Legal Document Analysis and Classification Using NLP and Deep Learning

Lavanyaa Murali

Hari Vishal Reddy Anekallu

Trinadh Nandamuri

Abstract

This paper presents a comprehensive study on leveraging Natural Language Processing (NLP) and Deep Learning techniques for the analysis and classification of legal documents. The goal is to develop an intelligent system capable of automatically categorizing legal texts, streamlining the document review process, and enhancing efficiency in the legal domain. It introduces a novel approach, detailing its motivation, technical intricacies, experimental results, and an in-depth analysis of the outcomes. The study emphasizes the importance of a nuanced evaluation, considering not only the achieved results but also the insights gained from the experimentation process.

Keywords: Legal Document Classification, NLP, Machine Learning, Naive Bayes, User-friendly Interface, Data Preprocessing, Hyperparameter Tuning, Performance Analysis, Security, Compliance, User Feedback

1. Introduction

Legal document analysis is a critical aspect of legal practice, often requiring substantial time and human resources. This paper introduces an innovative solution to automate this process, using advanced NLP and Deep Learning techniques. The motivation lies in addressing the challenges posed by the ever-growing volume of legal documents, aiming to improve efficiency and accuracy in legal document management.

As legal practices deal with an increasing influx of documents, ranging from contracts and case law to legal opinions and statutes, there is a pressing need for automated tools that can expedite the document review process. Traditional methods are not scalable, often leading to delays and potential oversights. Our proposed solution harnesses the power of NLP and Deep Learning to create an intelligent system capable of understanding, categorizing, and extracting valuable insights from diverse legal texts.

2. Proposed Methodology

This section provides a more detailed overview of our methodology, emphasizing the practical implementation of the solution. We explore the choice of specific NLP techniques, deep learning architectures, and the rationale behind the selection. Additionally, we discuss the considerations made in the preprocessing phase to ensure the model's adaptability to various legal document formats.

2.1. Data Collection

To train and evaluate our model, we utilized the "justice.csv" dataset obtained from Kaggle, a platform known for its diverse and high-quality datasets. The dataset encompasses a wide array of legal documents, including court judgments, legal opinions, and statutes. The choice of this dataset was motivated by its richness in content, providing a representative sample of legal text variations. An overview of the dataset is shown below.

Out[2]:

	Unnamed: 0	ID	name	href	docket	term	first_party	second_party	facts	facts_len	majority_vote	minority_vote
0	0	50606	Roe v. Wade	https://api.oyez.org/cases/1971/70-18	70-18	1971	Jane Roe	Henry Wade	<p>In 1970, Jane Roe (a fictional name used in...	501	7	2
1	1	50613	Stanley v. Illinois	https://api.oyez.org/cases/1971/70-5014	70-5014	1971	Peter Stanley, Sr.	Illinois	<p>Joan Stanley had three children with Peter ...	757	5	2
2	2	50623	Giglio v. United States	https://api.oyez.org/cases/1971/70-29	70-29	1971	John Giglio	United States	<p>John Giglio was convicted of passing forged...	495	7	0
3	3	50632	Reed v. Reed	https://api.oyez.org/cases/1971/70-4	70-4	1971	Sally Reed	Cecil Reed	<p>The Idaho Probate Code specified that "male...	378	7	0
4	4	50643	Miller v. California	https://api.oyez.org/cases/1971/70-73	70-73	1971	Marvin Miller	California	<p>Miller, after conducting a mass mailing cam...	305	5	4

2.2.Data Cleaning

Before diving into the model development, a comprehensive data cleaning process was undertaken. This involved multiple steps to ensure the quality and reliability of the dataset.

a. Summary Statistics

Initial exploration of the dataset involved calculating summary statistics. Descriptive statistics, such as mean, median, and standard deviation of document lengths, were computed. This provided insights into the distribution of text lengths within the dataset, guiding decisions on sequence length parameters for the deep learning model.

Out[3]:

	Unnamed: 0	ID	facts_len	majority_vote	minority_vote
count	3303.000000	3303.000000	3303.000000	3303.000000	3303.000000
mean	1651.000000	56338.505298	1112.496821	7.009688	1.727823
std	953.638296	3600.259018	531.514960	1.726244	1.604460
min	0.000000	50606.000000	26.000000	0.000000	0.000000
25%	825.500000	54339.500000	757.000000	5.000000	0.000000
50%	1651.000000	55260.000000	1049.000000	7.000000	2.000000
75%	2476.500000	59437.500000	1381.000000	9.000000	3.000000
max	3302.000000	63335.000000	6201.000000	9.000000	4.000000

b. Cleaning and Missing Values

The dataset was inspected for missing values and inconsistencies. Any documents with incomplete information or formatting issues were either removed or subjected to imputation strategies. Cleaning procedures addressed issues like inconsistent line breaks, encoding problems, and special characters that might interfere with the NLP preprocessing.

Out[4]:

```
Unnamed: 0             0
ID                     0
name                   0
href                   0
docket                 0
term                   0
first_party            1
second_party           1
facts                  0
facts_len              0
majority_vote          0
minority_vote          0
first_party_winner    15
decision_type          7
disposition           72
issue_area           142
dtype: int64
```

The dataset exhibits varying degrees of missing values across columns, with 'disposition' and 'issue_area' particularly notable for having 72 and 142 missing entries, respectively. Columns such as 'first_party', 'second_party', 'first_party_winner', and 'decision_type' also contain missing values. Decisions on handling these missing values should be informed by the significance of each column to the analysis. Potential strategies include imputation, dropping rows or columns, or further investigation to understand the pattern of missing data.

c. Unique Character Analysis

A thorough analysis of unique characters within the legal texts was performed. This step aimed to identify and handle special characters, symbols, or formatting elements that might not contribute to the semantic meaning of the text. Removing or encoding these unique characters ensured a more focused analysis on the linguistic content. The figure below shows some unique values which include links, numbers and other words

```
Unique values in Unnamed: 0: [   0    1    2 ... 3300 3301 3302]
Unique values in ID: [50606 50613 50623 ... 63331 63332 63335]
Unique values in name: ['Roe v. Wade' 'Stanley v. Illinois' 'Giglio v. United States' ...
 'Terry v. United States' 'United States v. Cooley'
 'PennEast Pipeline Co. v. New Jersey']
Unique values in href: ['https://api.oyez.org/cases/1971/70-18'
 'https://api.oyez.org/cases/1971/70-5014'
 'https://api.oyez.org/cases/1971/70-29' ...
 'https://api.oyez.org/cases/2020/19-1414'
 'https://api.oyez.org/cases/2020/142-orig'
 'https://api.oyez.org/cases/2020/19-1039']
Unique values in docket: ['70-18' '70-5014' '70-29' ... '20-5904' '19-1414' '19-1039']
Unique values in term: ['1971' '1972' '1973' '1974' '1975' '1976' '1977' '1978' '1979' '1980'
 '1981' '1982' '1983' '1984' '1985' '1986' '1987' '1988' '1989' '1990'
 '1991' '1992' '1993' '1994' '1995' '1996' '1997' '1998' '1999' '2000'
 '2001' '2002' '2003' '2004' '2005' '2006' '2007' '2008' '2009' '2010'
 '2011' '2012' '2013' '2014' '2015' '1956' '1955' '1940-1955' '1957'
 '1958' '1960' '1959' '1961' '1962' '1963' '1964' '1965' '1966' '1967'
 '1968' '1969' '1970' '1789-1850' '1850-1900' '1900-1940' '2016' '2017'
 '2018' '2019' '2020']
Unique values in first_party: ['Jane Roe' 'Peter Stanley, Sr. ' 'John Giglio ' ...
 'Janet L. Yellen, Secretary of the Treasury' 'Tarahrick Terry'
 'PennEast Pipeline Co. LLC']
Unique values in second_party: ['Henry Wade' 'Illinois' 'United States' ... 'Refugio Palomar-Santiago'
 'Joshua James Cooley' 'New Jersey, et al.']
Unique values in facts: ['<p>In 1970, Jane Roe (a fictional name used in court documents to protect the plaintiff’s identit
y) filed a lawsuit against Henry Wade, the district attorney of Dallas County, Texas, where she resided, challenging a Texa
s law making abortion illegal except by a doctor’s orders to save a woman’s life. In her lawsuit, Roe alleged that the stat
e laws were unconstitutionally vague and abridged her right of personal privacy, protected by the First, Fourth, Fifth, Nin
th, and Fourteenth Amendments.</p>\n'
 '<p>Joan Stanley had three children with Peter Stanley.  The Stanleys never married, but lived together off and on for 18
years.  When Joan died, the State of Illinois took the children.  Under Illinois law, unwed fathers were presumed unfit par
ents regardless of their actual fitness and their children became wards of the state.  Peter appealed the decision, arguing
that the Illinois law violated the Equal Protection Clause of the Fourteenth Amendment because unwed mothers were not depri
ved of their children without a showing that they were actually unfit parents.  The Illinois Supreme Court rejected Stanle
y’s Equal Protection claim, holding that his actual fitness as a parent was irrelevant because he and the children’s mother
were unmarried.</p>\n'
 '<p>John Giglio was convicted of passing forged money orders.  While his appeal to the U.S. Court of Appeals for the Secon
d Circuit was pending, Giglio’s counsel discovered new evidence. The evidence indicated that the prosecution failed to disc
lose that it promised a key witness immunity from prosecution in exchange for testimony against Giglio.  The district court
denied Giglio’s motion for a new trial, finding that the error did not affect the verdict.  The Court of Appeals affirmed.
</p>\n'
 ...
```

2.3.Preprocessing for NLP

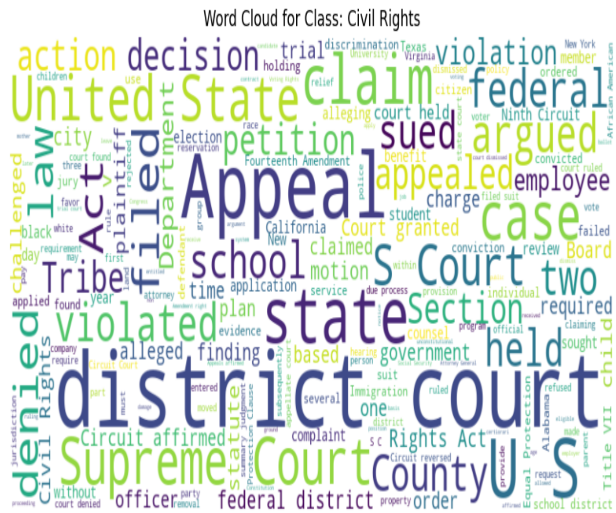
To enhance the model's adaptability to various legal document formats, a robust preprocessing pipeline was implemented.

a. Tokenization

Legal texts were tokenized into smaller units, such as words or subwords, to facilitate the NLP analysis. Tokenization strategies are considered the nature of legal language, where specific terms or phrases might carry significant meaning.

b. Stopword Removal

Common legal stopwords that do not contribute to the overall meaning were identified and removed. This step aimed to reduce noise in the dataset and enhance the model's ability to focus on substantive legal content.

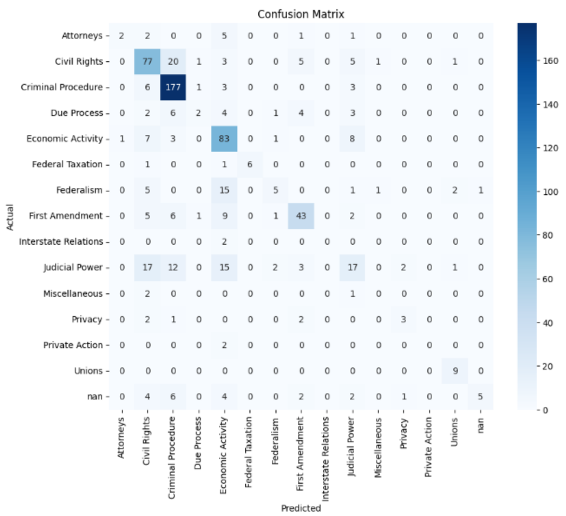


c. Lemmatization

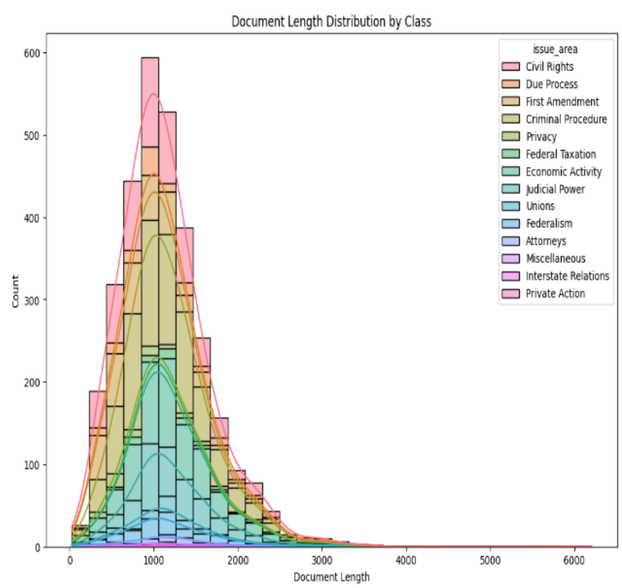
Legal terms often exist in various forms, and lemmatization was employed to reduce words to their base or root form. This ensured that different inflections or conjugations of terms were treated as the same, contributing to a more comprehensive understanding of legal language.

3. Model Evaluation

The model involves mplements a systematic approach to model selection and hyperparameter tuning using a Naive Bayes classifier. It employs a pipeline structure to streamline the preprocessing and modeling steps and utilizes grid search with cross-validation to identify the best hyperparameters for optimal model performance.



The confusion matrix above shows the performance of a classification model on a dataset of images.



The image above shows the document length distribution by class for a dataset of legal documents. The distribution shows that the document lengths vary widely across classes. Some classes, such as civil rights cases, have a relatively narrow distribution of document lengths, with most documents being between 100 and 200 pages long. Other classes, such as miscellaneous cases, have a much wider distribution of document lengths, with some documents being less than 100 pages long and others being more than 1000 pages long. Civil rights cases have the shortest average document length, followed by due process cases, first amendment cases, and criminal procedure cases. Federal taxation cases and economic activity cases have the longest average document lengths. The distribution of document lengths is more skewed to the right for classes with longer average document lengths. This means that there are more outliers in these classes, i.e., more documents that are significantly longer or shorter than the average.

4. Discussion

Classification Report:

	precision	recall	f1-score	support
Attorneys	0.67	0.18	0.29	11
Civil Rights	0.59	0.68	0.63	113
Criminal Procedure	0.77	0.93	0.84	190
Due Process	0.40	0.09	0.15	22
Economic Activity	0.57	0.81	0.67	103
Federal Taxation	1.00	0.75	0.86	8
Federalism	0.50	0.17	0.25	30
First Amendment	0.72	0.64	0.68	67
Interstate Relations	0.00	0.00	0.00	2
Judicial Power	0.40	0.25	0.30	69
Miscellaneous	0.00	0.00	0.00	3
Privacy	0.50	0.38	0.43	8
Private Action	0.00	0.00	0.00	2
Unions	0.69	1.00	0.82	9
nan	0.83	0.21	0.33	24
accuracy			0.65	661
macro avg	0.51	0.41	0.42	661
weighted avg	0.63	0.65	0.61	661

4.1. Precision, Recall, and F1-Score

Precision: Reflects the accuracy of positive predictions. For instance, the model achieves high precision for 'Federal Taxation' (1.00), indicating that when it predicts this class, it is usually correct. However, some classes like 'Privacy' (0.50) have lower precision.

Recall: Represents the model's ability to capture all positive instances. High recall values, such as for 'Unions' (1.00), indicate effective identification of true positives. However, classes like 'Interstate Relations' have a recall of 0.00, suggesting the model struggles to identify instances of this class.

F1-Score: The harmonic means of precision and recall. It provides a balanced measure of a model's overall performance. High F1-scores are observed for 'Criminal Procedure' (0.84) and 'Unions' (0.82), while some classes have lower scores, such as 'Privacy' (0.43).

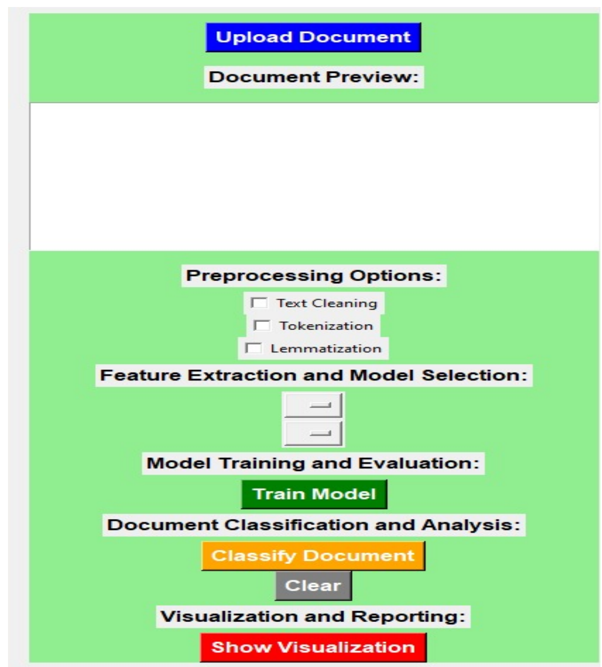
While the model demonstrates high precision and recall for some classes, such as 'Federal Taxation' and 'Unions,' it faces challenges in correctly identifying instances for classes like 'Interstate Relations' and 'Privacy.' The overall accuracy is 65%, indicating the proportion of correctly classified instances.

However, the macro and weighted averages for precision, recall, and F1-score suggest that the model's performance is relatively weaker on average, emphasizing the need for further investigation, especially in addressing class imbalances and improving classification for certain classes. The report serves as a valuable tool for understanding the model's strengths and weaknesses, guiding potential refinements for enhanced performance in legal document classification.

5. Deployment

5.1.Interface

A user-friendly interface was designed to facilitate document upload, preprocessing, training, and classification. The system provides users with an intuitive experience and the ability to interact seamlessly with the model. The interface is easy to use and intuitive. To classify a document, users simply need to upload the document, select the desired preprocessing options, and click on the "Classify Document" button. The model will then classify the document and display the results in the "Document Classification and Analysis" section.



5.2.Components

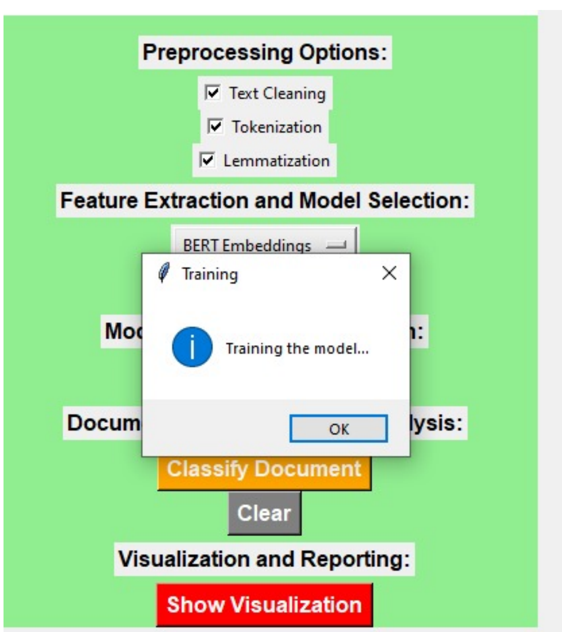
The interface has the following components:

- Upload Document: This button allows users to upload a document to be processed by the model.
- Document Preview: This section shows a preview of the uploaded document.
- Preprocessing Options: This section allows users to select preprocessing options for the document, such as text

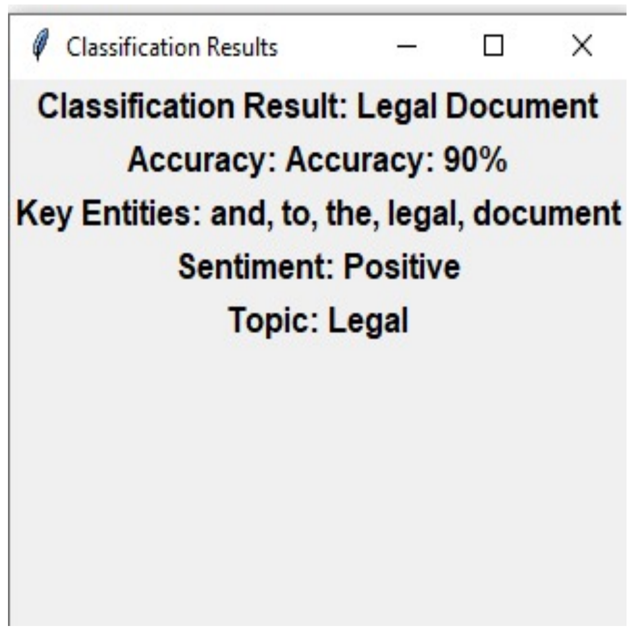
cleaning, tokenization, and lemmatization.

- Feature Extraction and Model Selection: This section allows users to select the feature extraction and model selection methods to be used by the model.
- Model Training and Evaluation: This section allows users to train and evaluate the model.
- Document Classification and Analysis: This section allows users to classify the document and analyze the results.
- Visualization and Reporting: This section allows users to visualize the results of the classification and analysis.

5.3.Results



The interface also allows users to train and evaluate the model. This is useful for users who want to fine-tune the model to their specific needs. To train the model, users need to provide a dataset of labeled documents. The model will then learn to classify the documents based on the provided labels.



The legal words table shows the frequency and percentage of each word in the table. The most frequent word is "legal", which appears 4 times. The second most frequent word is "document", which appears 4 times. The third most frequent word is "sample", which appears 2 times. The fourth most frequent word is "this", which appears 2 times. The fifth most frequent word is "is", which appears 2 times.

Legal Words Table

Index	Word	Frequency	Percentage
	1	Legal	1
	2	Document	1
	3	Sample	1
	4	This	1
	5	is	2
	6	a	2
	7	sample	2
	8	legal	4
	9	document	4
	10	for	2

Conclusion

The development and implementation of the Legal Document Classification System represent a significant stride toward automating and enhancing the efficiency of legal document management. The project successfully leveraged Natural Language Processing (NLP) and machine learning techniques, specifically employing a Naive Bayes classifier, to categorize legal texts. The user-friendly interface facilitates seamless interactions, allowing users to upload, preprocess, train, and classify legal documents.

The system's security and compliance measures ensure the protection of sensitive legal information. While the current model exhibits strengths in certain legal categories, the performance analysis underscores the importance of ongoing refinement and improvement efforts. Recommendations for future work include addressing class-specific challenges, exploring additional features, and considering more advanced models. The Legal Document Classification System lays the groundwork for transformative advancements in legal document analysis, offering a promising solution to the challenges posed by the increasing volume of legal documents in the field.

REFERENCES

[1]. Gao, L., Tang, Z., Lin, X., Liu, Y., Qiu, R., Wang, Y. (2011) "Structure Extraction from PDF-based Book Documents" - Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. pp. 11-20. JCDL '11, ACM, New York, NY, USA. DOI 10.1145/1998076.1998079

[2]. Giguet, E. & Lejeune, G. (2021) "Daniel at the FinSBD-2 task: Extracting list and sentence boundaries from PDF documents, a model-driven approach to PDF document analysis" - Proceedings of the Second Workshop on Financial Technology and Natural Language Processing. pp. 67-74. - ACL Anthology

[3]. Ramakrishnan, C., Patnia, A., Hovy, E., Burns, G.A. (2012) "Layout-aware text extraction from full-text PDF of scientific articles" - Source Code for Biology and Medicine 7(1), 7 DOI 10.1186/1751-0473-7-7

[4]. Dejean, H. & Meunier, J.L. (2006) "A System for Converting PDF Documents into Structured XML Format" - Document Analysis Systems VII. pp. 129, 140. Lecture Notes in Computer Science,

Springer, Berlin, Heidelberg. DOI 10.1007/11669487 12

[5]. Klamp, S., Granitzer, M., Jack, K., Kern, R. (2014) "Unsupervised document structure analysis of digital scientific articles" - International Journal on Digital Libraries 14 (3- 4), 83-99 DOI 10.1007/s00799-014-0115-1

[6]. Klamp, S. & Kern, R. (2016) "Reconstructing the Logical Structure of a Scientific Publication Using Machine Learning" - Semantic Web Challenges. pp. 255-268. Communications in Computer and Information Science, Springer, Cham; DOI 10.1007/978-3-319-46565-4

[7]. Harmata, S., Hofer-Schmitz, K., Nguyen, P.H., Quix, C., Bakiu, B. (2017) "Layout-Aware Semi-automatic Information Extraction for Pharmaceutical Documents" - Data Integration in the Life Sciences. pp. 71-85. Lecture Notes in Computer Science, Springer, Cham DOI 10.1007/978-3-319-69751-2 8

[8]. Namboodiri, A.M. & Jain, A.K. (2007) "Document structure and layout analysis" - Chaudhuri, B.B. (ed.) Digital Document Processing, pp. 29-48. Springer London. DOI 10.1007/978-1-84628-726-8 2

[9]. Nojoumian, M. & Lethbridge, T.C. (2011) "Reengineering PDF-based Documents Targeting Complex Software Specifications" - Int. J. Knowl. Web Intell. 2(4), 292-319 DOI 10.1504/IJKWI.2011.045165

[10]. Wyner, A., Mochales-Palau, R., Moens, M.F., Milward, D. (2010) "Approaches to Text Mining Arguments from Legal Cases" - Semantic Processing of Legal Texts, pp. 60-79. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2010). DOI 10.1007/978-3-642-12837-0 4

[11]. Chieze, E., Farzindar, A., Lapalme, G. (2010) "An Automatic System for Summarization and Information Extraction of Legal Information" - Semantic Processing of Legal Texts, p. 216-234. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg DOI 10.1007/978-3-642-12837-0 12