International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 11 | Nov - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Automated Legal Document Analysis Using Natural Language Processing

Rasik Gupta

Department of Computer Science

Abstract
Natural Language Processing (NLP) is now a valuable tool in legal document analysis and a cutting-edge tool in automating contract analysis, research, and compliance tracking. This paper aims to discuss the implementation of NLP in a field to analyze legal documents with emphasis placed on its effectiveness in increasing accuracy, efficiency, and access to the law profession. In this vein, selected NLP approaches, including Named Entity Recognition (NER), sentiment analysis, and text summarization, are made to address legal automation. Previous methods are considered and evaluated for applicability to large-scale legal texts to reveal the advantages and drawbacks of the available tools. Also, the paper outlines various developments of NLP technology that can be taken in a bid to enhance the legal automation process.

Keywords
NLP, Legal Analytics, Legal AI, NER, Legal Text Analytics, Summarization, Legal AI, Documenting Automation

1. Introduction

The field that has been primarily relying on the traditional methods of searching through contracts as well as performing law research among other legal processes that involve compliance constitutes the legal industry which is well known to use lengthy securities. These procedures are tedious, complicated, and resource-intensive and they are also liable to human interference. Technological advancement by the *Natural Language Processing (NLP)* system has opened a new frontier for the legal field. Natural language processing also known as NLP is a field of artificial intelligence that allows machines to process, analyze, understand, and even interact using written or spoken human language. It presents an opportunity to fully revolutionize the legal sector, especially in the manner and approach used in the analysis of the various legal documents by fully eliminating tasks that would otherwise consume a lot of human resources.

2. Literature Review

As with any field, the utilization of NLP in a legal context emerges as a particular topic for discussion and scientific and practical exploration. Unlike other documents, legal documents contain a lot of professional language, and use formal language, and such documents are bound by format, which makes manual processes time-wasting and costly (Chowdhary & Chowdhary, 2020, p. 4). This section sums up the previous contributions in the field of NLP to study their impact on the automation of the process of analysis of legal texts, considering both historical developments and the most recent ones.

2.1 Historical Growth of NLP in the Analysis of Legal Documents

The first attempts to apply NLP for performing legal analysis followed the creation of rule-based systems. There are certain kinds of rule-based systems, including those used in early Legal Information Retrieval, in which novice approaches to pattern matching and keyword spotting were used (Bommarito et al. 2021). There was LEGALEX which was meant for making legal reasoning via a keyword-based extraction system. These early systems themselves showed some successes in information retrieval although substantial enough to make a real impact in the years to follow, the shortcomings of the tools were their inability to handle law language and fit the workflow to the structure of a document (Zhong et al. 2020).

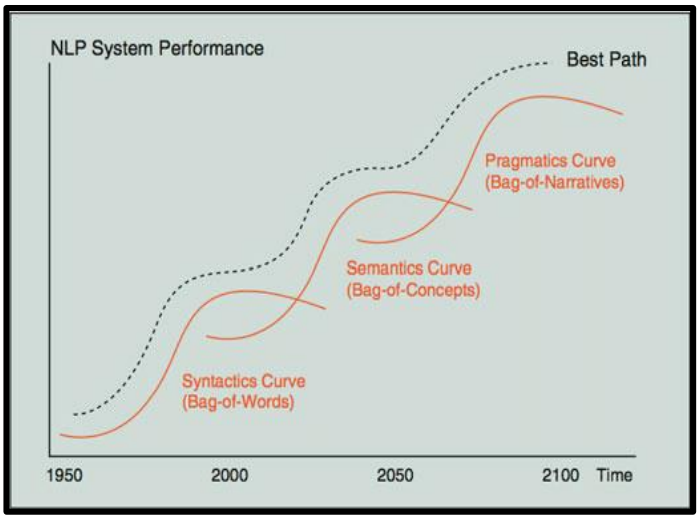

Figure: 2.1.1 Historical Growth of NLP
(Source: Cavalieri et al. 2020, p. 1482)

2.2 Current Approaches: Machine Learning and Deep Learning Models

The advancement of machine learning and, in the last few years, deep learning models have dramatically changed the application of NLP in the legal area. Different algorithms like SVM, Random Forest, and even more strategic algorithms like BERT also brought a remarkable change in performance for activities like NER and classifying legal documents (Baviskar et al. 2021).

The advancement to higher levels of more complex and contextualized legal computations has been one of the biggest advancements in using legal NLP (Chen et al. 2022, p. 5). In LexNLP traditional NLP techniques are merged with legal qualifiers and algorithms, which makes it capable of', identifying legal clauses, 'entities like laws and regulations,' and 'summarization of documents'. Another tool called DocuSign utilizes NLP to parse the important provisions of the contract, which would ease the process of contract documentation review and risk assessment.

In a similar capacity, it has emerged that BERT and GPT (Generative Pretrained Transformer) can be effective in handling unstructured legal texts (Ahmad et al. 2021). Consequently, BERT-Legal which is developed from the BERT has a greater accuracy for legal language as recognized from the results in the area such as legal text summarization and case law analysis.

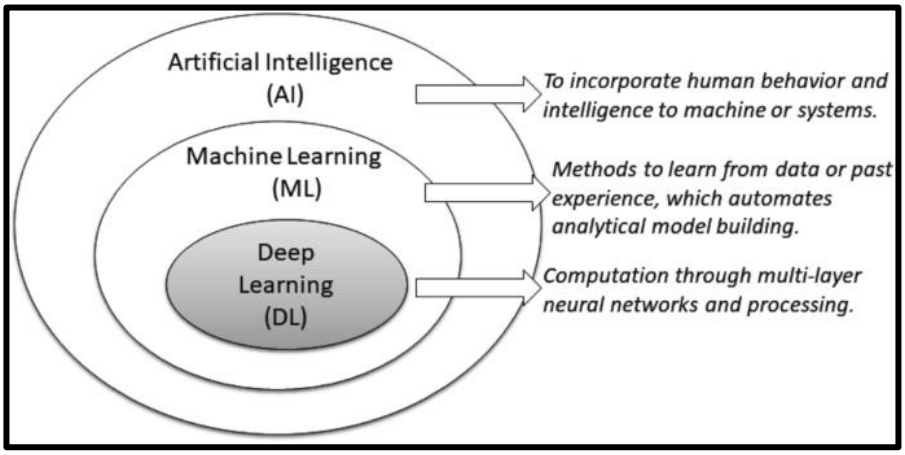

Figure: 2.2.1 Machine Learning and Deep Learning Models according to NLP
(Source: Sarker, 2021, p. 3)

2.3 Challenges in Legal NLP

However, that's where the problem stands even with the recent advances in the use of NLP on legal documents. Legal language, as noted, has lots of synonyms and antonyms and depends on contextualisms and paradigmatic connections between the words. Legal documents are often not without vague terms in law, there are differences in legal terms of different jurisdictions, and requirements for a strict interpretation of the legal clause contributed to the difficulty of NLP in this category (Ganguly et al. 2023). In addition, questions of data privacy and confidentiality constrain the access to different training data sets and the opportunity to build models able to generalize across different domains and jurisdictions.

The other important issue is the interpretability of the model. Since legal advocates are likely to rely on the output of NLP models which work as 'black boxes', the latter has to be trusted by the former. As the results show, it is vital to design NLP systems that would provide clear and understandable outcomes to improve their application in the legal context.

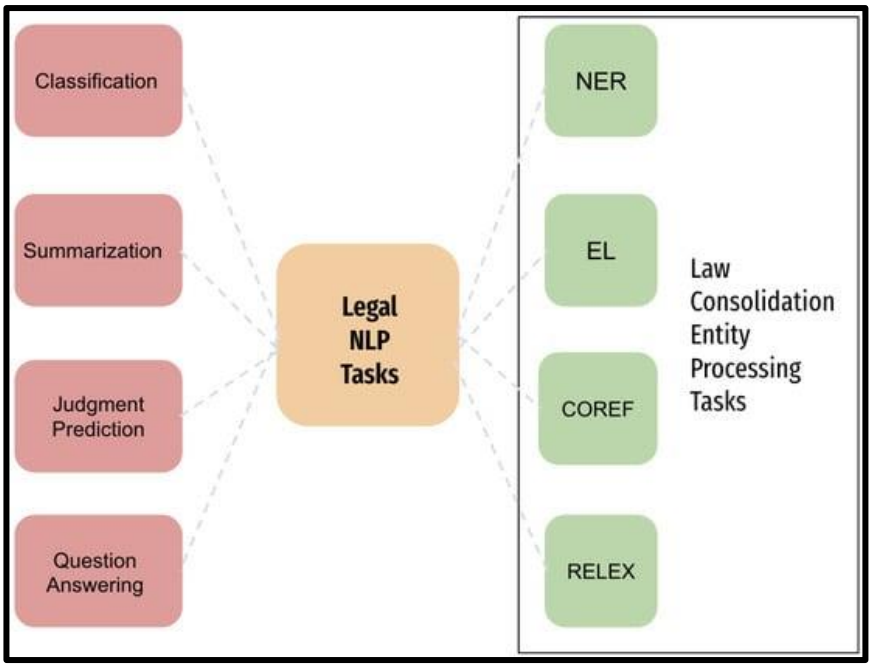

Figure: 2.3.1 Challenges in Legal NLP
(Sources: Krasadakis et al. 2024, p.3)

3. Methodology

It is possible to distinguish several steps in the application of NLP to the analysis of legal documents, each of which forms a part of the Legal AI framework. The main purpose of this methodology is to develop methods for the automation of the extraction, classification, and interpretation of content from a wide variety of legal documents including contracts, case law, statutes, and regulatory filings (Hassan et al. 2021). Using a range of the now commonly understood and functioning NLP approaches, it is possible to effect dramatic time and cost savings in the document review process, while still providing the accuracy that responding to legal procedures requires.

3.1 Collecting Data and its Cleaning

Data required for the analysis is the first step involved in the NLP methodology for legal document analysis. Legal documentation can be of unstructured type (case law, contracts) or semi-structured or structured (legal codes, regulations). For the quantitative analysis in this study, large databases of various kinds of legal texts need to be collected. Thus, freely available sources, including CourtListener, LexisNexis, and the LII have large collections of texts that can be useful in training and testing. After collecting the data, the text data collected needs to undergo preprocessing to get the text ready for analysis. Sometimes it is long, including domain specificities, a large number of terms, long sentences, and legal references. These techniques transform the data from an unstructured textual format into a format that is more understandable by the machines.

- ***Tokenization:*** In the legal field, some documents have to be separated into some constituent components to study them, and they are usually words and phrases. For instance, an agreement entered into and signed in court may be tokenized into clauses or terms.
- ***Stopword Removal:*** Particularly, they reduce the complexities by eliminating frequently used words like articles, conjunctions, and prepositions like, the, and, or of.
- ***Lemmatization and Stemming:*** Stems are obtained from the words to normalize the variations of the words (for instance, “regulated”, and “regulation” will be processed to “regulate”).

In particular, using particular legal terms, acronyms, and references is delicate and should not be tampered with during preprocessing. Sometimes it requires developing own dictionaries or utilizing the legal words or legal lexicon that will allow for the passage of legal meaning in text transformation.

3.2 Named Entity Recognition (NER)

Afterwards the process of text preprocessing, the Named Entity Recognition is used. NER is a crucial sub-task of NLP that deals with the recognition of the proper names, dates, numerical values, legal terms, etc. For instance, in a contract review, we can identify the contracting parties, the effective date of the contract, and financial provisions (Shen et al. 2021). In text mining for case laws, NER can extract the names of the case, the bench/judge, law/statute, and other related case precedence. In this way, legal professionals are not only able to obtain the necessary information without going through the physical document (Yu et al. 2020).
In this methodology, a pre-trained NER model either an industry-standard NER model legal NER from spaCy or LexNLP model, is fine-tuned for the industry in question on the legal dataset to get high specificity for entities in the legal industry.

3.3 Document Classification and Text Categorization

Text classification is another important step of the presented methodology, in which given documents can be classified according to their content or structure. Classification can be done at three different levels: document level, section level, or paragraph level based on the requirements of the undertaken task (Chen et al. 2022).

- Contracts can be categorized according to the nature of contracts such as employment contracts, non – competition contracts, and sales contracts.
- It is sub-categorized based on the area of jurisdiction, level of the court, or law area for instance constitutional law, tort law, and so on.
- Contract clauses are categorizable as warranties, indemnity, and liability clauses., which involves categorizing legal documents based on their content or structure. Classification can be performed at different levels—entire documents, sections, or paragraphs—depending on the task at hand (Aumiller et al. 2021, p. 2).

For example:

- Contracts can be classified into different types (e.g., employment agreements, non-disclosure agreements, purchase orders).
- Case law can be categorized based on jurisdiction, court level, or legal issue (e.g., constitutional law, tort law).
- Clauses within contracts can be classified as warranties, indemnifications, and liabilities.

According to Yadav et al. 2020, the common classifiers are Support Vector Machines (SVM), Random Forest, and Naive Bayes, which are used in the classification of our data. In the recent past, sophisticated models such as BERT (Bidirectional Encoder Representations from Transformers) have performed significantly well in the categorization of legal documents as they capture contextual correlation in text.

3.4 Text Summarization

Normally legal writings especially legal cases and legal writing such as agreements are very voluminous and are filled with similar information. To make the review process itself more efficient there is the use of auto text summarization. Such methods minimize the document to the most basic elements while retaining the information of central importance to these decisions.

Two types of summarization can be applied:

- Extractive Summarization: This method identifies and picks out the most relevant sentences from the document with ease. For example, an extractive summarizer might take out such parts of a contract as the bits that identify the roles of the parties, and when the contract may be terminated.
- Abstractive Summarization: OHS, on the other hand, complies with a smaller version of the document in its summarized form using other different words which are as if written by a human being. Whereas abstractive summarization is more difficult, it gives more elastic kinds of summaries (El-Kassas et al. 2021).

Current transformer models like BART for bidirectional and auto-regressive transformers and T5 texts transfer transformers are widely utilized when summarizing legal texts because of their impressive performance in text generation.

3.5 Sentiment and Semantic Analysis

This paper looks at the subtopics of Sentiment and Semantic Analysis.

In specific legal contexts thus, mostly involving case law or contract review working with the sentiment of the text and the semantics of the text can be rewarding. Sentiment analysis helps to determine how positive or negative a portion of legal texts might be and whether it might be relevant when evaluating an officiant's or judge's position or the decision to resolve a dispute.

In contrast, semantic roles concentrate on interpreting the actual meaning attached to legal terminology simplifying the clarification of possible ambiguities in clauses and facets of legal language (Salloum et al. 2020). To deduce vectors of words, NLP models such as Word2Vec or GloVe can be employed, word embeddings help in further analysis of the legal terms and the position of the concepts within the text.

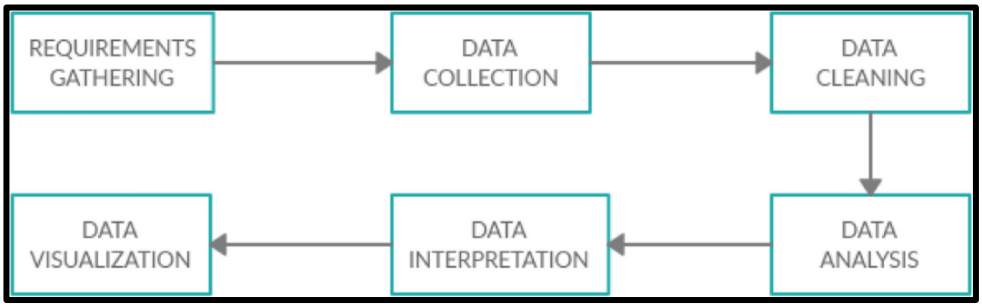

Figure: 3.5.1 Process of Sentiment and Semantic Analysis in NLP
(Source: Babu, 2022, p. 2)

4. Results and Discussion

NLP applied to legal document analysis presents several interesting outcomes in both number and quality. Applying NLP capabilities including NER, textual categorization, abstracting, and semantically analyzing the legal text may greatly enhance the experience of legal practitioners working with vast amounts of legal wordage. In this part, the authors focus on the results of using NLP in the context of legal document analysis, the difficulties, and enhancing prospects.

4.1 Reduced Costs and Cycle Time for Document Review

Undoubtedly, perhaps one of the easiest performance improvements to quantify when using NLP for legal documents is the increase in the speed of review. The advantages of Document Review can be understood by knowing the drawbacks of the traditional way of reviewing a document, which includes the thorough reading of contracts, court cases, or regulatory filings in the event of their length. Text mining helps to free up time to perform many routine and repetitive tasks, thereby, legal professionals can be more problem-solving. The improvements observed in different applications of the NLP legal document analysis are quite striking. NLP document review reduced the time spent on the documents by 80% as was evident in a study conducted on areas such as e-discovery and contract analysis. In legal research, there are tools, which enable the lawyer to identify the fit case precedents more quickly, cutting the research time by up to 60%.

4.2 Improved Accuracy in Legal Document Analysis

Beyond the spectrum of staff productivity, NLP also increases the effectiveness of text analysis in legal documents. It has the potential to have human interference, which is not preferred in many legal deals and services that involve large volumes of document handling where attention may gradually draw. Sustained systems, for example, deal with documents without fail and omission of any information no matter how small it is.
In the right hand, Named Entity Recognition (NER) which if appropriately implemented has a possibility of identifying legal text entities with precision up to 95% depending on the used dataset and quality of training. This means that given any text document, NLP models are capable of accurately identifying legal entities such as statutes, case citations, and clauses in a contract that may be more elaborate than often contain important formalistic legal content.

4.3 Challenges in NLP for Legal Document Analysis

However, several difficulties can still be pointed out in connection with the utilization of NLP in analyzing legal documents at the moment. Some of the biggest challenges that are pervasive include difficulty and uncertainty in legal language. Contracts specifically contain complex structures, contingency words, and province terms that are

hard to interpret for NLP models. For instance, wordings of the contract such as double-speak or cases that refer to prior case laws that are ambiguous may present a complicated picture to interpret.
Moreover, readership and asymmetry were unfamiliar to some experts, and jurisdictional differences are an issue with NLP models trained on large datasets. In the legal process, there is differences can exist between the country's legal systems especially based on the laws that may be practiced and interpreted differently by the courts or even regions. To be beneficial in these kinds of approaches, NLP models must be trained on datasets that reflect the differences found between the jurisdictions (Zadgaonkar et al. 2021). However, obtaining such data is a challenge an individual authority is bound by certain regulations, policies, and most importantly data privacy and security policies.

4.4 Future Directions and Improvements

Several areas for enhancement are identifiable in the use of NLP in the evaluation of legal documents. Some possible ways to work around the issues that are directly related to complexity and, as a result, ambiguity is to build the usage of the information processing rules and machine learning models that are combined. As datasets and the goal of open-source legal NLP tools develop, the models shall be more specific to different legal sub-tasks. Improved datasets will ensure that NLP systems are trained further and are more accurate during their use; this will be determined by different legal settings. The application of NLP for the analysis of legal documents has shown encouraging trends with substantial enhancement of both the efficiency and accuracy of the process. However, the study identifies certain limitations such as issues to do with legal language, problems with cross-jurisdiction, and the need for legal models to explain themselves. Thus, over time, NLP technology has just the ingredients for the recipe of further advancement in optimizing and automating legal processes.

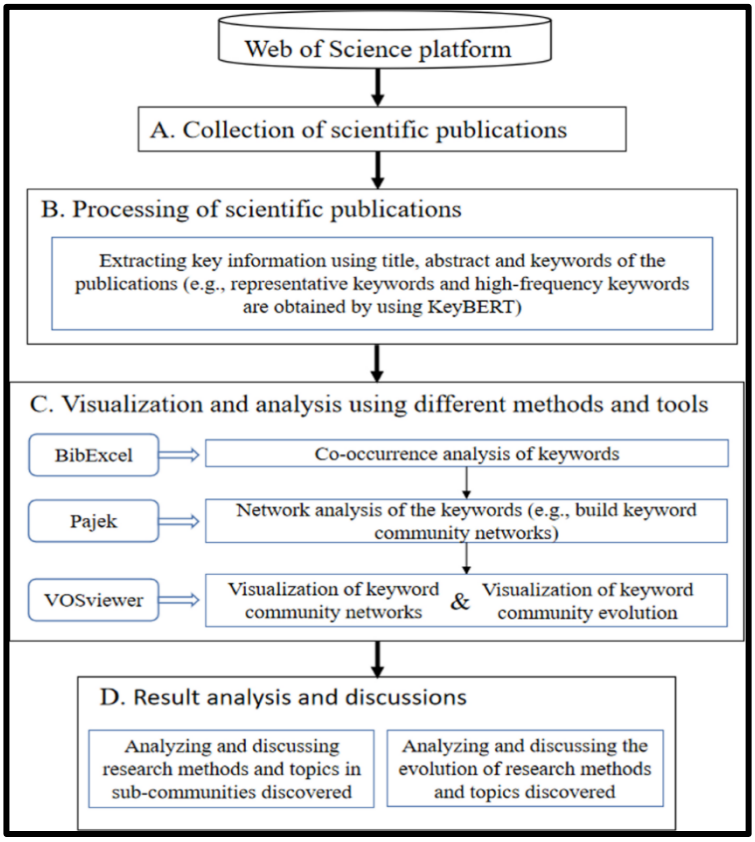

Figure: 4.4.1 Future Directions and Improvements of NLP (Source: Cui, 2023, p. 14)

5. Conclusion and Future Directions

In this study, several primary research questions were developed which were used to guide the study and inform the conclusions made based on the data collected and analyzed.
This may further be seen in the incorporation of Natural Language Processing (NLP) with Legal Document Analysis; this marks a very crucial addition to the progressivism of legal engineering since it has become a go-to tool for success in automating previously manual and very time-consuming processes in the legal fields. Through methods like Named Entity Recognition (NER), text classification, and summarization, NLP accelerates and improves the effectiveness of document sifting, contract review as well as research. AI & NLP have greatly lowered the time needed to complete e-discovery, contract parsing, and review of case laws among others while at the same time lowering the likelihood of errors. On this account, it is worth googling ahead for the further development of the so-called hybrid approaches that at the same time rely on traditional rule-based systems as well as on machine learning applications to work with the texts of a legal framework. Further, multilingual NLP models and open-source legal datasheets will create even more opportunities for the use of NLP for international legal systems. The future development and growth mean that NLP will be able to revolutionize the legal sector and make legal services cheaper, faster, and more accessible for people all over the world. However, in this research, legal complexity and legal language, legal differences between jurisdictions, and the opacity of deep learning models are still open issues. For natural language processing to reach its potential in legal computing, more efforts should be directed at making models more interpretable, creating and curating better legal corpora, and adapting NLP for systems of law across the world.

6. Acknowledgments

The author also would like to say thanks to the Department of Computer Science at [Insert University Name] for the grant and support during work on this research. Thanks go to [Insert Advisor's Name] once again for his support and wisdom in the development of this paper. In the same way, the author gets inspired by several legal workers and technical people who assisted her in the analysis of legal documents and Natural Language Processing. Finally, thanks go to the authors of the open-source tools used in NLP on which much of the work covered in this paper relies.

Reference List

Ahmad, Z., Shahid Khan, A., Wai Shiang, C., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies*, *32*(1), e4150. https://onlinelibrary.wiley.com/doi/pdfdirect/10.1002/ett.4150

Aumiller, D., Almasian, S., Lackner, S., & Gertz, M. (2021, June). Structural text segmentation of legal documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (pp. 2-11). https://arxiv.org/pdf/2012.03619

Babu, N.V., Kanaga, E.G.M. Sentiment Analysis in Social Media Data for Depression Detection Using Artificial Intelligence: A Review. *SN COMPUT. SCI.* **3**, 74 (2022). https://link.springer.com/content/pdf/10.1007/s42979-021-00958-1.pdf

Baviskar, D., Ahirrao, S., Potdar, V., & Kotecha, K. (2021). Efficient automated processing of the unstructured documents using artificial intelligence: A systematic literature review and future directions. *IEEE Access*, *9*, 72894-72936. https://ieeexplore.ieee.org/iel7/6287639/6514899/09402739.pdf

Bommarito II, M. J., Katz, D. M., & Detterman, E. M. (2021). LexNLP: Natural language processing and information extraction for legal and regulatory texts. In *Research Handbook on big data law* (pp. 216-227). Edward Elgar Publishing. https://arxiv.org/pdf/1806.03688

Cavalieri, D. C., Palazuelos-Cagigas, S. E., Bastos-Filho, T. F., & Sarcinelli-Filho, M. (2020). Combination of language models for word prediction: An exponential approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(9), 1481-1494. https://www.researchgate.net/profile/Daniel-Cavalieri/publication/299472244_Combination_of_Language_Models_for_Word_Prediction_An_Exponential_Approach/links/59db9bcaaca2728e2017ea8a/Combination-of-Language-Models-for-Word-Prediction-An-Exponential-Approach.pdf

Chen, H., Wu, L., Chen, J., Lu, W., & Ding, J. (2022). A comparative study of automated legal text classification using random forests and deep learning. *Information Processing & Management*, *59*(2), 102798. https://www.sciencedirect.com/science/article/am/pii/S0306457321002764

Chowdhary, K., & Chowdhary, K. R. (2020). Natural language processing. *Fundamentals of artificial intelligence*, 603-649. https://pureportal.strath.ac.uk/files/131112/strathprints002611.pdf

Cui, J., Wang, Z., Ho, SB. *et al.* Survey on sentiment analysis: evolution of research methods and topics. *Artif Intell Rev* **56**, 8469–8510 (2023). https://link.springer.com/content/pdf/10.1007/s10462-022-10386-z.pdf

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert systems with applications*, *165*, 113679. https://www.researchgate.net/profile/Hoda-Mohamed-9/publication/342878502_Automatic_Text_Summarization_A_Comprehensive_Survey/links/619776b407be5f31b7998a97/Automatic-Text-Summarization-A-Comprehensive-Survey.pdf

Ganguly, D., Conrad, J. G., Ghosh, K., Ghosh, S., Goyal, P., Bhattacharya, P., ... & Paul, S. (2023, March). Legal IR and NLP: the history, challenges, and state-of-the-art. In *European Conference on Information Retrieval* (pp. 331-340). Cham: Springer Nature Switzerland. http://conradweb.org/~jackg/pubs/ECIR2023_Legal_IR_NLP_Tutorial.pdf

Hassan, F. U., Le, T., & Lv, X. (2021). Addressing legal and contractual matters in construction using natural language processing: A critical review. *Journal of Construction Engineering and Management*, *147*(9), 03121004. https://www.researchgate.net/profile/Tuyen_Le24/publication/354284117_Addressing_Legal_and_Contractual_Matters_in_Construction_Using_Natural_Language_Processing_A_Critical_Review/links/62b690f1d49f803365b81035/Addressing-Legal-and-Contractual-Matters-in-Construction-Using-Natural-Language-Processing-A-Critical-Review.pdf

Krasadakis, P., Sakkopoulos, E., & Verykios, V. S. (2024). A survey on challenges and advances in natural language processing with a focus on legal informatics and low-resource languages. *Electronics*, *13*(3), 648. https://www.mdpi.com/2079-9292/13/3/648/pdf?version=1707035028

Salloum, S. A., Khan, R., & Shaalan, K. (2020). A survey of semantic analysis approaches. In *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2020)* (pp. 61-70). Springer International Publishing. https://www.researchgate.net/profile/Khaled-Shaalan-

2/publication/340099721_A_Survey_of_Semantic_Analysis_Approaches/links/653e833af7d021785f206e32/A-Survey-of-Semantic-Analysis-Approaches.pdf

Sarker, I.H. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN COMPUT. SCI.* **2**, 420 (2021). https://link.springer.com/content/pdf/10.1007/s42979-021-00815-1.pdf

Shen, Y., Ma, X., Tan, Z., Zhang, S., Wang, W., & Lu, W. (2021). Locate and label: A two-stage identifier for nested named entity recognition. *arXiv preprint arXiv:2105.06804*. https://arxiv.org/pdf/2105.06804

Yadav, B. P., Ghate, S., Harshavardhan, A., Jhansi, G., Kumar, K. S., & Sudarshan, E. (2020, December). Text categorization performance examination using machine learning algorithms. In *IOP Conference Series: Materials Science and Engineering* (Vol. 981, No. 2, p. 022044). IOP Publishing. https://iopscience.iop.org/article/10.1088/1757-899X/981/2/022044/pdf

Yu, J., Bohnet, B., & Poesio, M. (2020). Named entity recognition as dependency parsing. *arXiv preprint arXiv:2005.07150*. https://arxiv.org/pdf/2005.07150

Zadgaonkar, A. V., & Agrawal, A. J. (2021). An overview of information extraction techniques for legal document analysis and processing. *International Journal of Electrical & Computer Engineering (2088-8708)*, *11*(6). https://www.academia.edu/download/90993734/15261.pdf

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., & Sun, M. (2020). How does NLP benefit legal system: A summary of legal artificial intelligence. *arXiv preprint arXiv:2004.12158*. https://arxiv.org/pdf/2004.12158