IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2023.0322000

Legal Natural Language Processing from 2015-2022: A Comprehensive Systematic Mapping Study of Advances and Applications

Ernesto Quevedo[1], Tomas Cerny[2], Alejandro Rodriguez [1], Pablo Rivas[1], (Senior Member, IEEE), Jorge Yero [1], Korn Sooksatra [1], Alibek Zhakubayev [1], Davide Taibi [3,4]

[1]Department of Computer Science, School of Engineering and Computer Science, Baylor University, Waco, Texas 76798, USA
[2]Systems and Industrial Engineering, University of Arizona, Tucson, Arizona, 85721, USA
[3]University of Oulu. Oulu, Finland (e-mail: davide.taibi@oulu.fi)
[4]Tampere University. Tampere, Finland

Corresponding author: Ernesto Quevedo (e-mail: ernesto_quevedo1@baylor.edu).

This article is based on work supported by the National Science Foundation under Grant No. 2039678, 2136961, and 2210091. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

ABSTRACT

The surge in legal text production has amplified the workload for legal professionals, making many tasks repetitive and time-consuming. Furthermore, the complexity and specialized language of legal documents pose challenges not just for those in the legal domain but also for the general public. This emphasizes the potential role and impact of Legal Natural Language Processing (Legal NLP). Although advancements have been made in this domain, particularly after 2015 with the advent of Deep Learning and Large Language Models (LLMs), a systematic exploration of this progress until 2022 is nonexistent. In this research, we perform a Systematic Mapping Study (SMS) to bridge this gap. We aim to provide a descriptive statistical analysis of the Legal NLP research between 2015 and 2022. Categorize and sub-categorize primary publications based on their research problems. Identify limitations and areas of improvement in current research. Using a robust search methodology across four reputable indexers, we filtered 536 papers down to 75 pivotal articles. Our findings reveal the diverse methods employed for tasks such as Multiclass Classification, Summarization, and Question Answering in the Legal NLP field. We also highlight resources, challenges, and gaps in current methodologies and emphasize the need for curated datasets, ontologies, and a focus on inherent difficulties like data accessibility. As the legal sector gradually embraces Natural Language Processing (NLP), understanding the capabilities and limitations of Legal NLP becomes vital for ensuring efficient and ethical application. The research offers insights for both Legal NLP researchers and the broader legal community, advocating for continued advancements in automation while also addressing ethical concerns.

INDEX TERMS Systematic-Mapping-Study, Legal-NLP, Deep Learning

I. INTRODUCTION

Much of the resources in the legal field are stored in text form. Privacy policies, state regulations, contracts, and judgments are examples of this storage. Recent advancements in Natural Language Processing (NLP) have been identified as suitable for enhancing and automating tasks in the legal domain. In these discussions, the term "Legal NLP" is frequently used. This term emphasizes the application of NLP methods, specifically in legal contexts.

An increasing volume of legal texts is being produced. This surge in legal information adds to the workload of legal professionals, making many tasks repetitive. Deep analysis and understanding are essential for numerous tasks within the legal domain. Consequently, even for experts, considerable time is required to retrieve the appropriate legal documents. The intricate nature of legal language occasionally leads to hesitations among professionals due to its inherent ambiguities. Comprehending such complex documents becomes even more challenging for individuals without legal expertise. A clear example is that many users routinely agree with organi-

zational security and privacy policies [R1], ignore them [R2], and lack security awareness [R3, R4] to comprehend what they have agreed to.

Legal complexity, ranging from descriptive to empirical methods, is explored in a study by Ruhl et al. [R5]. The research illustrates that in 2012, American entities spent 6.1 billion hours and 168 billion dollars complying with the United States of America (USA) Tax Code, a document containing nearly 4 million words. This text underwent over 5000 amendments between 2001 and 2012.

Legal documents frequently utilize a specialized language that necessitates expert interpretation. For most native speakers, such language can appear as indecipherable jargon. Consequently, individuals without a legal background often struggle with interpreting legal texts and responding to simple inquiries. The extensive nature of these documents further compounds this challenge. Automating the process of answering questions or summarizing legal documents becomes crucial. Challenges with language translation arise as in other sectors dealing with human languages in text form.

In the context of complexity, references can be made to studies such as Ruhl et al. [R5] concerning the USA's Tax Code. Moreover, Friedrich et al.[R6] investigate the linguistic attributes of legal codes across diverse nations and traditions, drawing from physics, algorithmic complexity theory, and information theory. Their research indicates that distinct legal texts, like acts, regulations, or literature, occupy specific areas on the complexity-entropy plane, which is defined by the measures of information and complexity.

Significant advancements in Legal NLP have been observed, particularly since 2015. This growth is attributed to the high performance of Deep Learning techniques in NLP and the invention of Large Language Models (LLMs), which have gained tremendous popularity in recent years due to their impressive results and applications in multiple domains. However, a comprehensive, systematic understanding of this progress from 2015 until 2022 remains uncharted. This paper endeavors to bridge this gap.

To systematically categorize and synthesize this expansive research area, a Systematic Literature Review (SLR) in the form of a Systematic Mapping Study (SMS) is undertaken. The focus is directed towards quantitative and qualitative dimensions, encompassing statistics and a meticulous categorization and examination of the extant literature. The efficacy of SMS in research has been demonstrated in its capacity to structure and categorize existing findings while also spotlighting areas requiring enhancement [R7].

Our objectives in this paper are:

- Present descriptive statistics of existing published research on Legal NLP published between 2015 and 2022.
- Systematically categorize and sub-categorize the primary publications based on the research problem they approach.
- Summarize the results presented in these studies in each category and detect current limitations and areas of improvement.

This paper is structured as follows. First, we present the related studies. Second, we explain our methodology for conducting this study. Next, we offer the results of our study with categorized and separate results per research question in our Systematic Mapping Study. After that, we present a discussion section, followed by the threads-to-validity section. Finally, we conclude the paper with the conclusions section.

II. RELATED WORK

Because our work takes the form of a Systematic Mapping Study, other related studies and surveys were excluded from our SMS. Previous literature reviews have focused on particular points, datasets, and approaches in the Legal NLP field. A comprehensive presentation of the current state of the art in every subarea, including its limitations and resources, is essential. This section showcases surveys conducted in the Legal NLP field and the two identified Systematic Literature Reviews (SLRs) in this domain.

Table 1 summarizes the research questions asked by other studies and the number of primary studies included in them. Some of these related works are surveys that do not follow a clear methodology. Therefore, they do not provide an exact number of primary studies analyzed, therefore, we will approximate those cases. Additionally, we provide Table 2 with a comparison with advantages and disadvantages among the related works.

In the study by Chalkidis et al. [R8], emphasis was placed on Deep Learning (DL) applications in law and on legal word embeddings from extensive corpora. An overview of DL architectures and feature representations in the Legal NLP field up to 2018 was given. Main tasks like Text Classification, Information Extraction, and Information Retrieval were outlined. Convolutional Neural Networks (CNN)-based models were identified as dominant for information retrieval tasks. A shift from detailed feature engineering to simpler networks, either standalone CNNs or combined with Long Short-Term Memories (LSTMs), was noted. Performance was enhanced by integrating features from methods such as Latent Dirichlet Allocation (LDA), BM25, and established word distances.

In the study by Montelongo et al. [R9], a bibliometric review was provided from 1987 to 2020 on tasks performed in the legal domain using Deep Learning. Emphasis was on Legal NLP tasks such as Classification, Feature Extraction, Information Extraction, Information Retrieval, Preprocessing, Summarization, Text generation, and Theoretical aspects. However, a greater focus was on article statistics and their growth rather than on the current state-of-the-art trends in the Legal NLP field.

In the review by Sheik et al. [R10], existing Deep Learning models for various summarization types in the legal domain were examined. It was concluded that text simplification often serves as a preliminary pre-processing step in many legal text-processing tasks to enhance sentence interpretability. The potential of transformers to decrease training time for extensive legal documents was noted. It was emphasized that adopting

TABLE 1. Related Work Research Questions

Ref.	Type	Primary Studies	Research Questions
[R8]	Survey	15	RQ1: What are the Deep Learning (DL) applications in law and the legal word embeddings trained on large corpora? RQ2: What is the current state of the art using DL in law in the NLP areas: Text Classification, Information Extraction, and Information Retrieval?
[R9]	Bibliometric Review	137	RQ1: What is the work done in Legal NLP in tasks like Classification, Feature Extraction, Information Extraction, Information Retrieval, Preprocessing, Summarization, Text generation, and Theoretical?
[R10]	Survey	12	RQ1: What are the state-of-the-art techniques in the Legal Text Summarization area?
[R11]	SLR	22	RQ1: What types of judicial decisions have been predicted using the machine learning methods? RQ2: What are the machine learning methods used to predict judicial decisions? RQ3: How was the performance of the machine learning method used to predict judicial decisions?
[R12]	Survey	60	RQ1: What are the state-of-the-art techniques in the Legal Information Retrieval area?
[R13]	SLR	59	RQ1: What are the main approaches for translating legal documents into formal specifications? RQ2: What legal ontologies have been used for the translation? RQ3: What annotation approaches are used for semantic annotation of legal text? RQ4: What are the main approaches for mining relationships from the annotated text? RQ5: What are the main techniques for formalizing NL terms into a domain model? RQ6: What kinds of techniques have been studied for translating NL expressions into formal ones for legal documents?

TABLE 2. Comparison among current studies

Ref.	Advantage	Disadvantages
[R8]	Emphasis on Deep Learning applications in law until 2018.	Do not cover recent years and is only focused on Deep Learning, not traditional Machine Learning approaches.
[R9]	RQ1: Covers a set of works on tasks like Multiclass Classification, Information Extraction, Information Retrieval, Summarization, Text Generation, and Theoretical from 1987 until 2020	The main focus of the article is on statistics and not state-of-the-art approaches. Also, it does not cover the period 2021-2022.
[R10]	Studies the Deep Learning's state of the art applied to the Legal NLP Summarization task.	Does not cover the year 2022 and is only focused on Summarization and Deep Learning approaches.
[R11]	Studies the Machine Learning approaches applied to the court decisions prediction.	Only focus on court decision prediction and missing the results of the year 2022.
[R12]	Studies the state-of-the-art techniques applied to the Legal NLP Information Retrieval task.	Only focused on Information Retrieval.
[R13]	The study focuses on state-of-the-art approaches in Legal NLP for translating legal documents into formal specifications with the use of legal ontologies	Focus only translating legal documents into formal specifications with the use of legal ontologies.

a hybrid (extractive and abstractive) summarization approach results in summaries more coherent than the original text.

Rosili et al. [R11], presented a systematic literature review (SLR) on predicting court decisions using machine learning methods was conducted. Machine learning methods utilized for predicting court decisions were determined and analyzed. It was concluded that including new case types and a combined classifier in machine learning methods was essential to enhance the performance of prediction tools.

In the research by Sansone et al. [R12], an overview of artificial intelligence approaches for the legal domain was presented, with an emphasis on Legal Information Retrieval systems utilizing NLP, Machine Learning (ML), and Knowledge Extraction (KE) techniques. Legal Information Retrieval systems were investigated from various perspectives, leading to a taxonomy of approaches. It was concluded that challenges remain in understanding legal document structures, their summarization, and their search and recommendation, especially considering the unique structures of document types like codes, case law, and articles.

Soavi et al. [R13], presented a SLR on transforming natural language legal contracts into formal specifications. The authors conclude with the need to focus on three main open challenges. First, the significant level of domain dependence of the approaches used. Second is the challenge of building tools that automatically identifying semantic and structural elements from the contract text with near-expert performance. Finally, it is essential to ensure that formal specifications fully capture legal documents intended to enhance adoption by legal practitioners and the quality of software systems that support the practice of Law.

In summary, our SMS can be separated from these literature reviews. First, our SMS explores 2015-2022, an interval of years that none of the related works cover individually. Furthermore, it focuses on all Legal NLP tasks and their resources to present a current state of the art in each of them and discuss limitations, future challenges, and trends. Second, it takes the form of an SMS.

III. METHODOLOGY

The Systematic Mapping Study presented in this paper follows the principled process as systematic literature reviews according to Kitchenham et al. [R7]. The process has three main phases: planning, conducting, and reporting. During the planning stage, we justify why we need to conduct this study, the protocol to follow is established, and the research

questions are defined. Next, the protocol is executed in the conducting phase, identifying the essential articles and categorizing and synthesizing the existing evidence. Finally, in the reporting stage, the results are reported in a structured format for the intended readers.

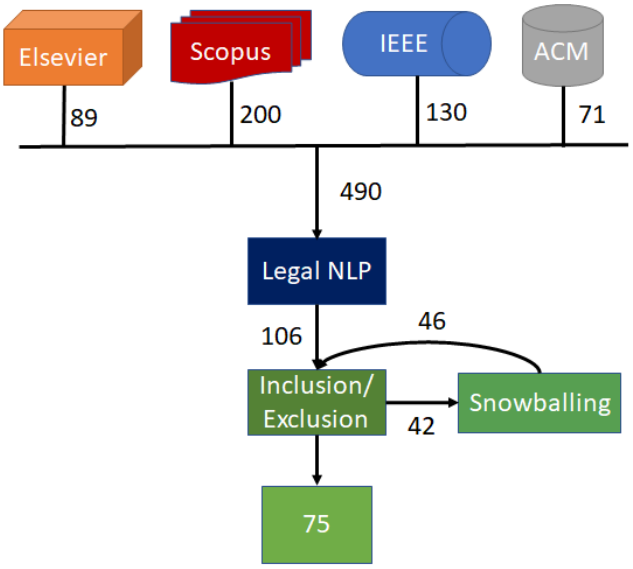

FIGURE 1. Search Process

A. RESEARCH QUESTIONS

As mentioned by Kitchenham et al. [R7], it is essential to identify the research questions before executing a study. Research questions allow us to clarify the scope and objectives of the study and, therefore, which papers should be included during the search process. In this study, our research questions are:

RQ 1: What is the current state of the art of every branch of NLP in the legal domain? With this research question, we want to determine every NLP branch's state of the art in the legal domain. We intend not only to assess the research community effort on this problem but also to detect if an NLP branch is applied to the legal field where the latest techniques are still not being used.

RQ 2: Which resources (datasets, ontologies, web scrapped, etc.) are being used from the legal domain to apply and enhance NLP? Under this question, we intend to extract the current resources used to train and improve NLP models applied to the legal domain. We want to categorize the purposes of each dataset, ontology, or any other resource extracted. We hope this will make it easier for future researchers to find a resource they can use easily instead of reinventing the wheel.

RQ 3: What are the limitations of the current work of NLP applied to the legal domain? With this question, we intend to identify areas of improvement or unexplored areas in the Legal NLP field. This analysis will provide future researchers with a roadmap and perspective on what can be done, resources that need to be built, and the branches of improvement. Furthermore, we will also summarize future work and challenge perspectives from the papers analyzed.

B. INCLUSION AND EXCLUSION CRITERIA

Inclusion and Exclusion criteria are used as the basis to select the essential studies according to the process of SMS and SLR [R7]. Our inclusion criteria in this paper are:

- Papers investigating the application of NLP techniques on the legal domain in general.
- Papers investigating the performance of NLP algorithms applied to any subdomain of the legal domain.
- Papers investigating the results of NLP algorithms applied to a legal domain dataset.
- Papers investigating the results of NLP algorithm using legal hand-made or automatically extracted ontologies.
- Papers investigating the limitation of some NLP areas on the legal domain or NLP in general.

On the other hand, our exclusion criteria to exclude studies irrelevant to our goals were:

- Papers not in English.
- Short papers (less than four pages), opinions papers, vision or roadmap or plan papers.
- Secondary studies, like existing SMS or SLR.
- Duplicate papers. In case of duplication, the most recent version was selected.
- Case Studies without generalization.
- Non peer-reviewed.
- No full text available.

C. SEARCH PROCESS: IDENTIFYING ESSENTIAL ARTICLES

In this work, we include Natural Language Processing (NLP) primary studies (not use cases or surveys) on the topic of NLP in the legal domain, published from 2015 to 2022. To identify these papers, we did the following. First, we searched three scientific databases: Scopus, ACM Digital Library, and IEEE Xplore. The search was applied to the title and abstract of papers. Figure 1 illustrates the process. Our search query can be described as follows:

```
((Legal OR Law OR Policies OR Regulations)
AND (NLP OR "Natural Language Processing"))
```

Based on the objectives of this SMS and the RQs defined, the final string was structured using the PICOC strategy [R14]. The **P**opulation regards Legal NLP research papers; because our goal is the Legal NLP domain in general, we need it to target as many sources as possible and not only NLP context. The **I**ntervention is Legal NLP (study's focus). The **C**omparison is done across methods, datasets, and state-of-the-art approaches in the Legal NLP field. The **O**utcome is a categorized summary of Legal NLP approaches and results. In addition, an overview of datasets, ontologies, and other resources. And the outcome is an analysis, categorization, and solutions to current challenges in the Legal NLP field.

Finally, the **C**ontext is broad due to the generality of our research. Therefore, any research done in Legal NLP is considered. Finally, the search string was constructed using the usual keywords we detected in surveys and other SLRs in Legal NLP.

The number of papers detected in the first stage was 502. We decided on the time scope for the period from 2015 to 2022 since 2015 is the year with the explosion in the use of Deep Learning and a massive advance in the state of the art of several NLP tasks. To mitigate the risks associated with paper duplication, we used bibtex duplication check of all the papers extracted by the search query and removed the duplicates, which led to a total of 490 unique articles Table 3 shows the distribution of papers and results per indexer.

Phase 1: To include and exclude papers, our first phase was that each article was examined by a pair of authors and applied inclusion and exclusion criteria to the title and abstract. During this process, 106 articles were included.

Phase 2: In case of conflicts between the pair of judges in Phase 1, a third author also reads the title and abstract of such paper and gives the final decision. From this phase, only three papers were excluded, giving a total of 103 papers.

Phase 3: In the third phase, every single author was assigned a subset of the included papers to read the full text and decide whether the article should be included in the study. After this phase, we ended with 42 papers relevant to our SMS.

Phase 4: In the fourth phase, every author who performed the full read of a paper did the data extraction following the established coding.

Phase 5: To reduce the risk of omitting relevant articles, we also performed a one-level lightweight forward and backward snowballing on the included papers [R15]. We inspected the articles cited by each of our included primary studies and the publications that subsequently cited the paper. Citations were located using Google Scholar. In total, 46 more papers were identified, and after applying phases 1 to 4 to them, it included 33 new relevant articles. This raised the number of primary studies in our SMS to 75 papers.

TABLE 3. Search Query Results for Various Index Sites

Indexer	Search Results	Filtered	Referenced	Relevant
ACM DL	71	17	6	6
IEEE Xplore	130	41	15	15
Scopus	200	34	13	13
Elsevier	89	14	8	8
Sub-Total	490	106	42	42
Snowballed	46	43	33	33
Total	536	149	75	75

D. QUALITY ASSESSMENT

We used the DARE [1] method criteria [R16]. This method appraises the quality of SLRs in Software Engineering (SE). However, the criteria used to score a study in DARE are not necessarily SE domain-dependent. Therefore, it is a good fit to use this criterion here.

Our SMS quality assessment was done through two steps, following good practices like in [R16]. First, the researchers responsible for the data extraction performed the quality assessment using DARE. After that, a second researcher checked this evaluation. In the presence of any conflicts, they discussed until agreement. This process was done during each paper's full reading and data extraction.

We did not use DARE as a criterion for inclusion or exclusion because we aim to analyze the results and evidence provided in every paper. We only use it to assess the quality of this SMS.

E. DATA EXTRACTION AND ANALYSIS PROCEDURES

We define a data extraction form to capture relevant Legal NLP information from the data sources, like datasets, other resources used, neural architectures, machine learning methods, word embeddings, language models, and relevant information for answering RQs as can be appreciated in Table 4. The 148 remaining papers were distributed and assigned to each researcher during the second phase. While reading the complete text, the researcher would also perform the data extraction on such an article if the decision were to include it.

TABLE 4. Data Extraction Form. Shows Extraction Field, Purpose of RQ involved, and the total of papers used for each Research Question.

Extraction Field [Field Description]	Purpose/RQs	Papers
Reference info. [Title, authors, year, source, abstract]	Demographic	-
Paper Goal/Method/Results	RQ.1	65
NLP Algorithm/Word Embedding/Language Models	RQ.1	65
Resources used/dataset/ontologies	RQ.2	14
Challenges and limitations in Research	RQ.3	13

This extraction was done following a coding schema as the best practices for SMS suggest [R7]. We identify the codes of our coding schema as follows:

General Information: This includes the goal and method followed from the paper. Which Legal NLP task were they tackling? Example: " To predict the judgment outcomes as a multiclass classification problem based on a court case filed in China."

NLP Approach: This includes the algorithm used if it was a neural network and which architecture. In addition, we include the word embeddings and language models used, if any. Finally, any dataset, ontology, or external resource is used. Example: "The Multiclass Classification model is a Neural Network Architecture based on

[1] https://www.ncbi.nlm.nih.gov/pubmedhealth/about/DARE/ Accessed on 02/04/2023

Recurrent Neural Networks. Authors use Chinese court cases, which they call articles, which describe facts of the case, along with outcomes."

Challenges: This relates to the difficulties and problems that can be found by dealing with legal text, lack of resources, and even challenges of the approach taken. Example: ''Ethical concerns associated with high-performance Legal NLP algorithms. When integrated into the legal system, such technology must remain free from issues like bias, racial discrimination, and uninterpretable results that fail to persuade individuals."

Future Directions: This describes the future directions that the field might take in the following years. Example: ''Inclusion of bias and fairness analysis and good interpretability."

F. DATA SYNTHESIS AND BIAS ASSESSMENT

The descriptive statistics presented in all our Research Questions (RQ1, RQ2, and RQ3) were analyzed by counting the number of selected primary studies published that fit in a given classification by NLP task, method, language model, word embeddings used, resources used, and limitations presented. Furthermore, for the qualitative analysis presented in RQ1 and RQ3, the NLP experts discussed and used as base similar studies [R9, P1] to decide the proper naming for each category and subcategory by selecting the most frequently used.

A similar approach was taken for the RQ1 classification of primary studies based on the task and method. Data extraction accuracy and category fit for each paper were cross-validated by an NLP expert. For papers that didn't align with defined NLP task categories, discussions led to consensus on classification, typically associating with known NLP tasks. The exact process was applied for subcategories, methods, word embeddings, and language models. An increased discussion was required when the method or neural architecture clarity was lacking.

In the case of RQ3, with the current gaps and challenges, we followed the same protocol as in RQ1. We considered information extracted from the Limitations or Threats sections in particular primary studies. This way ensured that provided information was not biased by the authors of this study.

IV. RESULTS

This section presents our results by providing the answers to each research question separately. Of 536 papers obtained from the search, only a few papers were relevant to our research. Most of the papers presented the use of one particular model in one specific use case. They cannot extract anything related to the current state of the art or limitations in general of the Legal NLP field. Table 5 shows the summarized results of Research Question 1.

A. RQ 1: WHAT IS THE CURRENT STATE OF THE ART OF EVERY BRANCH OF NLP IN THE LEGAL DOMAIN?

To make the discussion easier to read and understand, we will separate the answer to this research question into multiple categories. One category for each Legal NLP area is explored in our selected papers. Furthermore, Table 5 outlines which papers addressed specific Legal NLP tasks, the methods employed in these papers, the types of legal documents utilized, and the total number of papers per task. The categories that will be discussed are:

Language Modeling (LM): predicts upcoming words from prior word context.

Multiclass Classification (Mult. Class.): in Machine Learning consists of classifying data instances into two or more selected classes.

Summarization (Sum.): in NLP is the task of producing a shorter version of one or several documents that preserves most of the input's semantics.

Information Extraction (IE): is the NLP task of extracting limited semantic content from text.

Question Answering and Information Retrieval (QA/IR): Question Answering is the NLP task where a given question is answered by using a set of documents as a knowledge base. Information Retrieval under NLP encompasses the retrieval of all media based on the user needs related to a topic.

Coreference Resolution (CR): is the task in NLP of finding all the expressions that refer to the same entity in a text.

Cross Lingual Transfer Learning (Cross-Lingual): Cross-lingual transfer refers to transfer learning using data and models available for one language with higher resources (e.g., English) to solve tasks in another, commonly more low-resource, language.

All the categories selected are well-known NLP tasks. We decided to follow this categorization based on similar Legal NLP studies [R9] and empirical studies [P1].

1) Language Modeling

Language modeling task consists of predicting upcoming words from prior word context. Formally, given A sequence of words $w_1, w_2, ..., w_n$ drawn from a vocabulary V, where n is the length of the sequence, and V is the set of all possible words. Then, the objective is to estimate the joint probability distribution $P(w_1, w_2, ..., w_n)$ of the sequence. This joint probability can be decomposed using the chain rule of probability:

$$P(w_1, w_2, ..., w_n) = \prod_{i=1}^{n} P(w_i | w_1, w_2, ..., w_{i-1}) \quad (1)$$

Where $P(w_i | w_1, w_2, ..., w_{i-1}$ is the conditional probability of word w_i given the previous $i-1$ words in the sequence.

In recent years the best results have been obtained by Neural Language Models [P1] where embeddings of the previous words represent the preceding context. Among the most

TABLE 5. Legal type of documents with the methods, embeddings, and PLMs used in summarization

Task	Methods	Document Types	Datasets	Articles	Total
LM	**LMs obtained:** Legal-BERT, CoLMQA, Lawformer, Legal-RoBERTa	Judgement Outcomes, General Legal Text	CAIL2018, CAIL-Long, Case-HOLD, LexGLUE	[P2, P3, P4, P5, P6]	5
Mult. Class.	**Classic ML:** LDA, Naive Bayes, Decision Trees, SVM. **Deep Learning (Not LLM-based):** Stack Attention, LSTM, GRU, BiLSTM, Reinforcement Learning, GNN, CNN, BiGRU-Att, HAN, LWAN, ZERO-CNN-LWAN, ZERO-BIGRU-LWAN. **LLM-based:** BERT, HIER-BERT, XLNet, T5, DistilBERT, RoBERTa, DeBERTa, BigBird, Longformer, CaseLaw-BERT	Privacy Policies, Judgement Outcomes, General Legal Text, Assigning Petitions	SCOTUS, CJO, PKU, CAIL, OPP-155, ECHR, SigmaLawABSA, Terms of Service, DMOZ, POSTURE50K, LExGLUE, ILSI.	[P1, P7, P8, P9, P10, P11, P12, P13, P14, P15, P16, P17, P18, P19, P20, P21, P22, P23, P24]	20
Summ.	**Classic ML:** Naive Bayes, Random Forests, SVM, TBS, Text Rank. **Deep Learning (Not LLM-based):** InferSent + FFNN, Sent2Vec + FFNN, LSTM, BiLSTM + Attention, Pointer Generator Networks. **LLM-based:** BERT, T5, BART, Custom LegalBERT, Dyploc, Global Aware.	Patent Documents, General Legal Text	Legal cases from the Federal Court of Australia, BillSUM, LegalSUM, Civil Trial Court Debate	[R10, P1, P25, P26, P27, P28, P29, P30, P31]	9
IE	**Classic ML:** Rule-Baed, LDA, SVM. **Deep Learning (Not LLM-based):** BiLSTM + Attention + CRF. **LLM-based:** BERT.	Privacy Policies Court Records, General Legal Text, License Terms	Case-HOLD, COLIEE Statute Law Task, License texts, Contracts, LexGLUE	[P32, P33, P34, P35, P36, P37, P38, P39, P40, P41]	10
QA and IR	**Classic ML:** BM25, SVM. **Deep Learning (Not LLM-based):** BiDAF, Word2Vec-based, Neural Attention, CNN. **LLM-based:** BERT, RoBERTa, LegalBERT, ALBERT, ELECTRA.	Civil Code, General Legal Text, Tax Ruling	Query Generation, EURLEX57K, ALQAC-2021, CJRC, JEC-QA, LeCARD, Bar Exam QA, CJO, LexGLUE	[P1, P42, P43, P44, P45, P46, P47, P48, P49, P50, P51, P52, P53, P54, P55, P56, P57]	17
CoRef-Res.	**Classic ML:** Rule-Based. **Deep Learning (Not LLM-based):** GNN, BiLSTM. **LLM-based:** SpanBERT, BERT.	Tax Law, General Legal Text	CRC, Luxembourg's Income Tax Law, License texts.	[P58, P59, P60]	3
Cross-Lingual	**LLM-based:** BERT, DistillBERT	General Legal Text	JRC-Acquis, EURLEX57K, License Texts	[P61]	1

widely known, big, and high-performance language models are transformer-based language models like Bidirectional Encoder Representations from Transformers (BERT) [R17], Generalized Autoregressive Pretraining Model based on the Transformer-XL (XLNet) [R18], and Generative Pre-trained Transformer (GPT) [R19] based.

Regarding approaching the language modeling task in the legal language, we cite the work of Chalkidis et al. [P2], which conducted a study to understand how to maximize BERT's performance in the legal domain. They released a new language model called Legal Bidirectional Encoder Representations from Transformers (Legal-BERT) by fine-tuning the original BERT model on legal data. For this they used as based data: the official database of the European Union Law (EURLEX) [2]; official place of publication for newly enacted legislation in the United Kingdom (LEGISLATION.GOV.UK) [3], the case-law database of the European Court of Human Rights (HUDOC) [4], CASE LAW ACCESS PROJECT [5]; the Electronic Data Gathering, Analysis, and Retrieval system used by the U.S. Securities and Exchange Commission (SEC-EDGAR) [6].

In the research by Huang et al.[P3], a language model for legal documents was developed based on GPT [R19]. Due to uncertainties, such as article numbers, slots were used instead of uncertain tokens. Legal documents, inclusive of these slots, were then used for training. A second phase, involving Key-Value Memory Networks enhanced by Transformer encoders, filled these slots, functioning as a question-answering task. The model's performance was evaluated using F-score and perplexity, revealing that the slotted model outperformed state-of-the-art models. Moreover, their slot-filling algorithm exhibited higher accuracy than Memory Networks (MemNN) [R20] and Whoosh [7].

[2]http://eur-lex.europa.eu Accessed on 02/04/2023
[3]http://www.legislation.gov.uk Accessed on 02/04/2023
[4]http://hudoc.echr.coe.int Accessed on 02/04/2023
[5]https://case.law Accessed on 02/04/2023
[6]https://www.sec.gov/edgar.shtml Accessed on 02/04/2023
[7]https://pypi.org/project/Whoosh/ Accessed on 02/04/2023

In the study by Zheng et al. [P4], a new dataset, the Case Holdings On Legal Decisions (CaseHOLD) dataset, was introduced, representing a crucial task for lawyers with legal significance and NLP challenges. Performance evaluation was conducted on CaseHOLD and other legal NLP datasets. Results indicated that domain pretraining might be beneficial when task alignment with the pretraining corpus is strong. The performance enhancement in three legal tasks was linked to task domain specificity. A comparison revealed similar performance between SVM and BERT for an overruling task, with more distinct differences in complex challenges.

In the research by Xiao et al. [P5], a pre-trained model, Lawformer, was introduced to analyze lengthy legal documents. Based on the Transformer designed to handle long sequences of data (Longformer) [R21], it blends sliding window attention, dilated sliding window attention, and global attention mechanisms instead of the full self-attention method, enabling linear complexity for extended sequences. The China AI and Law Challenge Long (CAIL-Long) dataset was proposed, resembling CAIL2018 but with average case lengths mimicking real-world scenarios. Additional datasets utilized included the Legal Case Retrieval Dataset for Chinese Law System (LeCaRD) [P49], Chinese Judicial Reading Comprehension (CJRC) [P46], and a Legal Question Answering dataset collected from the National Judicial Examination of China (JEC-QA) [P62].

In the study by Qin et al. [P6], the performance of four pre-trained models on general and legal domain corpora was compared regarding classification accuracy on three Chinese legal document datasets. The primary datasets evaluated included CAIL2018, a legal judgment prediction dataset from the Chinese AI and Law challenge; CAIL-Long, also from CAIL, featuring lengthier civil cases comprising shorter texts, most of which are under 256 tokens. The leading language model showcasing the best results was:

Lawformer: which utilizes Longformer as a basic encoder and collects tens of millions of case documents published by the Chinese government for pretraining.

Legal-RoBERTa: same dataset as Lawformer, but with the Robustly optimized BERT approach (RoBERTa) architecture. The main limitations mentioned are long text, quadratic complexity of self-attention, and position embeddings not generalizable (like BERT, max 512).

The authors also mentioned as a limitation, that the classification effectiveness of the Pretrained Language Model (PLM) decreases when the semantic composition and complexity of the documents increase. They propose solutions to these problems using variants of transformers with approximations of attention with lower complexity.

2) Multiclass Classification

The multiclass classification task in Machine Learning consists of classifying data instances into two or more selected classes. Formally, given an input sequence of words $w_1, w_2, ..., w_n$ from the vocabulary V and a set of classes $C = c_1, c_2, ..., c_k$ where k is the number of classes. Then, it is assigned the input sequence $w_1, w_2, ..., w_n$ to a given class in C based on the estimated probabilities $P(c_i|w_1, w_2, ..., w_n)$ for $i = 1, 2, ..., k$. Usually, the optimization problem to solve is:

$$c^* = arg\ max_{c_i \in C} P(c_i|w_1, w_2, ..., w_n) \quad (2)$$

Where c^* is the predicted class for the given input sequence.

Several classifications you will want to automate to legal text exist, as in many NLP domains. As part of the high research on this area of the Legal NLP field, we cite the following works.

In the research by Luo et al. [P21], judgment outcomes of Chinese court cases were predicted as a multiclass classification problem. Data was sourced from the court data and judgments available online from China Judgement Online (CJO)[8]. A two-stack attention mechanism was employed: one for fact embedding and another for dynamically generated article embedding, using fact-side clues for guidance. Word embeddings were created with "Words to Vectors" (Word2Vec) [R22] on various legal sources. A project limitation noted was its confinement to single-defendant cases, as multiple defendants complicated fact-to-defendant mapping.

In the study by Shulayeva et al. [P63], automatic identification of legal principles and facts within common law citation was addressed. The Naive Bayesian Multinomial Classifier was used to classify features like part of speech tags, unigrams, dependency pairs, sentence length, text position, and citation presence. A corpus derived from 50 British and Irish Legal Institute common law reports was introduced. This corpus features annotated areas with predefined citation names, with sentences labeled as fact, principle, or neither. The authors indicated that the corpus can be accessed upon request.

Following a similar approach, legal text classification techniques were investigated by Undavia et al. [P7] on a dataset containing manually-categorized SCOTUS legal opinions (Supreme Court Database or SCDB) corpus, from Washington University School of Law [9]. The approaches tested are:

- Latent Diritshclech Allocation (LDA) + Logistic Regression.
- ‘‘Documment to Vector" Doc2vec + Logistic Regression.
- Bag-of-Words + Support Vector Machine.
- Word2vec + Convolutional Neural Networks (CNN)

[8]https://wenshu.court.gov.cn Accessed on 02/04/2023
[9]http://supremecourtdatabase.org/ Accessed on 02/04/2023

- Word2vec + LSTM
- Word2vec + GRU

Also, in addition to the Word2Vec embeddings [R22], they explored other pre-trained word embeddings like FastText vectors [10] from Facebook AI Research and also Glove vectors [11] from Pennington et al. Finally, they also explored the publicly available pre-trained word vectors trained on about 100 billion words from part of the Google News dataset [12]

From the comparisons, the best method from this work for automated legal document classification is the combination of CNN with the word embeddings from a general domain (Google News) with (72.4% accuracy for 15 general categories and 31.9% accuracy for the 279 more specific categories).

In the research by Zhong et al. [P18], TOPJUDGE, a model for predicting judgments from Chinese legal documents, was introduced. A unique multitasking methodology using DAG-based architectures was proposed for legal judgment prediction. Compared to baselines, which encompassed CNNs and LSTMs with a softmax activation function, a multitasking strategy for training the neural model was developed. New datasets were introduced and utilized by the authors: China Judgement Online (CJO) [13], by Peking University Law Online (PKU) [14], and (Chinese AI and Law Challenge) CAIL [15].

In alignment with previous approaches, a study was conducted by Harkous et al.[P20] to determine if a privacy policy addresses users' general privacy concerns. The 115 Online Privacy Policies (OPP-115) dataset[P19] was employed. A neural architecture comprising neural embeddings, a CNN, and dense layers with a classification head was utilized. Custom word embeddings for the privacy-policy domain, named "Policies Embeddings," were trained on a corpus of 130K privacy policies from Google Play Store apps, reflecting app companies' data practices. Bag-of-word techniques were also applied for representing judicial documents and extracting features for subsequent learning.

In the study by Fang et al.[P24], manifold learning-based dimensionality reduction methods were assessed for judicial document classification. Their dataset, sourced from Aletras et al.[R23], incorporated features from the European Court of Human Rights (ECHR) case texts using N-gram and topic models [16].

Several dimensionality reduction techniques were evaluated, including autoencoder, factor analysis, Gaussian processes latent variable model (GPLVM), Isomap, principal component analysis (PCA), kernel version of PCA, probabilistic version of PCA, Landmark Isomap, locally-linear embedding (LLE), multidimensional scaling (MDS), Sammon mapping, stochastic neighbor embedding (SNE), symmetric SNE, and t-distributed stochastic neighbor embedding (t-SNE). Methods such as Bagging, K Nearest Neighbors (KNN), Logistic Regression, Random Forest, and Support Vector Machine (SVM) were tested for classification.

A limitation highlighted by the authors was the often lower count of labeled judicial documents compared to their feature dimensionality, potentially compromising prediction performance when using directly extracted text features. It was noted that the Bag of Words model can yield numerous features, possibly affecting the Natural Language Processing (NLP) algorithm's performance. While linear dimension reduction techniques are expected to offer improved vectors with fewer dimensions, non-linear dimension reduction techniques were suggested as a remedy, aiming to preserve distances between points in reduced dimensions.

In the study by Chalkidis et al.[P9], various neural models were assessed on the introduced English legal judgment prediction dataset, sourced from the European Court of Human Rights (ECHR)[17]. Apart from dataset introduction and achieving notable multiclassification results, the superiority of Pre-trained Language Models (PLMs) over simpler methods was demonstrated. Evaluated neural models included Bidirection Gated Recurrent Unit with Attention (BiGRU-Att) [R24], Hierarchical Attention Network (HAN)[R25], Label-Wise Attention Network (LWAN)[R26], and BERT. A hierarchical version of BERT (HIER-BERT) was introduced to overcome BERT's length constraints. Notably, HAN and HIER-BERT exhibited the most prominent performance.

In another study by Chalkidis et al.[P10], Large-Scale Multi-label Text Classification (LMTC) in the legal field was addressed. A dataset, consisting of 57K legislative documents from EUR-LEX, was introduced[18]. Methods such as BERT, BIGRU-ATT, HAN, CNN-LWAN, BIGRU-LWAN, and their Zero-Shot versions like ZERO-CNN-LWAN, ZERO-BIGRU-LWAN were evaluated. It was found that Pre-trained Language Models (PLMs) outperformed several other techniques, with BERT achieving the highest results across most metrics.

In the study by Pillai et al.[P11], verdict classification of court cases was addressed as a text classification challenge. A dataset comprising Indian civil and law judicial cases was introduced[19]. Words were represented using Bag of Words [R22], and a CNN network served as the text encoder. Limitations noted included the absence of standard legal procedures across countries, limited cross-country data, and the presence of irrelevant content in legal texts. A future challenge identified was the reduced accuracy when considering multiple verdicts.

[10] https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md Accessed on 02/04/2023
[11] https://nlp.stanford.edu/projects/glove/ Accessed on 02/04/2023
[12] https://code.google.com/archive/p/word2vec/ Accessed on 02/04/2023
[13] http://wenshu.court.gov.cn/ Accessed on 02/04/2023
[14] http://www.pkulaw.com/ Accessed on 02/04/2023
[15] http://cail.cipsc.org.cn/index.html Accessed on 02/04/2023
[16] https://figshare.com/s/6f7d9e7c375ff0822564 Accessed on 02/04/2023
[17] https://archive.org/details/ECHR-ACL2019 Accessed on 02/04/2023
[18] http://nlp.cs.aueb.gr/software_and_datasets/EURLEX57K Accessed on 02/04/2023
[19] https://lawrato.com/indian-kanoon/ipc and https://devgan.in/ipc/ Accessed on 02/04/2023

In the research conducted by Noguti et al. [P16], the automation of assigning petitions to their relevant law areas was explored. The dataset, sourced from the ‘‘Public ‘‘Prosecutor's Office of the Ministério Público" (PRO-MP) system covering public petition registrations from 2016 to 2019, was meticulously labeled by the ‘‘Ministério Público do Estado do Paraná" (MPPR) prosecutors. Standard text preprocessing, including lowercasing, lemmatization, and punctuation removal, was employed. Texts were represented using Term Frequency-Inverse Document Frequency (TF-IDF) or word embeddings (Word2Vec, FastText, Glove).

In addition, the authors tested classification models like Logistic Regression, SVM, Gradient Boosting, and various neural networks were evaluated. Recurrent neural networks, particularly LSTM, achieved up to 90% accuracy, surpassing human performance. However, a clear reference for the dataset was not provided.

In the study by Jayasinghe et al.[P22], the application of sentence embeddings for multiclass classification was investigated to pinpoint critical sentences in legal cases. An adapted dataset based on the SigmaLaw ABSA Dataset from[P64] was utilized. The BERT-cased model [20] was fine-tuned, and an average of the hidden states was taken, followed by prediction using a fully connected layer. A unique loss function was devised due to the specific nature of their classes.

In the research by Hamdani et al. [P8], an effort was made to bridge aspects of compliance checking through a dual-focused study. Firstly, a framework for a document-centric approach to compliance checking in the data supply chain was conceptualized. Secondly, methods for automated compliance checking of privacy policies were devised. A Hierarchical Multi-Label Classification problem using privacy policies was introduced. Two classical approaches, local classifiers and text-to-text, were experimented with.

The local classifier approach achieved state-of-the-art by transformers, reproducing Polisis paper previous arch [P20], using XLNet [R18] instead of CNN as a base classifier. Finetune XLNet on 21 tasks, one predicting categories and the rest for each attribute's values.

In the text-to-text approach proposed, the Hierarchical Multi-Label Classification (HMTC) problem was transformed into text-to-text tasks for each label hierarchy level, effectively capturing label dependencies within the same level. It was noted that training a unique algorithm per level ensures linear classifier scaling with hierarchy depth. Two finetuning methods were explored: independent task finetuning and multitask finetuning to grasp the global label hierarchy. The T5 [R27] big LLM transformer classifier was utilized. The OPP-115 [P19] dataset was employed, and a ground truth dataset, encompassing policies from both OPP-115 and other GDPR sources, was introduced.

In the study by Akcca et al. [P23], the prediction of crime labels in Turkish court decisions was explored. The authors developed both supervised and unsupervised datasets. Models ranging from traditional machine learning to transformers were tested. Hyperparameters were explored through grid search. Comparisons were made among models like Naïve Bayes, Logistic Regression, SVM, Bidirectional Long Short-Term Memory (BiLSTM), Distilled BERT (DistilBERT), and BERT. Word embeddings were investigated, including Bag of Words + TF-IDF and Fast Text. Two datasets, an unlabeled collection for transformer pretraining and a labeled set of court cases, both derived from Turkish legal documents, were used.

A novel legal extreme multi-label classification dataset, POSTURE50K [21], containing 50,000 legal opinions and associated legal procedural postures, was introduced by Song et al. [P12]. A deep learning architecture, utilizing domain-specific pre-training and a label attention mechanism, was proposed for multi-label document classification. Evaluations were conducted on both the released dataset and the Large-Scale Multi-Label Text Classification on English Union Legislation dataset (EUROLEX57K), with state-of-the-art results observed. The methodology employed was based on a RoBERTa-driven deep learning architecture, using label embeddings and multi-task learning strategies.

The absence of a standard benchmark dataset and the high complexity of the legal domain were cited as limitations by the authors. It was suggested that future work should expand the classification and text representation experiments, fine-tuning, and pretraining of neural language models to capture the domain-specific characteristics of law. Enhancing the dataset using additional sources and data augmentation was also proposed.

In the study by De et al. [P14], a hybrid system for multi-label classification of judgments was proposed, incorporating visual and natural language descriptions for explanation in Spanish legal documents. Text processing was the initial step, involving document cleaning, lemmatization, and redundancy removal. Parts were identified using regular expressions, and classifications utilized knowledge of Spain's legal documents. Anonymization followed, with proper names replaced by tags like @Corporate, @Judge, etc. Two classification approaches, Binary transformation strategy and Multi-class transformation strategy, were evaluated.

Different models, such as Decision Trees, Random Forest, and Extra Tree Classifier, were tested by the authors. The Random Forest model with the MTS strategy was found to have superior performance, though the recall was higher with the BTS strategy. Given the decision tree model's use, an algorithm was proposed to traverse the tree's branches from roots to leaves. The final decision was derived from the majority of labels determined by the decision trees.

A custom dataset containing 106,806 judgment texts, annotated by lawyers familiar with the Spanish legal system, was utilized by the authors. No name or link for this dataset

[20] https://huggingface.co/bert-base-cased Accessed on 02/04/2023

[21] https://forms.office.com/Pages/ResponsePage.aspx?id=ZLjMYhpqXUuOHDl97BqCWEQaso-9T_JFiLjD7N8NqbNUMjEyQ0JRTDhGQlM4VVUzS0Q2TFRRWEFCMy4u. Accessed on 02/04/2023

was provided. Each document was marked with up to three labels, including a substantive order and three law categories spanning 47 classes.

In the study by Lyu et al. [P65], attention was given to distinguishing similar law articles and confusing fact descriptions in the legal judgment prediction (LJP) task. The unique challenge of concurrently addressing both issues was undertaken for the first time. A novel reinforced Criminal Element Extraction Network (CEEN) was introduced, comprising: (i) a fact description encoder, (ii) an RL-based element extractor, (iii) a criminal element discriminator, and (iv) a multitask judgment predictor.

Sentences of fact descriptions were projected into latent spaces by the fact description encoder using the hierarchical BiLSTM [R25]. Distinctive criminal elements, including criminals and targets, were uncovered using the reinforced criminal extractor. Meanwhile, an element discriminator was designed to distinguish law articles with similar TF–IDF representations. The effectiveness of the proposed method for Legal Judgement Prediction (LJP) was verified through extensive experiments on benchmark datasets CAIL-small and CAIL-big.

The use of the statute citation network with textual descriptions for Legal Statute Identification was first introduced by Paul et al.[P17]. (Legal Statute Identification using Citation Network (LeSICiN) was proposed, employing an Attribute Encoder for both Facts and Sections based on a Hierarchical Attention Network (HAN) network[R25]. A Structural Encoder, which utilized meta paths from the citation network and functioned as a Graph Neural Network (GNN), was also incorporated. A large-scale Legal Statute Identification (LSI) dataset derived from Indian court case documents was constructed. The goal was to identify sections of the Indian Penal Code, a primary criminal law in India. Both the dataset and codes were made available. [22].

The empirical study from Song et al. [P1] showed that in binary classification and multilabel classification tasks, the best PLMs are used in four different datasets. Table 6 offers the models that get the best results in each dataset.

3) Summarization

Summarization in NLP is the task of producing a shorter version of one or several documents that preserves most of the input's semantics. Formally, given an input document D which consists of sentences $s_1, s_2, ..., s_m$ where each sentence s_i is a sequence of words $w_{i1}, w_{i2}, ..., w_{in}$ from the vocabulary V. The objective is to produce a concise summary S that retains the essential information from D.

Approaching this task in the legal domain, we cite the following works:

In the study by Polsley et al.[P25], TF-IDF and Part of Speech Tagging (POS-tag) were employed to determine weights in various sections of a legal document. Summaries were subsequently generated based on these weights. The dataset sourced from the Federal Court of Australia[23] was utilized.

In the research by Merchant et al.[P26], legal text summarization was investigated through latent semantic analysis. Singular Value Decomposition (SVD) was applied to identify essential sentences from singular vectors, selecting them based on importance. The Bag of word embeddings[R30] was utilized as a vector representation of sentences. Criminal judgments were used for multiple-document tasks, while civil judgments were applied for single-document tasks. A shift from the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [R31] evaluation method was considered by the authors, and the potential application on mobile phones was planned.

Duan et al.[P28] introduced a method for quantifying court debates through multi-view utterance representation. An end-to-end model was developed, multitasking the learning process to tackle multi-role and multi-focus court debate summarization. By leveraging a legal knowledge graph, the model was designed to uncover legal concepts and align controversial focuses with the debate. Central to the model's architecture are BiLSTM-Attention mechanisms, emphasizing sentence representation, role representation, and legal knowledge representation. For evaluation purposes, a civil trial court debate dataset was constructed and utilized by the authors[24].

Tran et al. [P29] focused on the legal case retrieval task from the Competition on Legal Information Extraction/Entailment 2019 (COLIEE 2019). A combination of lexical features and latent features, termed decided summarization, was introduced. Different views were utilized to compare a query case with its candidates. While the summary and paragraph were employed to represent each query, the candidate was characterized by its summary, the lead sentence per paragraph, and the subsequent paragraphs.

In COLIEE 2019, many candidates lacked summaries from encoded summarization. Six matching options were utilized to compare the query and candidates. Various text matching techniques such as N-gram, skip-bigram, and a combination of unigram + skip-gram were employed. The issue was approached as a ranking challenge, with linear-SVM addressing the optimization, producing the top k results. The significance of catchphrases, key legal points typically drafted by experts, was emphasized as they correlated with summaries. A phrase scoring model was proposed to pinpoint these phrases, designating them as summaries.

Tran et al. [P30] advanced their earlier work by crafting a system for legal case summarization to aid in legal information retrieval. They sourced their primary datasets from COLIEE 2018 and 2019. Their phrase scoring model integrated Word Embedding (GloVe), CNN, and Multi-layer

[22]https://github.com/Law-AI/LeSICiN Accessed on 02/04/2023

[23]https://archive.ics.uci.edu/ml/datasets/Legal+Case+Reports Accessed on 02/04/2023

[24]https://github.com/zhouxinhit/Legal_Dialogue_Summarization Accessed On 02/04/2023

TABLE 6. Results from empirical study Song et al. [P1]

Dataset	PLM	F1	m-F1	M-F1	W-F1
Overruling [P4]	Custom LegalBERT [P4]	0.973			
Terms of Service [P15]	Custom LegalBERT	0.812			
POSTURE50K	BigBird [R28]				0.809
POSTURE50K	LightXML [R29] + Custom LegalBERT		0.820		
POSTURE50K	LAMT_MLC [P12]			0.263	
EUROLEX57K	LightXML + Custom LegalBERT		0.727		0.700
EUROLEX57K	LAMT_MLC			0.326	

Perceptron. They employed a four-step process to generate text summaries: firstly, they ranked document phrases based on their scores. Subsequently, high-scoring phrases were chosen, overlapping phrases were merged, and the process halted when a set summary length was surpassed.

Trappey et al. [P31] targeted summarization of patent documents. They amassed Chinese and English patents, segregating them into training and test datasets based on topic domains using Doc2Vec. The documents underwent preprocessing, which included lowercasing alphabets, removing stop words, and splitting into tokens. An attention-based connection, seen in machine translation, was established between the bi-directional LSTM encoder and the LSTM decoder. The model highlighted words with peak attention scores and crafted a summary sentence. The model's performance was gauged using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) against a reference summary, drawing data from the Derwent Innovation (DI) platform [25], and its target is collected from quick views in [26].

In this research, Norkute et al. [P66] highlighted the advantages of explainable automatic summarization in the legal domain, emphasizing expedited user reviews. Two distinct methods for explainable AI were introduced. The attention vector approach leverages attention scores from a deep learning summarizer to pinpoint influential tokens. Conversely, the source attribution approach employs a heuristic independent of the Pointer Generator model to determine the sentences most impacting the summary. The study revealed that users benefit from increased efficiency using these explainable models, particularly the attention vector approach.

In this study, Anand et al. [P67] reframed legal document summarization as a binary classification challenge, distinguishing sentences as either vital or non-vital. Observing that some judgments include a preliminary summary known as a headnote, they introduced a unique dataset generation technique utilizing this reference summary. This approach sidesteps the need for domain experts. Their proposed methodology allows for creating legal document summaries without requiring intricate feature engineering or specific domain expertise.

The authors delineated their methodology into two pivotal stages: first, creating labeled datasets to predict sentence significance, and then, employing multiple deep learning models to extract a document's essential elements for summarization. Sentence embedding showcased superior performance among the four labeled data generation methods proposed. Furthermore, when classifying or predicting, the LSTM-based Neural Network architecture surpassed other techniques in most scenarios.

The other approaches tested by the authors are:

- InferSent + FFNN
- Sent2Vec + FFNN
- LSTM + Glove
- LSTM + Word2vec
- Naive Bayes
- Random Forests
- TBS

Song et al. [P1] empirical study approached the summarization task using the JRC-Acquis and BillSum datasets [P27]. The PLM that obtained the best results in every metric and test set on the JRC-Acquis dataset was the DYPLOC [R32]. On the other hand, in the BillSum dataset, the Global Aware [R33] obtained the best results in most of the metrics and test sets, except the California test set where the TextRank [R34] obtained the best results.

4) Information Extraction

Information Extraction is the NLP task of extracting limited semantic content from text. It turns the unstructured information embedded in texts into structured data. Formally, given an input document D which consists of sentences $s_1, s_2, ..., s_m$ where each sentence s_i is a sequence of words $w_{i1}, w_{i2}, ..., w_{in}$ from the vocabulary V. The objective is to identify and extract structured pieces of information (entities, relationships, events, etc.) from D. This can be formalized as a function that maps the document D to a set of structured tuples $T = (e_1, r_1, e_2), (e_3, r_2, e_4)...$ where e_i are the entities and r_i are the relationships or attributes connecting the entities.

As part of this research in Legal NLP, we cite the following works:

In the study by Dragoni et al.[P38], rules were extracted from legal documents using NLP. Each rule can be described as a logical statement of the form $A \implies B$. The methodology

[25] https://clarivate.com/products/ip-intelligence/patent-intelligence-software/derwent-innovation/ Accessed on 02/04/2023

[26] https://clarivate.com/products/ip-intelligence/ip-data-and-apis/derwent-world-patents-index/ Accessed on 02/04/2023

employed was based on the StanfordParser[27]. WordNet was utilized to address language variability. Lastly, logical dependencies were extracted using the Boxer framework [R35].

In the research by Alohaly et al.[P34], an approach was developed to measure the volume of data gathered from an application by examining its privacy policy text with NLP techniques. Data collection practices were identified in the privacy policy text. A rule-based classifier was employed that examined all sentences, detecting those containing the term "collect" or its synonyms through WordNet[28] by using Core Natural Language Processing (CoreNLP) [R36], semantic relations linked with data collection practices in the privacy policy were analyzed, filtering out occurrences in a negative context.

The main limitations highlighted included a significant dependence on the types of information listed in the lexicon to discern a noun phrase in policy text as a gathered data item based on phrase comparison and similarity scores. Data collection practices were also quantified by tallying the number of ordered items. However, discrepancies were noted in the category of information type and the level of detail provided across different policies.

The task of contract element extraction was defined and automated by Chalkidis et al. [P41]. A new dataset was introduced, enabling the development of contract element extraction models. Two linear classifiers, Logistic Regression (LR) and Support Vector Machine (SVM), were tested using hand-crafted features, pre-trained word embeddings, and pre-trained POS tag embeddings. Optimal results were achieved through a hybrid method, integrating machine learning (LR or SVM with hand-crafted features, word, and POS tag embeddings) and manually formulated post-processing rules. Two datasets encompassing 11 contract element types were released. One labeled dataset contained 3500 English contracts, while the other unlabeled set included 750,000 contracts. Both were encoded to ensure privacy. Both datasets are accessible [29].

Zhang et al. [P33] constructed a statutes ontology and a case ontology to perform legal information retrieval. The authors define ontological structures that capture the Chinese legal system's statutes and judicial cases. Test on information extraction: i.e., user inputs query and system outputs related case/statutes. For extraction, use a proposal based on genetic algorithms and K Nearest Neighbours (KNN).

Kapitsaki et al. [P35] implement the Free Open Source Software License Term Extraction system (FOSS-LTE) for identifying license terms from the text software open-source licenses. The authors applied the FOSS-LTE approach to a set of license texts. To have an initial set of terms that are representative and commonly encountered in licenses, they performed a manual analysis on 25 licenses. For this reason, the input data (of license texts) are split into two sets, on which different steps are applied. The input data available are: (1) the sentences of all license texts gathered (excluding a test license set used for evaluation purposes), and (2) the sentences of the 25 manually analyzed licenses.

An initial data preprocessing phase common to all cases is used. The main steps followed were: data gathering, followed by data preprocessing (noise removal and sentence segmentation), after the creation of license terms and mapper, subsequent topic modeling and map creation, and final term to-topic matching. The topic modeling is achieved by combining Latent Dirichlet Allocation LDA [R37] with Doc2Vec [R30] and using cosine similarity. The authors evaluate their methodology in a curated License text dataset [30].

The Commercial Law Information Extraction based on Layout (CLIEL) system, designed for extracting information from legal documents irrespective of their format, structure, or layout, was introduced by Garcia et al. [P36]. Emphasis was placed on context. A Rule-based Layout Detection (RLD) phase was first applied, succeeded by integrating a proposed Rule-based Layout Detection Tree (RLDT) data structure. The RLD phase was tasked with annotating, extracting, and parsing document parts into the RLDT structure, facilitating the organized storage of identified parts and entities for subsequent processing. This study considered five data point types: (i) "Date of document", (ii) "Name of party", (iii) "Name of counterparty", (iv) "Governing law", and (v) "Jurisdiction".

Evaluation was carried out using a data set of 97 commercial law documents, with data points of interest manually identified by a domain expert to establish a benchmark dataset. From this, 20 documents were selected as a training set for generating Java Annotation Pattern Engine (JAPE) rules. Three approaches were assessed: (i) Majority Sense Baseline, (ii) Layout Insensitive, and (iii) CLIEL. The unique feature of CLIEL was its utilization of document layout for context, a method not adopted by the other two. Precision, recall, and the F-measure were the evaluation metrics used.

In the study by Sleimi et al. [P32], the goal was to generate semantic metadata using NLP techniques, encompassing a tokenizer, sentence splitter, POS-tagger, NER, and parser. A rule-based extraction was proposed for semantic metadata generation. An approach was developed to tag metadata for each phrase by establishing rules. Each document was analyzed phrase by phrase, consulting a rule table to determine the appropriate concept for the phrase, which was then labeled accordingly. Twelve concepts with their respective rules were presented. A total of 150 traffic laws were manually annotated, complemented by 200 other pre-annotated laws, resulting in 1127 ground-truth annotations.

Evaluation metrics included perfect match, partial match, misclassified, and missed. From 1100 predicted annotations, 873 were perfectly matched, 196 were partially matched, 31 were misclassified, and 58 were omitted. The utilized

[27]http://nlp.stanford.edu/software/lex-parser.shtml Accessed on 02/04/2023

[28]https://wordnet.princeton.edu/ Accessed on 02/04/2023

[29]http://nlp.cs.aueb.gr/software_and_datasets/CONTRACTS_ICAIL2017/index.html Accessed on 02/04/2023

[30]https://www.cs.ucy.ac.cy/index.php Accessed on 02/04/2023

dataset was the Traffic laws dataset for Luxembourg; the exact reference was unspecified, with the primary source being truncated in the provided information. However, the main source should be from [31].

In the research by Ji et al.[P40], information extraction from court record documents was addressed. The task was formulated as a joint learning of two tasks: paragraph classification and sequence labeling, common for NER. A BiLSTM + Attention-based architecture with a shared core was jointly trained. This architecture had two independent heads for the tasks and a final Conditional Random Field (CRF)[R38] layer. Compared to prior methods, a 72% achievement was recorded in legal evidence information extraction using this method. A primary limitation identified was the extended length of law documents. To counter this, a paragraph classification task was suggested for joint future training.

The study by Ge et al. [P39] centered on discerning fact-article correspondence, evaluating the relevance of a Law Article L to fact F of a case. A corpus with manually annotated fact-article correspondences was developed. This correspondence was treated as a text-matching problem, a binary output text classification with two inputs. To address the intricacy of legal text, articles were parsed into premise-conclusion pairs using random forests. A relevant corpus was presented in the paper. References were provided to other legal resources, such as ECHR (cases by the European Court of Human Rights) and CJO (China Judgments Online). Embeddings from the legal language model, Legal-Roberta, were utilized for word vector representation in the research.

Complex texts that models find challenging to comprehend were identified as limitations by the authors. To address these limitations, the specific structure of legal articles was proposed for exploitation. Specifically, it was observed that articles typically follow a premise-conclusion pair format.

```
(if <circumstances,crime,etc> then <penalty,sentence, etc>)
```

A significant advancement in the field was made by the work of Yoshioka et al.[P37], which surpassed the state-of-the-art in the Competition on Legal Information Extraction/Entailment (COLIEE)[32] statute law legal textual entailment task (also known as task number 4 in the competition). A BERT-based ensemble method coupled with data augmentation was proposed to address the COLIEE's statute law legal textual entailment task. For this task, a system was to be developed to determine if a provided legal article confirms a given question statement. Multiple BERT fine-tuning models were constructed, and an appropriate model ensemble was selected, considering the non-deterministic nature of BERT fine-tuning and question variability.

An accuracy of 0.7037 was achieved for the statute law legal textual entailment task using their proposed method. The implementation utilized an ensemble of BERT models. The question and article were concatenated using a sentence-separator token ([SEP]) and inputted into the BERT model to determine whether the article entailed the question (positive:1) or not (negative:0). Ten additional models were explored without data augmentation for the ensemble model creation; the average probability of positive and negative from the target models was used. The dataset was sourced from task number 4 in COLIEE, known as the COLIEE Statute Law Task, available in Japanese and English on the official site.

On the other hand, the authors mentioned a set of limitations, like the fact that the system performs poorly for difficult questions, suggesting common problems that nearly all submitted systems cannot handle at this moment. Hard to handle cases like:

1) Main terms are found in both the question and the first sentence. Systems typically indicate a positive (entailment) for such questions. Yet, a match is also observed with the last sentence that details an exceptional case of the articles. Consequently, the provided article doesn't entail the question. Given the prevalence of such exceptional cases in many articles, a data augmentation method to address these articles might be beneficial.
2) Creating a straightforward data augmentation method to address this type of logical mismatch is challenging.

5) Question Answering and Information Retrieval

Question Answering is the NLP task where using a set of documents as a knowledge base answers a given question. Information Retrieval under NLP encompasses the retrieval of all media based on the user needs related to a topic. Formally, given a question Q and a sequence of words $q_1, q_2, ..., q_j$ from the vocabulary V; and a set of documents or knowledge base $KB = D_1, D_2, ..., D_n$ of documents where each document D_i is a sequence of sentences $s_1, s_2, ..., s_m$ where each sentence s_i is a sequence of words $w_{i1}, w_{i2}, ..., w_{in}$ from the vocabulary V. The objective is to find an answer A to the question Q based on the information provided in KB. The answer A can be a sequence of words, a specific value, or a pointer to a segment in one of the documents in KB. In the specific case of Information Retrieval, the problem finishes before obtaining A; the main objective is obtaining the set $R \subset KB$ of documents that are most relevant to the given question Q.

We cite the following articles as the more relevant works in these tasks under the Legal NLP field:

In QA and IR tasks, techniques capturing text similarity effectively are crucial. This is exemplified in the study by Landthaler et al. [P43], where a word embedding approach was proposed. The aim was to determine the similarity between a vector representing the entire search via summation and another vector of identical size that retains the original order through summation.

Word2Vec was utilized by the authors to achieve word embedding, and cosine similarity was employed to calculate the similarity between vectors. The vector of a search was

[31] https://police.public.lu/en/legislation/code-de-la-route.html Accessed on 02/04/2023

[32] https://sites.ualberta.ca/~rabelo/COLIEE2021/ Accessed on 02/04/2023

computed by summing the vectors of each word in the query. All words in the documents were then iterated over, and a comparison vector was computed using a window size of n/2, where n is the search query length. After identifying the top X results similar to the search query through cosine similarity, all X results were concatenated. The selected words were then shifted sequentially, and similarity was recalculated to determine if a superior result existed.

The model was trained using the German Civil Code (GCC) [33]. The results were tested using a set of 10 German rental contracts, but the authors do not refer to the source. Also, it was tested on the e EU Data Protection Directive (EU-DPD) [34]

A system that addresses a Bar Examination written in Natural Language was introduced by John et al. [P50]. A BiLSTM + Attention architecture, combined with Glove word embeddings, was employed for the task. The dataset was derived from the MultiState Bar Examination (MBE). [35]. However, their corpus is available only on request.

Two primary objectives were set out by Locke et al [P48]. The first was to explore the utility of keyword extraction or query reduction methods in automatically generating queries for case law retrieval. The second was to address the absence of a suitable test collection for evaluating these methods. A test collection was subsequently created and made available by the authors [36] for this purpose. They also assessed existing keyword extraction methods using their dataset. Results indicated that while these methods matched the efficacy of average Boolean queries crafted by experts, they fell short when compared to keyword queries and the optimal Boolean queries from experts.

Methods for information retrieval and answering legal questions were proposed by Do et al.[P54]. The Competition on Legal Information Extraction/Entailment (COLIEE) 2016 data[37] served as their dataset. Six features, namely TF-IDF, Euclidean, Manhattan, Jaccard, LSI, and LDA, were utilized for information retrieval. A Ranking SVM was trained using pairs of queries and articles, drawing from some of these features. At inference, scores for articles relative to the given query were generated using the trained SVM. The experimental findings indicated that an amalgamation of LSI, Manhattan, and Jaccard yielded the most effective results.

Question Answering is perceived as a form of textual entailment, characterized as binary classification. Word embedding is derived using Word2Vec through the Continuous Bag Of Words (CBOW) technique. Sentence embedding is then achieved by summing all word embeddings and normalizing by the sentence's word count. The embeddings of the question and article are merged and fed into a CNN. Post-network, the output is amalgamated with LSI and TF-IDF features, which are then channeled through two perceptron layers for output prediction. An ablation study indicates that the fusion of LSI and TF-IDF with the CNN's output yields the most optimal results.

The study by Nejadgholi et al. [P45] centers on a semantic search designed to locate legal cases with facts mirroring a provided query. A two-step model is proposed by the authors. Initially, Word2Vec, Skip-gram, and FastText word embeddings are acquired. Using the FastText library's supervised model and an annotated dataset, a binary classifier is trained to identify fact-asserting sentences in immigration cases automatically. The model then calculates the cosine similarity between the input sentence and all recognized fact sentences, discarding non-fact sentences.

The authors contend that, despite the potential similarity, non-fact sentences can yield inaccurate results, as portions of a case could discuss hypothetical situations merely considered, not finalized. The utilized dataset comprised 46,000 immigration and refugee cases sourced from Canada's Federal and Supreme Court websites. From these, 150 cases were randomly chosen and manually annotated. However, this dataset was neither released nor named.

In their research, Sugathadasa et al. [P42] sought methods for generating meaningful vectors from legal documents. Their strategy melded two techniques for vectorizing a legal document to identify and rank pertinent legal records: doc2vec_NV and doc2vec_SSM. The Word2Vec technique was subsequently employed, where a document's one-hot encoding served as the input and its associated mention list as the output. The resultant vector in the latent space was produced by doc2vec_NV. Conversely, doc2vec_SSM assessed each word's significance to the entire dataset utilizing TF-IDF equations and then formed a document vector using these scores. In their testing, recall was chosen as the evaluation metric.

It was observed that doc2vec_SSM lagged behind both doc2vec_NV and doc2vec_NN. Additionally, doc2vec_NN, an ensemble version of doc2vec_NV and doc2vec_SSM, surpassed doc2vec_NV. The dataset comprised over 2,500 legal cases sourced from Findlaw [38].

In contemporary NLP, Information Extraction's leading methods have encompassed the Named-Entity Linking (NEL) subtask. It was demonstrated by Elnaggar et al.[P53] that transfer learning could be effectively applied from high-performing NEL models to legal documents. Furthermore, the authors introduced a NEL dataset specific to the legal domain named EURLEX 20k obtained from the European Union law and other public documents of the European Union (EU) web. The primary deep learning architecture they drew upon was devised by Ganea et al.[R39]. This structure employs a method rooted in entity embedding and a local model with

[33]https://www.gesetze-im-internet.de/englisch_bgb/ Accessed on 02/04/2023

[34]https://www.datenschutz-grundverordnung.eu/wp-content/uploads/2016/05/CELEX_32016R0679_EN_TXT.pdf Accessed on 02/04/2023

[35]http://www.ncbex.org/exams/mbe/ Accessed on 02/04/2023

[36]https://github.com/ielab/ussc-caselaw-collection Accessed on 02/04/2023

[37]https://webdocs.cs.ualberta.ca/~miyoung2/COLIEE2016/ Accessed on 02/04/2023

[38]https://www.findlaw.com/ Accessed on 02/04/2023

neural attention, collaboratively considering entities' semantic meanings and the context of words.

The challenge of discerning entailment relationships between case law documents, a task within the Competition on Legal Information Extraction and Entailment (COLIEE), was addressed by Rabelo et al. [P56]. An F-score of 0.70 was achieved on the COLIEE test dataset. Their strategy centered on extracting similarity measures between two text segments using the cosine similarity of the BERT representations. Subsequently, a threshold-based classifier was employed, and the outcomes were post-processed, considering the a priori probability dictated by the training samples' data distribution.

The Legal Case Retrieval Task of COLIEE 2019 was addressed by Shao et al. [P55]. A new BERT-based model, including Paragraph-Level Interactions (BERT-PLI), was introduced by the authors, modeling paragraph-level interactions in case documents using BERT. These interactions were then consolidated to deduce document relevance via a sequential modeling process. It was demonstrated through experimental results that their method surpassed existing solutions at that time.

The employed methodology initially pruned the candidate set based on BM25 rankings. To amplify the capability to understand semantic ties between legal paragraphs, the BERT model was fine-tuned using a readily available entailment dataset in the legal field before its integration with BERT-PLI. Such fine-tuning facilitated BERT's inference of supportive paragraph relationships, proving beneficial for the legal case retrieval task.

A retrieval-based legal Question Answering model, which learns attentive neural representations of both the input question and legal articles, was presented by Kien et al.[P44]. The authors demonstrated the model's efficacy by offering an annotated corpus and performing experiments comparing their model to leading methods in the domain. The proposed model comprises two distinct encoders: the Sentence Encoder and the Paragraph Encoder. The Sentence Encoder is crafted using word embeddings and a CNN framework. The Paragraph Encoder determines a sentence's attention weight by averaging the attention weights of its constituent words. Observing that not every sentence adds to the paragraph's meaning, the authors substituted softmax with sparsemax[R40].

Two datasets were constructed by the authors: first, the legal document corpus comprising Vietnamese legal documents; second, the QA dataset encompassing a collection of legal questions (queries) and associated relevant articles for each inquiry. Raw legal documents were initially sourced from official online sites. While the paper furnishes links to all utilized websites for dataset creation, a method to access the complete dataset remains to be provided by the authors.

The Document to Document (Doc2Doc) problem, aimed at Information Retrieval, is defined by Chalkidis et al.[P57]. Given legislation from the European Union (EU) and the United Kingdom (UK), the objective is to identify pertinent documents when a document from one legislation serves as the query for the other. Typically, a two-step approach is adopted. The initial step, document prefetching, retrieves the k most relevant documents to boost recall. The algorithms employed to derive a document's embedding include Best Match 25 (BM25), Words to Vector Centroids (W2VCent) [R41], BERT, Sentence-BERT (S-BERT) [R42], Legal-BERT, C-BERT [39], and Ensemble. It's worth noting that C-BERT is a BERT version pre-trained via a classification task based on the multilingual thesaurus maintained by the Publications Office of the European Union and hosted on the portal Europa (EUROVOC). [40].

Consequently, in the EU2UK context, C-BERT attains the highest accuracy, while for US2EK, BM25 emerges as the most accurate. Moreover, the Ensemble of C-BERT and BM25 surpasses other methods. The subsequent step involves reranking the k relevant documents, but this doesn't enhance performance from the initial step. The dataset utilized comprises Legal documents sourced from China Judgement Online. [41]

The Chinese Legal Case Retrieval Dataset was developed by Ma et al. [P49], introducing a range of relevant judgment criteria formulated by domain experts, specifically a legal team. Over 46,000 documents were collected from China Judgments Online, which the authors then processed by discarding smaller samples and anonymizing content. The data was annotated by human specialists, ensuring each entry was reviewed by a minimum of three annotators.

The annotation was based on relevance, utilizing judgment criteria that accounted for both subjective and objective evaluations. Several existing Information Extraction models, spanning traditional Bag of Words models to deep learning, were applied to the dataset for assessment. Notable challenges were an uneven distribution of charges and the intricacies of sampled queries for information retrieval. The dataset is made available by the authors [42].

Vold2021 et al.[P51] detail the deployment of a RoBERTa Base[R43] question-answer classification model for production use. They juxtapose the performance of a RoBERTa-base classifier with a conventional machine learning model in the legal sphere, assessing the disparity in performance between a trained linear SVM and the publicly sourced Privacy QA dataset. The authors demonstrated that RoBERTa registers a 31% enhancement in F1-score and a 41% increment in Mean Reciprocal Rank compared to the conventional SVM.

In their participation in the Automated Legal Question Answering Competition (ALQAC) 2021, Tieu et al. [P52] addressed three tasks. These tasks utilized data primarily sourced from legal webpages provided for the competition [43]:

[39] No specified on the paper what is the actual meaning of C. Gives the feeling to be related to the author's name.

[40] http://www.lt-innovate.org/lt-observe/resources/eurovoc-%E2%80%93-eus-multilingual-thesaurus Accessed on 02/04/2023

[41] https://wenshu.court.gov.cn/ Accessed on 02/04/2023

[42] https://github.com/myx666/LeCaRD/tree/main/data Accessed on 02/04/2023

[43] https://www.jaist.ac.jp/is/labs/nguyen-lab/home/alqac-2021/ Accessed on 02/04/2023

Task 1 Document Retrieval: First, retrieve documents using elastic search. Then, apply a more fine-grained filter via a fine-tuned BERT for domain-specific Vietnamese legal text.

Task 2 Textual Entailment: Framed as text classification, using the same model as above with different fine-tuning. Augment data by crawling web sources.

Task 3 Framed as a combination of the first two tasks (alternative 1) or as a sentence classification problem (alternative 2), in which case, using the same BERT-based model.

Abualhaija et al. [P47] put forward an automated Question Answering (QA) method aimed at aiding requirements engineers in identifying legal text segments pertinent to compliance requirements. They employed large-scale language models, notably BERT, A Lite BERT (ALBERT), RoBERTa, and Efficiently Learning an Encoder that Classifies Token Replacements Accurately (ELECTRA), which were fine-tuned for QA. The authors constructed a dataset encompassing 107 questions along with their corresponding answers. Yet, in their paper, a method to access this dataset remains to be provided.

Strkak et al.[P68] developed NLP-centric techniques to search for semantically analogous Polish private tax rulings. They refined a pre-existing BERT model for the Polish Individual Tax Interpretations dataset sourced from the Ministry of Finance website[44]. Ultimately, clusters were discerned by harnessing the contextual embeddings derived from BERT and applying the cosine similarity metric. One constraint highlighted was that substantial NLP language models demand an extensive corpus, yet no such corpus is present in the legal domain.

The empirical study of Song et al. [P1] showed that in the Question Answering dataset, the best PLM in the accuracy metric is the Custom LegalBERT.which shows a remarkable superiority on this task was mainly because it acquired domain knowledge via pre-training on 3.5 million U.S. cases. On the Information Retrieval task, the authors use the COLIEE-2021 dataset [45] where RoBERTa shows superiority on the Macro F1 metric. On the other hand, the LMIR[R44] showed the best results in the micro F1 metric.

[44]https://sip.mf.gov.pl Accessed on 02/04/2023

[45]https://github.com/sophiaalthammer/dossier_coliee Accessed on 02/04/2023

6) Coreference Resolution

Coreference Resolution is the task in NLP of finding all the expressions that refer to the same entity in a text. This task is essential to high-level NLP tasks like Natural Language Understanding (NLU), IR, IE, Question Answering (QA), and Summarization. Formally, given an input document D which consists of sentences $s_1, s_2, ..., s_m$ where each sentence s_i is a sequence of words $w_{i1}, w_{i2}, ..., w_{in}$ from the vocabulary V; and a set of mentions M within D, where each mention $m \in M$ refers to a segment of text in D. The objective is to group the mentions in M into clusters $C_1, C_2, ..., C_k$, such that all the mentions in a given cluster refer to the same real-world entity.

As part of the ongoing research of this task in the Legal NLP field, we reference the following relevant works:

Sannier et al.[P58] centered on detecting and resolving cross-references in legal texts via NLP. The Luxembourg's Income Tax Law and PHIPA legal corpus[46] were employed for the task. Initially, a legal text was annotated using Tokenizer, Sentence Splitter, and NER modules. Subsequently, the Java Annotation Patterns Engine (JAPE), a rule-driven language, was applied. The researchers intended to assess their strategy for cost-effectiveness and applicability beyond legal contexts. Given that a rule-based method is utilized for cross-reference detection, completeness is not assured. A primary limitation of this study is the absence of a provided standard for cross-references, necessitating the authors to establish one themselves.

Ji et al. [P60] aimed to identify speaker coreference resolution in legal texts, eschewing the use of external domain knowledge. The proposed model comprises three primary modules: a span-representation module encoding contextual data, a Graph Neural Network (GNN) module integrating established relations, and a multi-scoring mechanism producing coreference scores.

The model presented follows several stages. Initially, word embeddings and the BERT output vector are amalgamated to form final word representations, followed by a multi-layer BiLSTM encoding sentence details. Subsequently, a graph neural network integrates the mentioned-by relation and entity mapping relation. Lastly, a multi-scoring mechanism, comprising a biaffine attention model and a feed-forward neural network, calculates candidate scores. The definitive scores for candidates merge predictions from both classifiers in a specific ratio. Should a candidate's score surpass a set threshold, its antecedent is preserved; otherwise, it's eliminated.

Pothong et al. [P59] focused on extracting semantic meaning through coreference resolution and Abstract Meaning Representation (AMR). The Conventional of the Rights of the Child (CRC) dataset, sourced from Refworld, Convention

[46]https://people.svv.lu/sannier/crossreferences/ Accessed on 02/04/2023

of the Rights of the Child [47], is employed. This dataset is segmented into three parts, each encompassing individual articles and their respective statements. Their approach involves preprocessing tasks, utilizing Regular Expressions for Roman and Arabic numbers and Spacy [48] for sentence segmentation. Dependency Parsing and Part of Speech tagging are executed using Spacy, while the Span of Text BERT(SpanBERT) model addresses Coreference Resolution. The Abstract Meaning Representation is managed by the amrlib [49] and Spacy. Evaluations are conducted via Smatch [50] and Bilingual Evaluation Understudy (BLEU).

7) Cross Lingual Transfer Learning

Cross-lingual transfer refers to transfer learning using data and models available for one language with higher resources (e.g., English) to solve tasks in another, commonly more low-resource, language.

Shaheen et al.[P61] establish a baseline for LMTC using two multilingual datasets with parallel documents in English, French, and German. English serves as the training set, with German and French used for testing. The primary datasets deployed are JRC-Acquis and EURLEX57K. For training and transferring learning, BERT and DistillBERT[R45] are the chosen language models.

The primary constraint highlighted pertains to dataset accessibility for LMTC tasks. To address this, the authors suggest training an LTMC for low-resource languages in zero-shot settings, leveraging data from other languages, and subsequently making predictions in the unseen target language. By employing transfer learning, a classifier can be trained with datasets in certain languages (source languages), and the knowledge is then transferred to different languages (target languages). Such a transfer learning strategy is advantageous for tasks demanding substantial data, especially in languages with limited resources.

B. RQ 2: WHICH RESOURCES (DATASETS, ONTOLOGIES, WEB SCRAPPED, ETC...) ARE BEING USED FROM THE LEGAL DOMAIN TO APPLY AND ENHANCE NLP?

This question was answered in the previous section since it would have made it harder to read to search for the dataset in one section and the approach in the other section to solve a Legal NLP task. However, this section provides a more focused discussion according to the resources available in the Legal NLP domain. We present primary studies primarily focused on creating and introducing a new dataset. In addition, we provide Table 7, which maps every dataset extracted from the previous papers to its respective Legal NLP task with the reference of where to find the resource. However, we wanted to mention that multiple of these works have been using the resources provided by the Competition on Legal Information Extraction/Entailment (COLIEE) [51].

Wilson et al. [P19] curate an extensive collection of privacy policy documents. Utilizing the Skip-Gram word embeddings, they pre-train it on their specific datasets. The authors introduce the OPP-115 dataset, paving the way for researchers to develop models tailored to online privacy policies. Additionally, they unveil a web-based tool designed for adept annotators to implement the annotation scheme on chosen privacy policies.

Wyner et al. [P70] present a legal corpus sourced from National Conference of Bar Examiners (NCBE) materials and adapted for a textual entailment task on the Excitement Open Platform. The initial dataset features one hundred questions, each with four potential answers. With an answer key provided by the NCBE, each question was paired with one of its potential answers, resulting in four hundred theory-hypothesis pairs.

Oltramari [P71] introduces an ontology that domain experts derived from the OPP-115 dataset. The authors crafted this ontology to depict unstructured policy content in line with frame-based structures detailed using Ontology Web Language - Description Logic (OWL-DL). This ontology was integrated into an Apache Jena Fuseki server to facilitate dynamic operations. The server, accessible and deployed, offers a web service framework enabling various applications to retrieve data via SPARQL Protocol and RDF Query Language (SPARQL) queries. [52].

Manor et al. [P72] present a dataset comprising 446 parallel text sets. The authors demonstrate the degree of abstraction by highlighting the increased count of unique words in the reference summaries compared to the abstractive single-document summaries from the 2002 Document Understanding Conference (DUC) [R46], a benchmark dataset for single document news summarization. Furthermore, using various prevalent readability metrics, they reveal an average reading level difference of 6 years between the original documents and the reference summaries within their legal dataset.

Kornilova et al. [P27] present the BillSum dataset comprising 22,218 US Congressional bills and their corresponding summaries, divided into training and test sets. Furthermore, an additional test set of 1,237 California bills with their summaries is included to promote model applicability to different legislatures. The authors set multiple benchmarks, indicating significant potential for innovative approaches to effectively summarize complex legislative verbiage.

Duan et al.[P46] present the CJRC dataset, marking the inaugural Chinese judicial reading comprehension dataset designed to address existing gaps in legal studies. This dataset spans a broad spectrum, encapsulating 188 distinct causes of action and 138 specific criminal charges. Its applications

[47] https://www.refworld.org/docid/3ae6b38f0.html Accessed on 02/04/2023
[48] https://spacy.io/
[49] https://amrlib.readthedocs.io/en/latest/
[50] https://github.com/snowblink14/smatch
[51] https://sites.ualberta.ca/~rabelo/COLIEE2022/ Accessed on 02/04/2023
[52] https://explore.usableprivacy.org/ Accessed on 02/04/2023

TABLE 7. List of datasets and their corresponding tasks, papers and access link.

Dataset	Legal NLP Tasks	References	Link
CAIL2018; CAIL-Long	LM	[P5]	https://paperswithcode.com/dataset/chinese-ai-and-law-cail-2018
Case-HOLD	LM, IE	[P1, P4]	https://paperswithcode.com/dataset/casehold
LexGLUE	LM, Mult. Class., IE, QA/IR	[P13]	https://huggingface.co/datasets/lex_glue
SCOTUS	Mult. Class.	[P7]	http://supremecourtdatabase.org
CJO	Mult. Class.	[P18]	http://wenshu.court.gov.cn/
PKU	Mult. Class.	[P18]	http://www.pkulaw.com/
CAIL	Mult. Class.	[P18]	http://cail.cipsc.org.cn/index.html
OPP-115	Mult. Class.	[P19]	https://www.usableprivacy.org/data
ECHR	Mult. Class.	[P9, P24, P39]	https://archive.org/details/ECHR-ACL2019
SigmaLaw ABSA	Mult. Class.	[P22]	http://www.cs.ucy.ac.cy/âĹijgkapi/foss.html
Terms of Service	Mult. Class.	[P15]	http://claudette.eui.eu/ToS.zip
DMOZ	Mult. Class.	[P69]	https://tinyurl.com/y43htvum
POSTURE50K	Mult. Class.	[P1, P12]	https://rb.gy/fzsp1
ILSI	Mult. Class.	[P17]	https://github.com/Law-AI/LeSICiN
Legal Cases from the Federal Court of Australia	Summ.	[P67]	https://archive.ics.uci.edu/ml/datasets/Legal+Case+Reports
BilSUM	Summ.	[P1, P27]	https://github.com/FiscalNote/BillSum
LegalSUM	Summ.	[P67]	https://github.com/lauramanor/legal_summarization
Civil Trial Court Debate	Summ.	[P28]	https://github.com/zhouxinhit/Legal_Dialogue_Summarization
COLIEE Statute Law Task	IE	[P1, P37, P54, P56]	https://sites.ualberta.ca/~rabelo/COLIEE2021/
License texts	IE	[P35]	http://www.cs.ucy.ac.cy/âĹijgkapi/foss.html
Contracts	IE	[P41]	http://nlp.cs.aueb.gr/software_and_datasets/CONTRACTS_ICAIL2017/index.html
Query Generation	QA/IR	[P48]	https://github.com/ielab/ussc-caselaw-collection
EURLEX57k	QA/IR, Cross-Lingual	[P1, P10]	https://paperswithcode.com/dataset/eurlex57k
ALQAC-2021	QA/IR	[P52]	https://www.jaist.ac.jp/is/labs/nguyen-lab/home/alqac-2021/
CJRC	QA/IR	[P46]	https://paperswithcode.com/dataset/cjrc
JEC-QA	QA/IR	[P62]	https://jecqa.thunlp.org/
LeCARD	QA/IR	[P49]	https://github.com/myx666/LeCaRD
Bar Exam QA	QA/IR	[P70]	http://www.kaptest.com/bar-exam/courses/mbe/multistate-bar-exam-mbe-change
CJO	QA/IR	[P18]	https://wenshu.court.gov.cn/
CRC	CoRef. Res.	[P59]	https://www.refworld.org/docid/3ae6b38f0.html
Luxembourg's Income Tax Law	CoRef. Res.	[P58]	https://people.svv.lu/sannier/crossreferences/
License texts	CoRef. Res., Cross-Lingual	[P35]	http://www.cs.ucy.ac.cy/âĹijgkapi/foss.html

are varied, encompassing areas like information retrieval and factor extraction. Benchmark results, tested against several potent baselines, including BERT-based methods and BiDAF[R47], suggest considerable room for enhancement compared to human annotator performance.

Lippi et al. [P15] unveiled a fresh corpus dedicated to Terms Of Services (ToS), comprising 50 contracts, equivalent to over 12,000 sentences, thus enhancing the potential for method training and evaluation. The authors probe into contemporary deep learning architectures tailored for text categorization and also explore a structured SVM crafted for collective classification, taking into account the sentence sequences within a document. Furthermore, they introduce the Automated Detector of Potentially Unfair Clauses in Online Terms of Service (CLAUDETTE) web server to the public, facilitating users in submitting their query documents and independently assessing the efficiency of the proposed methods.

Zhong et al. [P62] developed a legal dataset tailored for question-answering, sourced from the National Unified Legal Professional Qualification Examination Counseling Book and various Chinese legal provisions. This dataset encapsulates five reasoning categories: word matching, concept comprehension, numerical evaluation, reading across multiple paragraphs, and intricate multi-hop reasoning. The latter two categories pose the greatest challenges, as they demand synthesizing information across various sections and executing several reasoning steps for answering.

The dataset is bifurcated into Knowledge-driven questions (KD-questions) and Case-analysis questions (CA-questions). In KD-questions, while the Co-matching model achieves a modest 25.37% accuracy, laypeople and experts secure scores of 71% and 77%, respectively. Despite Co-matching boasting the best accuracy, it lags considerably behind human performance. In the broader scope of CA questions, the Multi-matching model leads with an accuracy of 29.06%. Yet, when compared to laypeople and experts who achieve 58% and 84% accuracy, respectively, it's evident there's a substantial gap. This dataset underscores the complexity of question-answering within legal documents, as current machine learn-

ing models still fall markedly short of human capabilities.

A manually annotated legal opinion text dataset, SigmaLaw-ABSA, was introduced by Mudalige et al. [P64], aimed at aiding ABSA tasks in the legal domain. Results on the performance of several deep learning-based systems on the SigmaLaw-ABSA dataset were also presented by the authors. The corpus encompasses 39,155 legal cases, of which 22,776 were sourced from the United States Supreme Court. Approximately 2000 sentences were collected for annotation, and the court cases were chosen without a focus on any particular category.

A new multilingual, diachronic Legal Judgement Prediction (LJP) dataset from the Federal Supreme Court of Switzerland (FSCS) cases was introduced by Niklaus et al. [P73]. This dataset spans 21 years (from 2000 to 2020) and comprises over 85K cases: 50K in German, 31K in French, and 4K in Italian. For the baseline, a classical classification architecture was employed, utilizing two different variants of BERT: a native variant and a multilingual one.

The largest privacy policies dataset at the moment was introduced by Nokhbeh et al.[P69]. Using DMOZ[53], an open-content directory of the web with 1.5 million manually categorized websites, hundreds of thousands of privacy policies associated with their categories were collected. From this collection, a new dataset [54] comprising a corpus of over 100K web privacy policies was constructed. There is an intention for future work to enhance the corpus by incorporating more granular subcategories from DMOZ.

LexGLUE, a new benchmark for Legal Natural Language Understanding (NLU), was proposed by Chalkidis et al. [P13]. Seven complex, publicly available English datasets were gathered by the authors, ensuring they were large enough for evaluating various models. Models assessed across these datasets included linear SVM and several PLMs such as BERT, RoBERTa, DeBERTa [R48], Longformer [R21], BigBird [R28], Legal-BERT, and CaseLaw-BERT [P4]. To address the challenges posed by long texts, a hierarchical variant of the PLMs was employed. In this approach, paragraphs were initially encoded, and subsequently, the representation of each paragraph was used as a sequence to encode the entire document.

The work of Listenmaa et al. [P74] introduced the CNL, a component of L4, a domain-specific language crafted for drafting laws and contracts. Along with other functionalities, Natural Language Generation and an interactive process for ambiguity resolution were also incorporated by the authors.

C. RQ 3: WHAT ARE THE LIMITATIONS OF THE CURRENT WORK OF NLP APPLIED TO THE LEGAL DOMAIN?

With this research question, we intend to show the current limitations in the Legal NLP field given our study. In addition, we also show the solutions and proposed future works from primary studies for these gaps and challenges.

[53]https://curlie.org/ Accessed on 02/04/2023
[54]https://tinyurl.com/y43htvum Accessed on 02/04/2023

It was observed by Fang et al. [P24] that the count of labeled judicial documents often falls below the dimensionality of features inherent to these documents. Such a scenario can degrade the prediction performance when directly utilizing these extracted text features. To address these challenges, non-linear dimension reduction techniques were proposed, aiming to preserve distances between points in reduced dimensions.

The primary limitations highlighted by Pillai et al. [P11] include the absence of unified, standard legal procedures across nations and the scarce availability of cross-country data for evaluating diverse legal texts. Furthermore, it was emphasized that legal text often harbors considerable irrelevant content.

Three challenges in Legal Artificial Intelligence (AI), especially in the Legal NLP domain, were explored by Zhong et al. [P75]. Which are:

Knowledge Modelling Legal texts are mainly well formalized due to their nature, and a lot of knowledge and concepts can be used with high importance. But are not used in multiple recent works due to a lack of proper modeling of all this knowledge across legal documents.

Legal Reasoning The reasoning of Legal NLP tasks differs from the usual in several other NLP tasks. The legal rationale must strictly follow the rules well-defined in law. This implies the need to consider the rules already present in the legal domain in the NLP approaches.

Interpretability As mentioned in previous limitations, the legal language is quite complex. On top, of that, due to the nature of the legal domain, any legal decision or prediction should be interpretable to be applied to the actual legal system.

The research by Zhong et al. [P75] highlighted the ethical concerns associated with high-performance Legal NLP algorithms. When integrated into the legal system, such technology must remain free from issues like bias, racial discrimination, and uninterpretable results that fail to persuade individuals. It was emphasized that developments in this domain should aim to assist, rather than replace, legal professionals.

The considerable length of law documents, a primary limitation, was discussed by Ji et al.[P40]. In anticipation of addressing this challenge, a paragraph classification task was proposed, emphasizing joint training. The issue of extended texts was reiterated by Shao et al.[P55]. The authors elaborated that in the legal context, the notion of relevance exceeds the conventional definition of topical relevance. Relevant cases often align with the current case's decision, encompassing analogous situations and applicable statutes.

Therefore, identifying similarities in the legal issues and processes of cases is crucial, necessitating a comprehensive semantic understanding of entire documents. The collection of a substantial dataset for this task might pose challenges. In many legal systems, the downloading of large-scale legal documents is restricted. Moreover, acquiring accurate relevance judgments is often costly due to the need for expert knowledge in the legal domain. The data scarcity hinders the training of deep neural models.

The limitations of legal resources for research in Legal NLP and multilingual constraints were examined by Shaheen et al. [P61]. It's not solely about resource scarcity; very few of these resources are available in multiple languages. The challenges presented by low-resource languages in various NLP domains further complicate achieving optimal performance in Legal NLP using current state-of-the-art approaches.

The uneven distribution of charges and the complexities associated with sampled queries for information retrieval in the legal domain was highlighted by Ma et al. [P49]. To address these challenges, a method was proposed by the authors to categorize the queries into two types, considering both the difficulty and distribution of the query.

Multilingual legal resource limitations were highlighted by Chalkidis et al. [P13], emphasizing the importance of developing Legal NLP models in languages other than English. The challenges of creating new datasets and resources in the legal domain, irrespective of the language, were detailed by the authors. Legal barriers impede dataset creation, including copyright protections for critical documents like contracts and trade secret designations.

Additionally, bureaucratic processes often restrict access to court decisions. The absence of human evaluation in existing legal datasets was identified as another challenge. Some datasets, such as LexGLUE presented by the authors, lean on ground truth labels derived automatically from sources like court decisions. The quality assessment of these resources lacks a definitive and reliable benchmark.

The confusion and ambiguity of legal languages in the criminal context were emphasized by Lyu et al. [P65]. Due to this issue, different criminals and targets were observed to have indistinguishable fact descriptions.

A shortage of resources in the Legal NLP field was highlighted by Akcca et al.[P23], ranging from missing benchmarks in certain tasks to the dearth of well-curated datasets across various Legal NLP subdomains. The complexity of legal language, as emphasized in works like[P37], results in challenges for tasks like Question Answering. When paired with complex questions, current state-of-the-art approaches often fall short in performance and desired outcomes.

A limitation highlighted by Qin et al. [P6] is that the classification efficacy of PLMs diminishes as the semantic composition and intricacy of documents escalate, a phenomenon often observed in legal documents. As a remedy, the authors proposed using transformer variants that utilize attention approximations with reduced complexity.

The lack of a standard benchmark dataset and the inherent complexity of the legal domain were noted as limitations by Song et al. [P12].

In a recent empirical study, limitations of domain-specific PLMs, such as Legal-BERT, were demonstrated by Song et al. [P1] due to variations in legal subdomains and language across diverse legal documents.

Table 8 summarizes the limitations found in Legal NLP and possible solutions according to the literature reviewed.

V. DISCUSSION

This section discusses our results, focusing mainly on our findings and implications. We also discuss in which areas we believe the Legal NLP domain should focus the most given its current state of the art. Additionally, we provide Figure 2 which shows the distribution of articles from all research questions in the following categories:

Technical: A research paper that presented a new approach for at least a particular Legal NLP task.

Resource: A research paper that presented a new dataset, ontology, or other resource to be used for Machine Learning models.

Multilingual: A research paper that worked with more than one single language.

LLM-based: A research paper that presented a new LLM-based approach for at least a particular Legal NLP task.

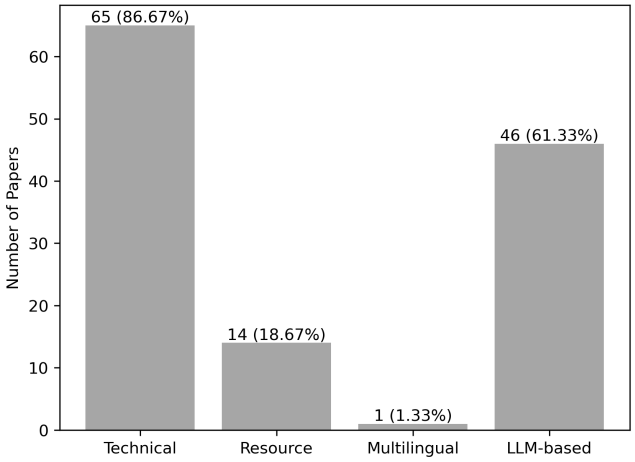

FIGURE 2. Papers distribution across the Technical, Resource, Multilingual, and LLM-based categories.

A. THE MAIN FINDINGS

First, there is a clear need for more resources, particularly labeled and curated datasets, to train supervised models. Current studies have demonstrated that even when using large transformers and language models trained on extensive general domain corpora, there is still room for improvement. Conversely, some research indicates that language models like Legal-BERT, trained exclusively on legal texts, do not consistently outperform general-purpose language models such as BERT or RoBERTa in all tasks. This suggests a significant divergence in the language used across different subdomains within the legal field.

Therefore, there is a growing demand for more curated and labeled data spanning as many legal subdomains as possible. By increasing the availability of such data, it becomes feasible to tackle challenges like resource scarcity, cases where the number of labeled judicial documents is less than the feature dimensionality, and the limited transferability of domain-specific PLMs.

TABLE 8. Legal NLP Limitations

Limitation	References	Proposed Solution
Lack of unified legal procedures across countries	[P11, P55]	Agree on a standardized legal procedure around the globe.
Knowledge Modelling	[P55, P75]	Design a proper model to represent the knowledge across legal documents.
Legal Reasoning	[P55, P75]	Include the rules already present in law when designing any solution.
Interpretability	[P75]	Explainability of the models used and a clear legal interpretation of any decision.
Ethical issues	[P75]	Inclusion of bias and fairness analysis and good interpretability.
Long Documents	[P30, P40, P55]	Paragraph classification task and train jointly in it.
Lack of resources	[P12, P55, P61]	Data augmentation
Complexity and ambiguity of the Legal language	[P1, P12, P37, P39, P55, P65]	Trained PLMs in more legal complex data. More separation of concerns in the solutions.
PLMs effectiveness decreases when the semantic composition of the documents increases	[P6]	Using variants of transformers with approximations of attention with lower complexity.
Number of labeled judicial documents is less than the dimensionality of features	[P24]	Using non-linear dimension reduction techniques.
Lack of Transferability of Domain-Specific PLMs	[P1]	Trained domain-specific PLMs in massive and more language-sparse legal data.

Second, from our prior analysis, we discern a need for more language models that are focused on the legal domain but trained on a diverse array of legal data. Models like Legal-BERT may not capture the full complexity and variation of legal language across different subdomains. Current advancements in NLP suggest that larger language models, trained with more data, tend to perform better. However, the intricacy of legal language remains a challenge, even for these sizable models. Yet, the prevailing trend in the NLP community indicates that training with an extensive amount of legal data will yield high-performing, domain-specific language models. Such training should also address one of their major limitations: the need for knowledge transferability to other legal datasets.

Furthermore, if more PLMs specialized in the legal domain emerge, we can leverage current state-of-the-art techniques, such as the Mixture of Experts, to address the limitations and challenges of PLMs, including handling language complexity. This strategy could also tackle other issues, such as processing lengthy documents, navigating the intricacies and ambiguities of legal language, and enhancing the transferability of domain-specific PLMs. Additionally, improved interpretability becomes achievable; when a group of high-performing models agrees on the importance of certain input parts, it reinforces confidence in that interpretation.

Third, based on the research reviewed, approaches using Pretrained Language Models (PLMs) typically outperform baseline systems such as SVM or BiLSTM. However, in certain tasks, SVM and BiLSTM not only yield commendable results but are also less computationally demanding than PLMs. Notably, these baseline systems have outperformed others in some information retrieval tasks. This underscores the need for increased focus on Information Retrieval within the Legal NLP field.

Fourth, we were struck by the evident need for more initiatives incorporating symbol-based methods to tackle Legal NLP challenges. This shortfall isn't exclusive to the Legal NLP domain; it's increasingly manifesting in the broader NLP landscape. Every year, larger transformer language models emerge, trained on ever-expanding datasets and boasting a growing number of parameters. Given the remarkable success of these models across various domains, a substantial proportion of researchers are pivoting towards exploring these vast models, often overlooking the existing knowledge and insights specific to fields such as Legal NLP. Research in Legal NLP must begin by integrating symbol-based methods. A compelling rationale for this inclusion is that these methods can directly address some of the prevailing challenges in the field, namely Knowledge Modeling, Legal Reasoning, and Interpretability.

On a related note, it's surprising to see a lack of research utilizing well-curated ontologies specific to the legal domain or its subdomains. Just as with datasets, there's a need for more legal domain ontologies, which will inevitably spark increased interest in implementing symbol-based methods within the Legal NLP arena. This approach will pave the way for a myriad of studies that merge state-of-the-art methods, integrating PLMs with Ontologies. Consequently, this can help address several challenges, including knowledge modeling, legal reasoning, and ethical dilemmas. This is particularly true since these solutions offer a formal and meticulously curated representation of the legal realm, as illustrated by Ontologies. Furthermore, such integration can potentially amplify the efficacy of leading-edge methods across a spectrum of Legal NLP tasks.

Fifth, as observed in our 'Results' section and based on the analysis of RQ1 within the scope of this SMS, the majority of research in Legal NLP centers on Multiclass Classification. Information Extraction, Question Answering, Information Retrieval, and Summarization are also prevalent subjects. However, Cross-Reference Resolution and Cross-Lingual Transfer remain largely unexplored, primarily because they aren't deemed critical for the more intricate legal processes targeted for automation. Moreover, within the scope of our SMS, we did not encounter articles addressing Discourse Coherence, an NLP task of significant relevance in Legal NLP. Discourse Coherence is vital for Legal NLP since the absence of either local or global coherence in legal documents regardless of their length can jeopardize the accurate understanding of the semantics and regulations outlined

in such documents.

Furthermore, leveraging Discourse Coherence as an initial preprocessing step for legal documents could be a key strategy to alleviate the challenges posed by complex legal language. By adopting this approach, several challenges associated with the intricacy and ambiguity of legal verbiage can be addressed. This includes the apt segmentation and understanding of convoluted sentences, maintaining context within legal documents, and aiding argumentation analysis. Additionally, this tactic can augment Legal NLP tasks by ensuring alignment with a query's intent in Information Retrieval and by assisting in generating summaries that not only highlight the primary themes but also preserve the logical progression of the original content.

A final observation that surprises us is that under the methodology of this SMS, we have yet to read any paper related to topics such as fairness and bias and how to mitigate them in the models of Legal NLP.

In addition to our primary findings, we aim to summarize what we regard as a valuable use case and applicability for each of the NLP branches, as well as their suitability.

Text Classification: This approach has numerous use cases and applications. In essence, it facilitates the categorization of legal documents. These categories can encompass anything deemed relevant within the legal context. For instance, they could relate to different areas of law (e.g., criminal, civil, cybersecurity). Within a specific legal domain, categorization can become even more granular, focusing on aspects like policy types, the quality of writing, and more. As indicated by the results, this method is the most extensively researched and often yields the best outcomes when trained on representative data. However, misclassification can occur, especially when content intersects with multiple categories. Consequently, decisions derived from these methods should invariably be reviewed by an impartial legal expert.

Summarization: The use case for this scenario is straightforward: summarizing extended documents such as court judgments and privacy policies. This aids both legal practitioners and the general public in understanding the crux of voluminous documents. However, as highlighted in the ‘‘Results" section, summarization is not flawless. It can omit pertinent details, introduce contradictions, or even include extraneous information absent in the original document. Thus, while an initial summary provides a useful starting point, it should invariably be reviewed and, if necessary, revised by legal professionals.

Information Extraction: The primary use case is the ability to identify and extract pertinent information such as names, places, dates, main concepts, and their relationships, significantly enhancing document analysis. This facilitates pinpointing specific details within expansive texts, promoting efficient data extraction and analysis. A notable feature of legal documents is their structured nature. This attribute makes such documents particularly amenable to high-performance results using state-of-the-art Information Extraction algorithms. Nonetheless, there remains a risk of misinterpretation, potentially leading to incorrect extractions of entities and their relationships. As always, the final output should not be accepted without question; consultation with legal experts remains essential.

Question Answering: The primary use case is to address legal inquiries. Such algorithms are commonly integrated into legal chatbots. Beyond assisting legal experts, these tools can aid the general public and junior legal practitioners, enabling them to swiftly find answers to legal questions without the need to comb through extensive documents. However, while they prove useful for routine legal queries, they might falter when faced with intricate questions requiring legal reasoning or interpretability. They are best suited for initial guidance rather than in-depth legal counsel.

Information Retrieval: The primary use case involves searching for pertinent case laws, policies, or legal documents. This aids legal practitioners and researchers in swiftly locating relevant information, thereby automating the traditionally manual legal research process. However, given their limitations, these algorithms might overlook nuances or contexts that a human researcher could discern. They are best employed for preliminary research or when seeking general categories of documents.

Coreference Resolution: This can be applied in scenarios such as document analysis, akin to Information Extraction, as it can associate various mentions of a single entity. A prime example is understanding a contract, wherein terms like ‘‘the party" and ‘‘such entities" can be linked to their respective references, such as actual party names or entities mentioned earlier in the document. Moreover, this task is pivotal in enhancing the results and applicability of all the previously discussed tasks. However, legal vernacular often employs specialized terminology and expressions with unique meanings. While Coreference Resolution tools can handle standard references, they might falter when interpreting nuanced or specialized legal references without targeted training. Thus, the assessment by a legal expert remains crucial.

Furthermore, the findings of this study regarding the various tools and techniques developed in Legal NLP offer insights into their potential role in bridging the understanding gap between legal experts and the general public concerning legal language. Addressing this issue is currently vital in the legal domain [P75].

Our recommendations on how advancements in Legal NLP could address this issue are as follows:

Adaptive Tools: Current legal tools and websites could be adapted based on the user's role. For instance, if a legal practitioner is using the tool or website, providing information in exact legal terminology is appropriate. How-

ever, for non-legal practitioners seeking to understand policies or laws, there should be an option to tailor the information to suit those needs. For example, tools could automatically translate legal jargon into plain language, making legal documents more comprehensible for the general public.

Question-Answering: Enhancing language comprehension in the legal domain through Legal NLP could lead to the development of tools that enable the public to pose questions about legal documents and receive straightforward, interpretable answers.

Information Retrieval: When a non-legal expert is searching for a document related to an application they're completing, or if they wish to understand a specific legal aspect, they should be able to phrase their query in everyday language and still access the precise legal documents relevant to their question.

Summarization: When a non-legal expert faces the task of agreeing to or signing a lengthy legal document, they often neglect its contents, as seen with cybersecurity policies [R1, R3, R4]. In such scenarios, an effective summarization tool is crucial. It can distill extensive legal documents into shorter, more digestible versions.

Legal Assistants: The advances in Legal NLP could lead to high-performance and trustworthy chatbots that guide individuals through legal procedures or answer frequently posed legal questions.

Furthermore, after examining the complexity of various approaches, another issue becomes evident: How can the results and insights from NLP research in the legal domain be effectively conveyed to legal professionals, policymakers, and the general public? Based on the comprehensive findings of this study, we offer the following recommendations.

1) Current tools, as well as future ones developed with state-of-the-art advancements in Legal NLP, need comprehensive documentation. They should also be equipped with accessible and user-friendly manuals and guides that detail the application and benefits of such tools.
2) User manuals should include examples and cater to diverse audiences. The presentation of information should vary depending on the reader, whether they are a policymaker, a member of the general public, or a legal professional with a specific task.
3) We believe that mere documentation and tutorials won't suffice. If an organization plans to use a Legal-NLP based tool with policymakers or legal professionals, conducting workshops should be mandatory. This approach will enable users to familiarize themselves with the tool and test it in real-life scenarios.
4) For the general public, the equivalent of workshops could be public lectures and webinars. We believe these would be beneficial, especially if they include demonstrations of various real-life scenarios.
5) Another vital consideration is that irrespective of the method employed to communicate these advancements, it's crucial to consistently highlight the limitations and risks associated with use. When feasible, providing a confidence percentage can be beneficial. It's imperative to address all ethical issues, limitations, and the potential pitfalls of over-dependence on the tool. Based on our study's current findings, any tool should be viewed as supplementary for specific phases, but it shouldn't be used for sensitive or decision centric matters.

VI. THREAT TO VALIDITY

Every SMS must include a validation process. Wohlin et al.[R49] defined four types of validity threats: construct, internal, external, and conclusion. We discuss each one of these threats in this section.

A. CONSTRUCT VALIDITY THREATS

Following the guidelines of [R49], construct validity threats are the threats linked to the generalization of the results to the concepts behind the study. To minimize this threat, more than one researcher rechecked the information provided in the paper and searched for references to code and datasets that might not appear in the original write-up. During our SMS, we could find some papers that needed more detailed information about their approach or neural architecture. Furthermore, many articles didn't provide a source code that could be used for reproducibility or checking the results. In such cases, we conclude with our joint assessments.

However, in a strict sense, our findings are valid only for our sample of selected primary studies, which were accessible by ACM, IEEE, Scopus, and Elsevier and depended on our search query. To try to mitigate the construct threats associated with the search query, we tried 12 different search queries by considering the concepts and acronyms without four indexers and saw how relevant the results were to our objectives. Our final search query was broad enough to enclose the state-of-the-art and critical primary studies from 2015 to 2022 in multiple NLP areas like Multiclass Classification, Information Extraction, Information Retrieval, Coreference Resolution, Summarization, Cross-Lingual Transfer, and Language Modeling.

In addition, another threat to construct validity was the inclusion of publications based on inclusion/exclusion criteria, which were defined before the study was conducted. During phase 1, only the title, keywords, and abstract were examined, and there was a possibility of excluding some relevant primary studies during this process. To mitigate this problem, we took two actions. First, whenever we were unsure whether a publication should be excluded, we opted for temporary inclusion. In some cases, we even did phase 4 in papers we needed clarification, and then we had a joint discussion to decide if we should keep them. However, even doing this process, we might have excluded a relevant publication. Second, we perform a one-level lightweight forward and backward snowballing on the included papers [R15] to find papers we might have missed due to the search query selected.

B. INTERNAL VALIDITY THREATS

Internal threats to validity are related to problems that might arise during the data extraction process. Extracting data is a complex process, which is more prominent if the information the study intends to extract is low or buried within the paper. Phase 4 in our SMS involves the extraction of relevant information is one of the more error-prone phases when performing an SMS because of the absence of standard terminology, missing information, or not presented in the paper. To mitigate this internal validity threat, the extracted data from the primary study were classified by the NLP experts. However, since this might introduce bias, the decision was only taken with research. To be more consistent, we explored related studies and even other types of studies like surveys [R9] and empirical [P1] to see what were their classification decision mostly aligns with ours.

Another internal threat to validity is possible errors and bias in our qualitative analysis since the quantitative is based on descriptive statistics and is less likely to contain errors. However, it was carefully reviewed. To mitigate this threat to the qualitative aspect of our SMS, we not only carefully reviewed it by more than one researcher, but also we made a great effort to be as faithful as possible to present the summary of the literal analysis inside the primary studies, without adding bias and opinions in the sections of our results.

C. EXTERNAL VALIDITY THREATS

Most of the conclusions in this SMS are focused on Legal NLP and cannot be generalized to other research topics. However, some of the conclusions, especially in the limitations and challenges, are also present in NLP in general [R50]. Still, in this study, we didn't make such general conclusions. Furthermore, we presented the results of multiple primary studies and did not validate them, implying that we also carried the same threats they had. The external validity threats are related to the generalization of results of the review to real-world scenarios [R49].

D. CONCLUSION VALIDITY THREATS

Threats related to inaccurate conclusions based on our findings and SMS and whether the SMS can be repeated enter this category. Among the factors involved in wrong conclusions, we can find incorrect data extraction or identify wrong relationships. A factor that helps us mitigate this threat is that we include more primary studies than any other SLR we have found to the best of our knowledge. To mitigate incorrect data extraction or wrong relations, we made the NLP experts double-check the final extraction from every researcher in this SMS. If something seems wrong or missing, the NLP expert will redo it with the other researcher. We realized this was a good decision since more than 10 of the primary studies included needed to include information or had a wrong extracted method or conclusion. Furthermore, with the measures we took, the absence of a wrongly classified primary study would not skew the presented descriptive statistics or conclusions supported by more than one paper. Hence, the probability of wrong conclusions is smaller. In addition, this study can be replicated by following our detailed methodology. Also, the raw data and protocol execution is available at [55].

VII. CONCLUDING REMARKS

In this Systematic Mapping Study, we reviewed a vast body of literature to ensure our findings comprehensively represented the progress in Legal NLP. Using a detailed query across four reputable indexers, we filtered an initial collection of 536 papers to 75 articles. This selection was based on research paper inclusion criteria, adhering to the best practices for Systematic Mapping Studies. We summarized the diverse approaches employed by the research community to enhance outcomes in various Legal NLP tasks, including Multiclass Classification, Language Modeling, Summarization, Information Extraction, Question Answering, Information Retrieval, and Coreference Resolution.

Additionally, we summarized the current resources, websites, and ontologies utilized to train new Machine Learning Models. We also outlined the existing limitations and gaps in the Legal NLP field. In conclusion, we offered an in-depth discussion on the primary findings of this SMS, highlighting its implications for Legal NLP researchers, legal practitioners, and the general public.

The results found on this SMS leave clear that NLP is already at the core of the legal sector, regardless of whether it's genuinely sufficient to obtain full automation of the tedious legal processes. It is evident the advantages that improvements in the Legal NLP field can bring. We are talking of complex tasks like reviewing long legal documents, retrieving related legal documents, and reviewing contracts and privacy policies, to mention a few, where a legal practitioner can spend a long time. Instead, a Legal NLP model could achieve it in minutes.

However, not everything is great; some challenges remain and gaps, as described in the results of our third research question. Several of today's models still need to be improved in processing long or language-complex legal documents. In addition, with the rise of PLMs to approach NLP tasks, including in the legal domain, there is the limitation of general PLMs not being enough to get good results and, on the other hand, domain-specific PLMs not being able to transfer to other legal subdomains too disparate from the one they were trained. Furthermore, the lack of resources like curated datasets and ontologies increases the difficulty of solving these problems. Also, different from several other NLP domains, the lack of unified legal procedures and the privacy inherent in the legal domain makes it difficult even to get new data.

More questions remain a challenge in Legal NLP, not only the performance of the models but also the ethical implications. Digitizing legal data as input for an NLP model could expose critical and private data. In addition, there are

[55] https://zenodo.org/record/7626621#.Y-VVM3bMJPY

bias issues to consider, like gender or race discrimination in delicate applications like judgment prediction. Furthermore, blindly depending on the recent and future Legal NLP tools may result in making high-impact mistakes or overlooking critical information.

Our findings present implications useful for Legal NLP researchers to deal with these challenges. Furthermore, our study shows which Legal NLP areas are more advanced in research and results, making them good candidates to start developing tools that might be useful in today's context if they consider the current limitations. From our NLP experience, several of these problems are neither new nor unique in the legal domain; some are general problems of the NLP field in general [R50]. We believe that the main effort from Legal NLP researchers should move to inherent difficulties in their field, like the lack of curated and accessible data. More curated datasets must be created, more unified data from multiple countries should be accessible, and more symbol methods that use the knowledge and rules already present in the legal domain should be included. Legal NLP is not far from a good automation that would be significantly helpful if we continue moving forward in these directions.

PRIMARY STUDIES

[P1] Dezhao Song, Sally Gao, Baosheng He, and Frank Schilder. ''On the Effectiveness of Pre-Trained Language Models for Legal Natural Language Processing: An Empirical Study''. In: *IEEE Access* 10 (2022), pp. 75835–75858 (cit. on pp. 6, 7, 11, 12, 17, 19, 21, 22, 25, 33, 34).

[P2] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. ''LEGAL-BERT: The muppets straight out of law school''. In: *arXiv preprint arXiv:2010.02559* (2020) (cit. on pp. 7, 33).

[P3] Weijing Huang, Xianfeng Liao, Zhiqiang Xie, Jiang Qian, Bojin Zhuang, Shaojun Wang, and Jing Xiao. ''Generating reasonable legal text through the combination of language modeling and question answering''. In: *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*. 2021, pp. 3687–3693 (cit. on pp. 7, 33).

[P4] Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. ''When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings''. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 2021, pp. 159–168 (cit. on pp. 7, 8, 12, 19, 20).

[P5] Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. ''Lawformer: A pre-trained language model for chinese legal long documents''. In: *AI Open* 2 (2021), pp. 79–84 (cit. on pp. 7, 8, 19, 33).

[P6] Ruyu Qin, Min Huang, and Yutong Luo. ''A Comparison Study of Pre-trained Language Models for Chinese Legal Document Classification''. In: *2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD)*. IEEE. 2022, pp. 444–449 (cit. on pp. 7, 8, 21, 22, 33).

[P7] Samir Undavia, Adam Meyers, and John E Ortega. ''A comparative study of classifying legal documents with neural networks''. In: *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*. IEEE. 2018, pp. 515–522 (cit. on pp. 7, 8, 19, 33, 34).

[P8] Rajaa El Hamdani, Majd Mustapha, David Restrepo Amariles, Aurore Troussel, Sébastien Meeùs, and Katsiaryna Krasnashchok. ''A combined rule-based and machine learning approach for automated GDPR compliance checking''. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 2021, pp. 40–49 (cit. on pp. 7, 10, 33, 34).

[P9] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. ''Neural legal judgment prediction in English''. In: *arXiv preprint arXiv:1906.02059* (2019) (cit. on pp. 7, 9, 19, 33, 34).

[P10] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. ''Large-scale multi-label text classification on EU legislation''. In: *arXiv preprint arXiv:1906.02192* (2019) (cit. on pp. 7, 9, 19, 33, 34).

[P11] V Gokul Pillai and Lekshmi R Chandran. ''Verdict Prediction for Indian Courts Using Bag of Words and Convolutional Neural Network''. In: *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*. IEEE. 2020, pp. 676–683 (cit. on pp. 7, 9, 20, 22, 33, 34).

[P12] Dezhao Song, Andrew Vold, Kanika Madan, and Frank Schilder. ''Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training''. In: *Information Systems* 106 (2022), p. 101718 (cit. on pp. 7, 10, 12, 19, 21, 22, 33, 34).

[P13] Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Martin Katz, and Nikolaos Aletras. ''Lexglue: A benchmark dataset for legal language understanding in English''. In: *arXiv preprint arXiv:2110.00976* (2021) (cit. on pp. 7, 19–21, 33).

[P14] Francisco de Arriba-Pérez, Silvia García-Méndez, Francisco J González-Castaño, and Jaime González-González. ''Explainable machine learning multi-label classification of Spanish legal judgements''. In: *Journal of King Saud University-Computer and Information Sciences* 34.10 (2022), pp. 10180–10192 (cit. on pp. 7, 10, 33, 34).

[P15] Marco Lippi, Przemysław Pałka, Giuseppe Contissa, Francesca Lagioia, Hans-Wolfgang Micklitz, Gio-

vanni Sartor, and Paolo Torroni. ''CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service''. In: *Artificial Intelligence and Law* 27 (2019), pp. 117–139 (cit. on pp. 7, 12, 19, 33).

[P16] Mariana Y Noguti, Eduardo Vellasques, and Luiz S Oliveira. ''Legal document classification: An application to law area prediction of petitions to public prosecution service''. In: *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2020, pp. 1–8 (cit. on pp. 7, 10, 33, 34).

[P17] Shounak Paul, Pawan Goyal, and Saptarshi Ghosh. ''LeSICiN: A heterogeneous graph-based approach for automatic legal statute identification from Indian legal documents''. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 36. 10. 2022, pp. 11139–11146 (cit. on pp. 7, 11, 19, 33, 34).

[P18] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. ''Legal judgment prediction via topological learning''. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2018, pp. 3540–3549 (cit. on pp. 7, 9, 19, 33, 34).

[P19] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. ''The creation and analysis of a website privacy policy corpus''. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016, pp. 1330–1340 (cit. on pp. 7, 9, 10, 18, 19).

[P20] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. ''Polisis: Automated analysis and presentation of privacy policies using deep learning''. In: *27th USENIX Security Symposium (USENIX Security 18)*. 2018, pp. 531–548 (cit. on pp. 7, 9, 10, 33, 34).

[P21] Bingfeng Luo, Yansong Feng, Jianbo Xu, Xiang Zhang, and Dongyan Zhao. ''Learning to predict charges for criminal cases with legal basis''. In: *arXiv preprint arXiv:1707.09168* (2017) (cit. on pp. 7, 8, 33, 34).

[P22] Sahan Jayasinghe, Lakith Rambukkanage, Ashan Silva, Nisansa de Silva, and Amal Shehan Perera. ''Critical sentence identification in legal cases using multi-class classification''. In: *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*. IEEE. 2021, pp. 146–151 (cit. on pp. 7, 10, 19, 33, 34).

[P23] Onur Akça, Gıyaseddin Bayrak, Abdul Majeed Issifu, and Murat Can Ganiz. ''Traditional Machine Learning and Deep Learning-based Text Classification for Turkish Law Documents using Transformers and Domain Adaptation''. In: *2022 International Conference on INnovations in Intelligent Systems and Applications (INISTA)*. IEEE. 2022, pp. 1–6 (cit. on pp. 7, 10, 21, 33, 34).

[P24] Xiaofan Fang and Xianghao Zhao. ''Nonlinear Dimensionality Reduction with Judicial Document Learning''. In: *2018 IEEE International Conference on Big Knowledge (ICBK)*. IEEE. 2018, pp. 448–455 (cit. on pp. 7, 9, 19, 20, 22, 34).

[P25] Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. ''Casesummarizer: a system for automated summarization of legal texts''. In: *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*. 2016, pp. 258–262 (cit. on pp. 7, 11, 34).

[P26] Kaiz Merchant and Yash Pande. ''Nlp based latent semantic analysis for legal text summarization''. In: *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE. 2018, pp. 1803–1807 (cit. on pp. 7, 11, 33, 34).

[P27] Anastassia Kornilova and Vlad Eidelman. ''BillSum: A corpus for automatic summarization of US legislation''. In: *arXiv preprint arXiv:1910.00523* (2019) (cit. on pp. 7, 12, 18, 19, 33).

[P28] Xinyu Duan, Yating Zhang, Lin Yuan, Xin Zhou, Xiaozhong Liu, Tianyi Wang, Ruocheng Wang, Qiong Zhang, Changlong Sun, and Fei Wu. ''Legal summarization for multi-role debate dialogue via controversy focus mining and multi-task learning''. In: *Proceedings of the 28th ACM international conference on information and knowledge management*. 2019, pp. 1361–1370 (cit. on pp. 7, 11, 19, 33, 34).

[P29] Vu Tran, Minh Le Nguyen, and Ken Satoh. ''Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model''. In: *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law*. 2019, pp. 275–282 (cit. on pp. 7, 11, 33, 34).

[P30] Vu Tran, Minh Le Nguyen, Satoshi Tojo, and Ken Satoh. ''Encoded summarization: summarizing documents into continuous vector space for legal case retrieval''. In: *Artificial Intelligence and Law* 28 (2020), pp. 441–467 (cit. on pp. 7, 11, 22, 33, 34).

[P31] Amy JC Trappey, Charles V Trappey, Jheng-Long Wu, and Jack WC Wang. ''Intelligent compilation of patent summaries using machine learning and natural language processing techniques''. In: *Advanced Engineering Informatics* 43 (2020), p. 101027 (cit. on pp. 7, 12, 33, 34).

[P32] Amin Sleimi, Nicolas Sannier, Mehrdad Sabetzadeh, Lionel Briand, and John Dann. ''Automated extraction of semantic legal metadata using natural language processing''. In: *2018 IEEE 26th International Requirements Engineering Conference (RE)*. IEEE. 2018, pp. 124–135 (cit. on pp. 7, 13, 34).

[P33] Ni Zhang, Yi-Fei Pu, Sui-Quan Yang, Ji-Liu Zhou, and Jin-Kang Gao. ''An ontological Chinese legal

consultation system''. In: *IEEE Access* 5 (2017), pp. 18250–18261 (cit. on pp. 7, 13, 34).

[P34] Manar Alohaly and Hassan Takabi. ''If You Can't Measure It, You Can't Manage It: Towards Quantification of Privacy Policies''. In: *2016 IEEE 2nd International Conference on Collaboration and Internet Computing (CIC)*. IEEE. 2016, pp. 539–545 (cit. on pp. 7, 13, 34).

[P35] Georgia M Kapitsaki and Demetris Paschalides. ''Identifying terms in open source software license texts''. In: *2017 24th Asia-Pacific Software Engineering Conference (APSEC)*. IEEE. 2017, pp. 540–545 (cit. on pp. 7, 13, 19, 34).

[P36] Matías García-Constantino, Katie Atkinson, Danushka Bollegala, Karl Chapman, Frans Coenen, Claire Roberts, and Katy Robson. ''CLIEL: context-based information extraction from commercial law documents''. In: *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law*. 2017, pp. 79–87 (cit. on pp. 7, 13, 34).

[P37] Masaharu Yoshioka, Yasuhiro Aoki, and Youta Suzuki. ''Bert-based ensemble methods with data augmentation for legal textual entailment in coliee statute law task''. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 2021, pp. 278–284 (cit. on pp. 7, 14, 19, 21, 22, 33, 34).

[P38] Mauro Dragoni, Serena Villata, Williams Rizzi, and Guido Governatori. ''Combining NLP approaches for rule extraction from legal documents''. In: *1st Workshop on MIning and REasoning with Legal texts (MIREL 2016)*. 2016 (cit. on pp. 7, 12, 34).

[P39] Jidong Ge, Yunyun Huang, Xiaoyu Shen, Chuanyi Li, and Wei Hu. ''Learning fine-grained fact-article correspondence in legal cases''. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 3694–3706 (cit. on pp. 7, 14, 19, 22, 34).

[P40] Donghong Ji, Peng Tao, Hao Fei, and Yafeng Ren. ''An end-to-end joint model for evidence information extraction from court record document''. In: *Information Processing & Management* 57.6 (2020), p. 102305 (cit. on pp. 7, 14, 20, 22, 34).

[P41] Ilias Chalkidis, Ion Androutsopoulos, and Achilleas Michos. ''Extracting contract elements''. In: *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*. 2017, pp. 19–28 (cit. on pp. 7, 13, 19, 34).

[P42] Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. ''Legal document retrieval using document vector embeddings and deep learning''. In: *Science and information conference*. Springer. 2018, pp. 160–175 (cit. on pp. 7, 15, 34).

[P43] Jörg Landthaler, Bernhard Waltl, Patrick Holl, and Florian Matthes. ''Extending Full Text Search for Legal Document Collections Using Word Embeddings.'' In: *JURIX*. 2016, pp. 73–82 (cit. on pp. 7, 14, 33, 34).

[P44] Phi Manh Kien, Ha-Thanh Nguyen, Ngo Xuan Bach, Vu Tran, Minh Le Nguyen, and Tu Minh Phuong. ''Answering legal questions by learning neural attentive text representation''. In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2020, pp. 988–998 (cit. on pp. 7, 16, 34).

[P45] Isar Nejadgholi, Renaud Bougueng, and Samuel Witherspoon. ''A Semi-Supervised Training Method for Semantic Search of Legal Facts in Canadian Immigration Cases.'' In: *JURIX*. 2017, pp. 125–134 (cit. on pp. 7, 15, 33, 34).

[P46] Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. ''Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension''. In: *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*. Springer. 2019, pp. 439–451 (cit. on pp. 7, 8, 18, 19, 33).

[P47] Sallam Abualhaija, Chetan Arora, Amin Sleimi, and Lionel C Briand. ''Automated question answering for improved understanding of compliance requirements: A multi-document study''. In: *2022 IEEE 30th International Requirements Engineering Conference (RE)*. IEEE. 2022, pp. 39–50 (cit. on pp. 7, 17, 33, 34).

[P48] Daniel Locke, Guido Zuccon, and Harrisen Scells. ''Automatic query generation from legal texts for case law retrieval''. In: *Information Retrieval Technology: 13th Asia Information Retrieval Societies Conference, AIRS 2017, Jeju Island, South Korea, November 22-24, 2017, Proceedings 13*. Springer. 2017, pp. 181–193 (cit. on pp. 7, 15, 19, 34).

[P49] Yixiao Ma, Yunqiu Shao, Yueyue Wu, Yiqun Liu, Ruizhe Zhang, Min Zhang, and Shaoping Ma. ''LeCaRD: a legal case retrieval dataset for Chinese law system''. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2021, pp. 2342–2348 (cit. on pp. 7, 8, 16, 19, 21, 34).

[P50] Adebayo Kolawole John, Luigi Di Caro, and Guido Boella. ''Solving Bar Exam Questions with Deep Neural Networks''. In: *Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Texts: co-located with the 16th International Conference on Artificial Intelligence and Law*. 2017 (cit. on pp. 7, 15, 34).

[P51] Andrew Vold and Jack G Conrad. ''Using transformers to improve answer retrieval for legal questions''. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*. 2021, pp. 245–249 (cit. on pp. 7, 16, 33, 34).

[P52] Truong-Thinh Tieu, Chieu-Nguyen Chau, Truong-Son Nguyen, Le-Minh Nguyen, et al. ''Apply Bert-based models and Domain knowledge for Automated Legal Question Answering tasks at ALQAC 2021''. In: *2021 13th International Conference on Knowledge and Systems Engineering (KSE)*. IEEE. 2021, pp. 1–6 (cit. on pp. 7, 16, 19, 33, 34).

[P53] Ahmed Elnaggar, Robin Otto, and Florian Matthes. ''Deep learning for named-entity linking with transfer learning for legal documents''. In: *Proceedings of the 2018 artificial intelligence and cloud computing conference*. 2018, pp. 23–28 (cit. on pp. 7, 15, 33, 34).

[P54] Phong-Khac Do, Huy-Tien Nguyen, Chien-Xuan Tran, Minh-Tien Nguyen, and Minh-Le Nguyen. ''Legal question answering using ranking SVM and deep convolutional neural network''. In: *arXiv preprint arXiv:1703.05320* (2017) (cit. on pp. 7, 15, 19, 33, 34).

[P55] Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. ''BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval.'' In: *IJCAI*. 2020, pp. 3501–3507 (cit. on pp. 7, 16, 20, 22, 33, 34).

[P56] Juliano Rabelo, Mi-Young Kim, and Randy Goebel. ''Combining similarity and transformer methods for case law entailment''. In: *Proceedings of the seventeenth international conference on artificial intelligence and law*. 2019, pp. 290–296 (cit. on pp. 7, 16, 19, 33, 34).

[P57] Ilias Chalkidis, Manos Fergadiotis, Nikolaos Manginas, Eva Katakalou, and Prodromos Malakasiotis. ''Regulatory compliance through Doc2Doc information retrieval: A case study in EU/UK legislation where text similarity has limitations''. In: *arXiv preprint arXiv:2101.10726* (2021) (cit. on pp. 7, 16, 33, 34).

[P58] Nicolas Sannier, Morayo Adedjouma, Mehrdad Sabetzadeh, and Lionel Briand. ''An automated framework for detection and resolution of cross references in legal texts''. In: *Requirements Engineering* 22.2 (2017), pp. 215–237 (cit. on pp. 7, 17, 19, 34).

[P59] Surawat Pothong and Nuttanart Facundes. ''Coreference Resolution and Meaning Representation in a Legislative Corpus''. In: *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*. IEEE. 2021, pp. 1–6 (cit. on pp. 7, 17, 19, 33, 34).

[P60] Donghong Ji, Jun Gao, Hao Fei, Chong Teng, and Yafeng Ren. ''A deep neural network model for speakers coreference resolution in legal texts''. In: *Information Processing & Management* 57.6 (2020), p. 102365 (cit. on pp. 7, 17, 33, 34).

[P61] Zein Shaheen, Gerhard Wohlgenannt, and Dmitry Mouromtsev. ''Zero-Shot Cross-Lingual Transfer in Legal Domain Using Transformer Models''. In: *2021 International Conference on Computational Science and Computational Intelligence (CSCI)*. IEEE. 2021, pp. 450–456 (cit. on pp. 7, 18, 21, 22, 33).

[P62] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. ''Jec-qa: A legal-domain question answering dataset''. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 9701–9708 (cit. on pp. 8, 19).

[P63] Olga Shulayeva, Advaith Siddharthan, and Adam Wyner. ''Recognizing cited facts and principles in legal judgements''. In: *Artificial Intelligence and Law* 25.1 (2017), pp. 107–126 (cit. on pp. 8, 34).

[P64] Chanika Ruchini Mudalige, Dilini Karunarathna, Isanka Rajapaksha, Nisansa de Silva, Gathika Ratnayaka, Amal Shehan Perera, and Ramesh Pathirana. ''SigmaLaw-ABSA: dataset for aspect-based sentiment analysis in legal opinion texts''. In: *2020 IEEE 15th international conference on industrial and information systems (ICIIS)*. IEEE. 2020, pp. 488–493 (cit. on pp. 10, 20).

[P65] Yougang Lyu, Zihan Wang, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Xiaozhong Liu, Yujun Li, Hongsong Li, and Hongye Song. ''Improving legal judgment prediction through reinforced criminal element extraction''. In: *Information Processing & Management* 59.1 (2022), p. 102780 (cit. on pp. 11, 21, 22, 33, 34).

[P66] Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. ''Towards explainable AI: Assessing the usefulness and impact of added explainability features in legal document summarization''. In: *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–7 (cit. on pp. 12, 33, 34).

[P67] Deepa Anand and Rupali Wagh. ''Effective deep learning approaches for summarization of legal texts''. In: *Journal of King Saud University-Computer and Information Sciences* (2019) (cit. on pp. 12, 19, 33, 34).

[P68] Tomasz Strąk and Michał Tuszyński. ''NLP Based Retrieval of Semantically Similar Private Tax Rulings''. In: *Procedia Computer Science* 207 (2022), pp. 2853–2864 (cit. on pp. 17, 33, 34).

[P69] Razieh Nokhbeh Zaeem and K Suzanne Barber. ''A large publicly available corpus of website privacy policies based on dmoz''. In: *Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy*. 2021, pp. 143–148 (cit. on pp. 19, 20).

[P70] Adam Zachary Wyner, Biralatei James Fawei, and Jeff Z Pan. ''Passing a USA national bar exam: a first corpus for experimentation''. In: *LREC 2016, Tenth International Conference on Language Resources and Evaluation*. LREC. 2016 (cit. on pp. 18, 19).

[P71] Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B Norton, N Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. ''PrivOnto: A semantic framework for the analysis of privacy policies''. In: *Semantic Web* 9.2 (2018), pp. 185–203 (cit. on p. 18).

[P72] Laura Manor and Junyi Jessy Li. ''Plain English summarization of contracts''. In: *arXiv preprint arXiv:1906.00424* (2019) (cit. on p. 18).

[P73] Joel Niklaus, Ilias Chalkidis, and Matthias Stürmer. ''Swiss-judgment-prediction: A multilingual legal judgment prediction benchmark''. In: *arXiv preprint arXiv:2110.00806* (2021) (cit. on pp. 20, 33).

[P74] Inari Listenmaa, Maryam Hanafiah, Regina Cheong, and Andreas KALLBERG. ''Towards CNL-based verbalization of computational contracts''. In: ACL. 2021 (cit. on p. 20).

[P75] Haoxi Zhong, Chaojun Xiao, Cunchao Tu, Tianyang Zhang, Zhiyuan Liu, and Maosong Sun. ''How does NLP benefit legal system: A summary of legal artificial intelligence''. In: *arXiv preprint arXiv:2004.12158* (2020) (cit. on pp. 20, 22, 23).

OTHER BIBLIOGRAPHY

[R1] Alessandro Acquisti and Jens Grossklags. ''Privacy and rationality in individual decision making''. In: *IEEE security & privacy* 3.1 (2005), pp. 26–33 (cit. on pp. 2, 24).

[R2] Nili Steinfeld. ''"I agree to the terms and conditions":(How) do users read privacy policies online? An eye-tracking experiment''. In: *Computers in human behavior* 55 (2016), pp. 992–1000 (cit. on p. 2).

[R3] Burcu Bulgurcu, Hasan Cavusoglu, and Izak Benbasat. ''Information security policy compliance: an empirical study of rationality-based beliefs and information security awareness''. In: *MIS quarterly* (2010), pp. 523–548 (cit. on pp. 2, 24).

[R4] Aggeliki Tsohou, Maria Karyda, Spyros Kokolakis, and Evangelos Kiountouzis. ''Managing the introduction of information security awareness programmes in organisations''. In: *European Journal of Information Systems* 24.1 (2015), pp. 38–58 (cit. on pp. 2, 24).

[R5] John B Ruhl and Daniel Martin Katz. ''Measuring, monitoring, and managing legal complexity''. In: *Iowa L. Rev.* 101 (2015), p. 191 (cit. on p. 2).

[R6] Roland Friedrich. ''Complexity and entropy in legal language''. In: *Frontiers in Physics* 9 (2021), p. 671882 (cit. on p. 2).

[R7] Barbara A Kitchenham, David Budgen, and O Pearl Brereton. ''Using mapping studies as the basis for further research–a participant-observer case study''. In: *Information and Software Technology* 53.6 (2011), pp. 638–651 (cit. on pp. 2–5).

[R8] Ilias Chalkidis and Dimitrios Kampas. ''Deep learning in law: early adaptation and legal word embeddings trained on large corpora''. In: *Artificial Intelligence and Law* 27.2 (2019), pp. 171–198 (cit. on pp. 2, 3).

[R9] Alfredo Montelongo and João Luiz Becker. ''Tasks performed in the legal domain through Deep Learning: A bibliometric review (1987–2020)''. In: *2020 International Conference on Data Mining Workshops (ICDMW)*. IEEE. 2020, pp. 775–781 (cit. on pp. 2, 3, 6, 25).

[R10] Reshma Sheik and S Jaya Nirmala. ''Deep Learning Techniques for Legal Text Summarization''. In: *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*. IEEE. 2021, pp. 1–5 (cit. on pp. 2, 3, 7).

[R11] Nur Aqilah Khadijah Rosili, Noor Hidayah Zakaria, Rohayanti Hassan, Shahreen Kasim, Farid Zamani Che Rose, and Tole Sutikno. ''A systematic literature review of machine learning methods in predicting court decisions''. In: *IAES International Journal of Artificial Intelligence* 10.4 (2021), p. 1091 (cit. on p. 3).

[R12] Carlo Sansone and Giancarlo Sperli. ''Legal Information Retrieval systems: State-of-the-art and open issues''. In: *Information Systems* 106 (2022), p. 101967 (cit. on p. 3).

[R13] Michele Soavi, Nicola Zeni, John Mylopoulos, and Luisa Mich. ''From Legal Contracts to Formal Specifications: A Systematic Literature Review''. In: *SN Computer Science* 3.5 (2022), p. 345 (cit. on p. 3).

[R14] Madhukar Pai, Michael McCulloch, Jennifer D Gorman, Nitika Pai, Wayne Enanoria, Gail Kennedy, Prathap Tharyan, and John M Colford Jr. ''Systematic reviews and meta-analyses: an illustrated, step-by-step guide.'' In: *The National medical journal of India* 17.2 (2004), pp. 86–95 (cit. on p. 4).

[R15] Claes Wohlin. ''Guidelines for snowballing in systematic literature studies and a replication in software engineering''. In: *Proceedings of the 18th international conference on evaluation and assessment in software engineering*. 2014, pp. 1–10 (cit. on pp. 5, 24).

[R16] Helvio Jeronimo Junior and Guilherme Horta Travassos. ''Consolidating a Common Perspective on Technical Debt and its Management Through a Tertiary Study''. In: *Information and Software Technology* (2022), p. 106964 (cit. on p. 5).

[R17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. ''Bert: Pre-training of deep bidirectional transformers for language understanding''. In: *arXiv preprint arXiv:1810.04805* (2018) (cit. on p. 7).

[R18] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. ''Xl-

net: Generalized autoregressive pretraining for language understanding''. In: *Advances in neural information processing systems* 32 (2019) (cit. on pp. 7, 10).

[R19] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. ''Improving language understanding by generative pre-training''. In: (2018) (cit. on p. 7).

[R20] Jason Weston, Sumit Chopra, and Antoine Bordes. ''Memory networks''. In: *arXiv preprint arXiv:1410.3916* (2014) (cit. on p. 7).

[R21] Iz Beltagy, Matthew E Peters, and Arman Cohan. ''Longformer: The long-document transformer''. In: *arXiv preprint arXiv:2004.05150* (2020) (cit. on pp. 8, 20).

[R22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. ''Efficient estimation of word representations in vector space''. In: *arXiv preprint arXiv:1301.3781* (2013) (cit. on pp. 8, 9).

[R23] Nikolaos Aletras, Dimitrios Tsarapatsanis, Daniel Preoţiuc-Pietro, and Vasileios Lampos. ''Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective''. In: *PeerJ Computer Science* 2 (2016), e93 (cit. on p. 9).

[R24] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. ''Show, attend and tell: Neural image caption generation with visual attention''. In: *International conference on machine learning*. PMLR. 2015, pp. 2048–2057 (cit. on p. 9).

[R25] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. ''Hierarchical attention networks for document classification''. In: *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2016, pp. 1480–1489 (cit. on pp. 9, 11).

[R26] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. ''Explainable prediction of medical codes from clinical text''. In: *arXiv preprint arXiv:1802.05695* (2018) (cit. on p. 9).

[R27] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. ''mT5: A massively multilingual pre-trained text-to-text transformer''. In: *arXiv preprint arXiv:2010.11934* (2020) (cit. on p. 10).

[R28] Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. ''Big bird: Transformers for longer sequences''. In: *Advances in neural information processing systems* 33 (2020), pp. 17283–17297 (cit. on pp. 12, 20).

[R29] Ting Jiang, Deqing Wang, Leilei Sun, Huayi Yang, Zhengyang Zhao, and Fuzhen Zhuang. ''Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification''. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 9. 2021, pp. 7987–7994 (cit. on p. 12).

[R30] Quoc Le and Tomas Mikolov. ''Distributed representations of sentences and documents''. In: *International conference on machine learning*. PMLR. 2014, pp. 1188–1196 (cit. on pp. 11, 13).

[R31] Chin-Yew Lin. ''Rouge: A package for automatic evaluation of summaries''. In: *Text summarization branches out*. 2004, pp. 74–81 (cit. on p. 11).

[R32] Xinyu Hua, Ashwin Sreevatsa, and Lu Wang. ''DYPLOC: Dynamic planning of content using mixed language models for text generation''. In: *arXiv preprint arXiv:2106.00791* (2021) (cit. on p. 12).

[R33] Ye Ma, Zixun Lan, Lu Zong, and Kaizhu Huang. ''Global-aware beam search for neural abstractive summarization''. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 16545–16557 (cit. on p. 12).

[R34] Rada Mihalcea and Paul Tarau. ''Textrank: Bringing order into text''. In: *Proceedings of the 2004 conference on empirical methods in natural language processing*. 2004, pp. 404–411 (cit. on p. 12).

[R35] James R Curran, Stephen Clark, and Johan Bos. ''Linguistically motivated large-scale NLP with C&C and Boxer''. In: *Proceedings of the 45th annual meeting of the Association for Computational Linguistics Companion volume proceedings of the demo and poster sessions*. 2007, pp. 33–36 (cit. on p. 13).

[R36] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. ''The Stanford CoreNLP natural language processing toolkit''. In: *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*. 2014, pp. 55–60 (cit. on p. 13).

[R37] David M Blei, Andrew Y Ng, and Michael I Jordan. ''Latent dirichlet allocation''. In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022 (cit. on p. 13).

[R38] John Lafferty, Andrew McCallum, and Fernando CN Pereira. ''Conditional random fields: Probabilistic models for segmenting and labeling sequence data''. In: (2001) (cit. on p. 14).

[R39] Octavian-Eugen Ganea and Thomas Hofmann. ''Deep joint entity disambiguation with local neural attention''. In: *arXiv preprint arXiv:1704.04920* (2017) (cit. on p. 15).

[R40] Andre Martins and Ramon Astudillo. ''From softmax to sparsemax: A sparse model of attention and multi-label classification''. In: *International conference on*

machine learning. PMLR. 2016, pp. 1614–1623 (cit. on p. 16).

[R41] Georgios-Ioannis Brokos, Prodromos Malakasiotis, and Ion Androutsopoulos. ''Using centroids of word embeddings and word mover's distance for biomedical document retrieval in question answering''. In: *arXiv preprint arXiv:1608.03905* (2016) (cit. on p. 16).

[R42] Nils Reimers and Iryna Gurevych. ''Sentence-bert: Sentence embeddings using siamese bert-networks''. In: *arXiv preprint arXiv:1908.10084* (2019) (cit. on p. 16).

[R43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. ''Roberta: A robustly optimized bert pretraining approach''. In: *arXiv preprint arXiv:1907.11692* (2019) (cit. on p. 16).

[R44] Jay M Ponte and W Bruce Croft. ''A language modeling approach to information retrieval''. In: *ACM SIGIR Forum*. Vol. 51. 2. ACM New York, NY, USA. 2017, pp. 202–208 (cit. on p. 17).

[R45] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. ''DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter''. In: *arXiv preprint arXiv:1910.01108* (2019) (cit. on p. 18).

[R46] Paul Over, Hoa Dang, and Donna Harman. ''DUC in context''. In: *Information Processing & Management* 43.6 (2007), pp. 1506–1520 (cit. on p. 18).

[R47] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. ''Bidirectional attention flow for machine comprehension''. In: *arXiv preprint arXiv:1611.01603* (2016) (cit. on p. 19).

[R48] Pengcheng He, Jianfeng Gao, and Weizhu Chen. ''Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing''. In: *arXiv preprint arXiv:2111.09543* (2021) (cit. on p. 20).

[R49] Claes Wohlin, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslen. *Experimentation in software engineering*. Springer Science & Business Media, 2012 (cit. on pp. 24, 25).

[R50] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. ''Natural language processing: State of the art, current trends and challenges''. In: *Multimedia tools and applications* (2022), pp. 1–32 (cit. on pp. 25, 26).

APPENDIX. MORE SPECIFIC DATA

This appendix shows all the information discussed in the paper but with more specific information. For instance, Tables 10, 11, 13, 12, 14 indicate which paper used which method for each Legal NLP task. In addition, Tables 15, 16, 17, 18, 19 show which papers work, which type of legal documents, and the embedding methods used. In addition, Table 9 maps the recent state-of-the-art language models to which papers in the Legal NLP field have used it.

TABLE 9. Language models used per paper

Language Model	Paper References
BERT	[P2, P9, P10, P13, P22, P23, P37, P52, P68] [P1, P27, P46, P47, P55, P56, P57, P60, P73]
RoBERTa	[P12, P13, P47, P51]
ALBERT	[P47]
ELECTRA	[P47]
Legal-BERT	[P1, P2, P13, P57]
Custom LegalBERT	[P1]
Legal-RoBERTa	[P6]
DistilBERT	[P23, P61]
SpanBERT	[P59]
Hier-BERT	[P9]
XLNet	[P8]
Lawformer	[P5, P6]
T5	[P1, P8]
DeBERTa	[P13]
BigBird	[P1, P13]
Longformer	[P5, P13]
CaseLaw-BERT	[P13]
GPT	[P3]

Table 15 summarizes the type of legal documents that were approached in these works and what have been the methods, word embeddings, and PLMs that have already been studied in the Binary and Multiclassification task. Furthermore, Table 10 summarizes which works have applied which methods in the area of Legal Text Multiclass Classification in general.

TABLE 10. Base Method and References of Multiclass Classification

Methods	References
Stack Attention	[P21]
LDA	[P7]
Naive Bayes	[P23]
Decision Trees	[P14]
SVM	[P7, P15, P16, P23]
LSTM, GRU, BiLSTM	[P7, P15, P16, P18, P23, P65]
Reinforcement Learning	[P65]
GNN	[P17]
CNN	[P7, P11, P15, P16, P18, P20, P30]
BiGRU-Att	[P9, P10]
HAN	[P9]
LWAN	[P9, P10]
ZERO-CNN-LWAN	[P10]
ZERO-BIGRU-LWAN	[P10]
BERT	[P8, P9, P10, P13, P22, P73]
HIER-BERT	[P9, P13]
XLNet	[P8]
T5	[P8]
DistilBERT	[P23]
RoBERTa	[P12, P13]
DeBERTa	[P13]
BigBird	[P13]
Longformer	[P13]
CaseLaw-BERT	[P13]

Table 16 summarizes the type of legal documents that were approached in these works and what have been the methods, word embeddings, and PLMs that have already been studied in the Summarization task. Furthermore, Table 11 summarizes which works have applied which methods in the area of Legal Text Summarization in general.

TABLE 11. Base Method and References of Summarization

Methods	References
InferSent + FFNN	[P67]
Sent2Vec + FFNN	[P67]
Naive Bayes	[P67]
Random Forests	[P67]
SVM	[P29]
TBS	[P67]
Text Rank	[P1]
SVD	[P26]
Pointer Generator Networks	[P66]
LSTM	[P31, P67]
BiLSTM + Attention	[P28]
BERT	[P1, P27]
T5	[P1]
BART	[P1]
Custom LegalBERT	[P1]
Dyploc	[P1]
Global Aware	[P1]

TABLE 12. Base Method and References of Question Answering and Information Retrieval

Methods	References
BM25	[P57]
BiDAF	[P46]
Word2Vec	[P43, P45]
Neural Attention	[P53]
CNN	[P54]
SVM	[P51, P54]
BERT	[P46, P47, P52, P55, P56, P57, P68]
RoBERTa	[P47, P51]
Legal-BERT	[P57]
ALBERT	[P47]
ELECTRA	[P47]

Table 18 summarizes the type of legal documents that were approached in these works and what have been the methods, word embeddings, and PLMs languages that have been already studied in the Question Answering and Information Retrieval tasks. Furthermore, Table 12 summarizes which works have applied which methods in the area of Legal Text Question Answering and Information Retrieval in general.

Table 17 summarizes the type of legal documents that were approached in these works and what have been the methods, word embeddings, and PLMs that have already been studied in the Information Extraction task. Furthermore, Table 13 summarizes which works have applied which methods in the area of Legal Text Information Extraction in general.

Table 19 summarizes the type of legal documents that were approached in these works and the methods, word embeddings, and PLMs that have already been studied in the Question Answering and Coreference Resolution task. Furthermore, Table 14 summarizes which works have applied which methods in the area of Legal Text Coreference Resolution in general.

TABLE 13. Base Method and References of Information Extraction

Methods	References
Rule-Based	[P32, P34, P36, P38]
LDA	[P35]
SVM	[P41]
BiLSTM+Attetion+CRF	[P40]
BERT	[P37]

TABLE 14. Base Method and References of Coreference Resolution

Methods	References
Rule-Based	[P58]
GNN	[P60]
SpanBERT	[P59]
BiLSTM	[P60]
BERT	[P60]

TABLE 15. Legal type of documents with the methods, embeddings, and PLMs used in Multiclass Classification

Doc. Type	Methods	Embeddings	PLMs	Reference
Privacy Policies	SVM, HMM, CNN, Naive Bayesian Multinomial	Word2Vec, Paragraph2Vec	XLNet, T5	[P8, P20]
Judgment Outcomes	Stack Attention, CNN, LSTM, BiGRU-ATT, HAN, LWAN	Word2Vec	BERT, HIER-BERT	[P9, P11, P14, P18, P21, P65]
General Legal Text	LDA, SVM, GNN, CNN, LSTM, GRU, BIGRU-ATT, HAN, CNN-LWAN, BIGRU-LWAN, ZERO-CNN-LWAN, ZERO-BIGRU-LWAN, SVM, Naïve Bayes, BiLSTM	Word2Vec, Doc2Vec, CBOW, Glove, FastText	BERT, Roberta, DistilBERT, DeBERTa, BigBird, Longformer, CaseLaw-BERT	[P1, P7, P10, P12, P17, P22, P23, P24, P63]
Assigning Petitions	SVM, GRU, LSTM, CNN, Gradient Boosting, Random Forest	Word2Vec, FastText, Glove		[P16]

TABLE 16. Legal type of documents with the methods, embeddings, and PLMs used in summarization

Doc. Type	Methods	Embeddings	PLMs	References
Patent Documents	BiLSTM	Doc2Vec		[P31]
General Legal Text	TF-IDF, Part of Speech Tagging, SVD, InferSent, LSTM, Naïve Bayes, Random Forests, TBS, BART, Text Rank, Global Aware, Dyploc	CBOW, Sent2Vec, Glove, Word2Vec	BERT, T5, Custom LegalBERT	[P1, P25, P26, P28, P29, P30, P66, P67]

TABLE 17. Legal type of documents with the methods, embeddings, and PLMs used in Information Extraction.

Doc. Type	Methods	Embeddings	PLMs	References
Privacy Policies	Rule-Based, CoreNLP	Word2Vec		[P34]
Court Records	BiLSTM, Attention, CRF	Word2Vec		[P40]
General Legal Text	StanfordParser, Rule-Based, Ontologies, JAPE rules, SVM	Word2Vec	BERT	[P32, P33, P36, P37, P38, P39, P41]
License Terms	LDA	Doc2Vec		[P35]

TABLE 18. Legal type of documents with the methods, embeddings, and PLMs used in Question Answering and Information Retrieval.

Doc. Type	Methods	Embeddings	PLMs	References
Civil Code	BiLSTM, Attention, CRF	Word2Vec		[P43]
General Legal Text	SVM, CNN, Neural Attention, BM25, W2VCent, BiDAF	Word2Vec, CBOW, FastText, Skip-Gram, Doc2Vec	BERT, BERT-PLI, C-BERT, S-BERT, RoBERTa, Custom Legal-BERT, LMIR, Legal-BERT	[P1, P42, P44, P45, P47, P48, P49, P50, P51, P52, P53, P54, P55, P56, P57]
Tax Ruling	BERT		BERT	[P68]

TABLE 19. Legal type of documents with the methods, embeddings, and PLMs used in Coreference Resolution

Doc. Type	Methods	Embeddings	PLMs	References
Tax Law	JAPE, Rule-Based	Word2Vec		[P58]
General Legal Text	GNN, BiLSTM BiDAF	Word2Vec, CBOW, FastText, Skip-Gram, Doc2Vec	BERT, Span-BERT	[P59, P60]

ERNESTO QUEVEDO received the B.S. degree in La Universidad de La Habana, Cuba, in 2020. Received the M.S. degree and currently doing its Ph.D. at Baylor University, USA.

From 2021 to 2023, he has been a Research Assistant with the Department of Computer Science at Baylor University. Since then, he has worked in multiple research areas, from Networking and Software Engineering to its specialty, which is Natural Language Processing (NLP). Currently, with a total of 6 publications. His research interests include Information Extraction, Question Answering, Information Retrieval, Large Language Models, and Ontology Learning. It has been awarded in two consecutive years with the award of Outstanding Graduate Student at Baylor University. It obtained a Best Paper Award in the LXAI workshop at NAACL 2022. He is a proud member of the Upsilon Pi Epsilon organization.

TOMAS CERNY is an Associate Professor of Systems and Industrial Engineering at the University of Arizona. His area of research is software engineering, code analysis, security, aspect-oriented programming, user interface engineering, and enterprise application design. He received his Master's and Ph.D. degrees from the Faculty of Electrical Engineering at the Czech Technical University in Prague and his M.S. degree from Baylor University.

From 2009 to 2017, he was an Assistant Professor of Computer Science at the Czech Technical University, FEE, Prague, Czech Republic. In 2017, he was a PostDoc at Baylor University, Texas, USA, and in the same year, he continued as an Assistant and Associate Professor with a concentration on Software Engineering till July 2023.

Dr. Cerny served 10+ years as the lead developer of the International Collegiate Programming Contest Management System. He authored nearly 100 publications mostly related to code analysis and aspect oriented programming. Among his awards is the Outstanding Service Award ACM SIGAPP 2018 and 2015 or the 2011 ICPC Joseph S. DeBlasi Outstanding Contribution Award. In the past few years, he chaired multiple conferences including ACM SAC, ACM RACS, or ICITCS. Furthermore, he led special issues and track on Code Analysis and Enterprise Applications.

ALEJANDRO RODRIGUEZ B.S. in Computer Science from the University of Havana, Cuba, in 2020. Graduate student at Baylor University, USA.

Exercised as Research Assistant for one and a half years at the Computer Science Department at Baylor University. Experienced in Software Engineering roles in the industry. Researches Natural Language Processing and Machine Learning applications in areas like fighting cybercrime, subaquatic organisms classification, and software engineering. Proud member of the Upsilon Pi Epsilon honors society.

PABLO RIVAS (S'01–M'11–SM'18) received the B.S. degree in computer systems engineering from the Nogales Institute of Technology, Nogales, Mexico, in 2003, the M.S. degree in electrical engineering from the Chihuahua Institute of Technology, Chihuahua, Mexico, in 2007, and the Ph.D. degree in electrical and computer engineering from The University of Texas at El Paso, El Paso, TX, USA, in 2011.

He has been an Assistant Professor of Computer Science with the School of Engineering and Computer Science, Baylor University, Waco, TX, USA, since 2020. Before that, he was with the School of Computer Science and Mathematics, Marist College, Poughkeepsie, NY, USA, from 2015 to 2020. He has more than eight years of industry experience as a Software Engineer and has been recognized for his creativity and academic excellence. He is currently in the planning phase of the Center for Standards and Ethics in Artificial Intelligence with funding from the National Science Foundation. He has published several peer-reviewed papers and authored a book on deep learning in 2020. He predominantly researches artificial intelligence and its ethical and social implications, focusing on computer vision, natural language processing, and quantum machine learning.

Dr. Rivas is a member of the IEEE Standards Association and is involved in the working groups developing the P70XX series standards for AI ethics. In 2011, he was inducted into the international honor society for IEEE Eta Kappa Nu; in 2021, he was inducted into Upsilon Pi Epsilon, the international honor society for the computing and information disciplines; and in 2022, he was elevated to a Senior Member of ACM.

JORGE YERO Received the B.S. degree in La Universidad de La Habana, Cuba, in 2020. Currently, he is pursuing his Ph.D. at Baylor University, USA.

Between 2022 and 2023, he was a Research Assistant with the Department of Computer Science at Baylor University. Experience in Software Engineering, DevOps, and Business intelligence roles in industry. His research interests include code analysis, microservice architecture and code generators.

KORN SOOKSATRA Korn Sooksatra is a Ph.D. candidate at Baylor University's Department of Computer Science. He holds a master's degree from Georgia State University.

His primary research focus lies in enhancing the robustness of machine learning models, with a special emphasis on tackling adversarial challenges. He utilizes his expertise in diverse areas, including computer vision, natural language processing, time-series data, and cybersecurity.

ALIBEK ZHAKUBAYEV He earned his B.S. and M.S. degrees from Nazarbayev University, Kazakhstan, in 2018 and 2020, respectively. Currently, he is pursuing his Ph.D. at Baylor University, USA.

Between 2018 and 2020, he served as a Research Assistant in the Department of Computer Science at Nazarbayev University. He then continued his research work at Baylor University's Department of Computer Science from 2020 to 2023. His research interests include activity recognition, computer vision, bioinformatics, and clustering.

DAVIDE TAIBI is a full Professor at the University of Oulu (Finland) where he heads the M3S Cloud research group. His research is mainly focused on Empirical Software Engineering applied to cloud-native systems, with a special focus on the migration from monolithic to cloud-native applications. He is investigating processes, and techniques for developing Cloud Native applications, identifying cloud-native specific patterns and anti-patterns. He is a member of the International Software Engineering Network (ISERN) since 2018. Before moving to Finland, he has been Assistant Professor at the Free University of Bozen/Bolzano (2015-2017), a post-doctoral research fellow at the Technical University of Kaiserslautern and Fraunhofer Institute for Experimental Software Engineering - IESE (2013-2014) and research fellow at the University of Insubria (2007-2011).