramanian, and T. Sudhakar Babu, “Comprehensive Research on Video

Imaging Techniques," All Open Access, Bronze, 2019.
[7] D. Katz, M. Bommarito, and J. Blackman, "A General Approach for Predicting the Behavior of the Supreme Court of the United States," PLOS ONE, 2017.
[8] K. Ashley, "Artificial Intelligence and Legal Analytics," Cambridge University Press, 2017.
[9] D. Jurafsky and J. Martin, "Speech and Language Processing,"3rded.,2023. https://web.stanford.edu/~jurafsky/slp3/
[10] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,"NAACL,2019. https://arxiv.org/abs/1810.04805
[11] R. Smith, "An Overview of the Tesseract OCR Engine," ICDAR, 2007. https://ieeexplore.ieee.org/document/4376991
[12] T. Brown et al., "Language Models are Few-Shot Learners," NeurIPS, 2020. https://arxiv.org/abs/2005.14165
[13] A. Chalkidis et al., "Legal-BERT: The Muppets Straight Out of Law School," Findings of EMNLP, 2020. https://arxiv.org/abs/2010.02559
[14] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing & Management, 1988. https://www.sciencedirect.com/science/article/pii/0306457388900210
[15] S. Bird, E. Klein, and E. Loper, "Natural Language Processing with Python," O'Reilly Media,2009. https://www.nltk.org/book/
[16] A. Graves et al., "Offline Handwriting Recognition with Multidimensional Recurrent Neural Networks," 2009. https://arxiv.org/abs/0907.0146
[17] S. Hochreiter and J. Schmidhuber, "Long ShortTerm Memory," Neural Computation, 1997. https://www.bioinf.jku.at/publications/older/2604.pdf
[18] A. Vaswani et al., "Attention is All You Need," NeurIPS,2017. https://arxiv.org/abs/1706.03762
[19] J. Zhong et al., "Extracting Contractual Obligations and Risks Using NLP," 2020. https://arxiv.org/abs/2008.01520
[20] G. Conneau et al., "Unsupervised Cross-lingual Representation Learning," ACL, 2018. https://arxiv.org/abs/1710.04087
[21] R. Moens et al., "Automatic Detection of Legal Concepts in Documents," Artificial Intelligence and Law,2007. https://link.springer.com/article/10.1007/s10506-007-9035-3
[22] A. Aletras et al., "Predicting Judicial Decisions of the European Court of Human Rights," PeerJ Computer Science,2016. https://peerj.com/articles/cs-93/
[23] P. Singh et al., "Text Extraction from Documents Using OCR Techniques," 2019. https://arxiv.org/abs/1909.09180
[24] C. Manning et al., "Stanford CoreNLP: A Toolkit for NLP," ACL, 2014. https://aclanthology.org/P14-5010/
[25] T. Mikolov et al., "Efficient Estimation of Word Representations in Vector Space," 2013. https://arxiv.org/abs/1301.3781