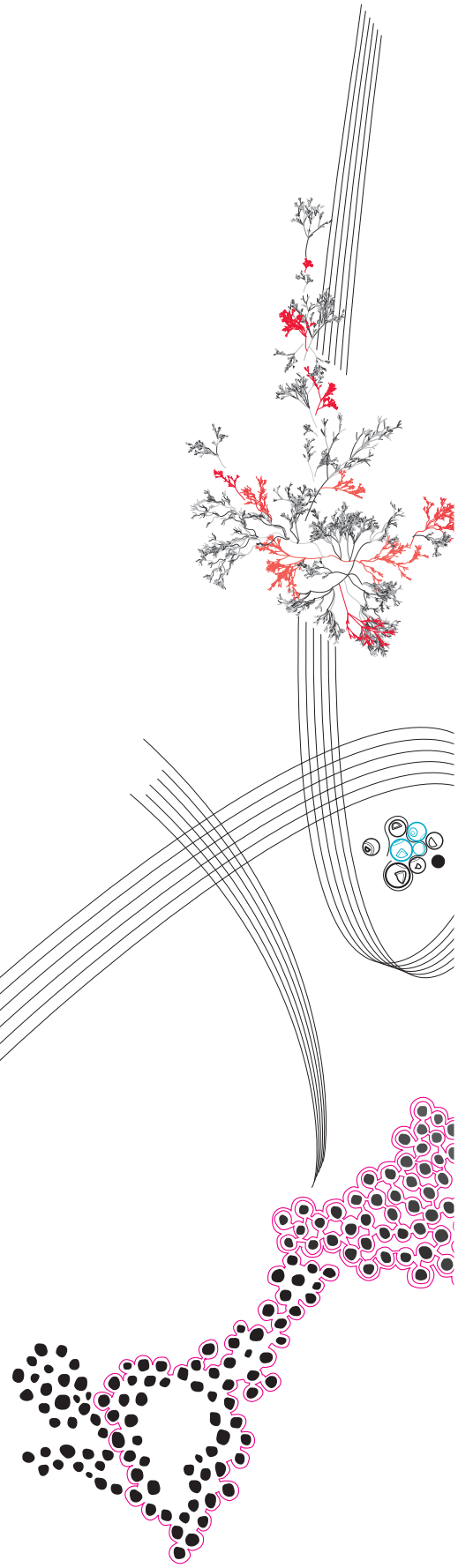

MSc Business Information Technology
Master Thesis

RechtBERT: Training a Dutch Legal BERT Model to Enhance LegalTech

Maarten S. Looijenga

Supervisors:
dr. F.A. Bukhsh (EEMCS)
dr.ir. J.M. Moonen (BMS)
N. Bouali MSc (EEMCS)

December 18, 2024

Faculty of Electrical Engineering
Mathematics and Computer Science,
University of Twente

UNIVERSITY OF TWENTE.

Acknowledgements

This thesis marks the conclusion of my journey at the University of Twente and the completion of my second master's degree. After earning a bachelor's degree in Business & IT and a master's degree in Science Education, I am proud to complete a master's in Business Information Technology, specialising in Implementation Management & Enterprise Architecture. Although my academic path turned out quite differently from what I initially anticipated, I am grateful for the milestones I have achieved, culminating in two remarkable master's degrees.

Before presenting this thesis, I would like to thank several individuals. First and foremost, I would like to thank my primary supervisor, Faiza. I sincerely appreciate her feedback and support. Faiza always made time for our meetings, was consistently well-prepared, and provided valuable input and guidance, for which I am truly grateful.

I would also like to thank my other supervisors, Hans and Nacir, for their input and insights. Their perspectives have been very helpful in establishing a broader context for this research and addressing the challenges I encountered along the way.

This research was conducted at a major Dutch law firm. To maintain the anonymity of this study, I will refrain from mentioning any names, but I am incredibly thankful for the opportunity to conduct IT research at this organisation. I am especially grateful to the director for granting me this opportunity, trusting me, and providing me with the chance to work at the office every day and fully engage with all employees.

I sincerely thank my daily supervisor for his guidance, support, and the enjoyable moments we shared, including countless coffee breaks and interesting discussions, whether about the research itself or other topics. I am grateful for being made to feel like a fully integrated part of the team and for being involved in all the office's activities.

Additionally, I want to thank all the employees for their interest in me and my research, the valuable insights they shared about their work, and, most importantly, for making my time at the firm so enjoyable. Although my research did not always excite me, I genuinely looked forward to going to work every day.

Finally, I thank my family and friends for their unconditional support, assistance, and encouragement throughout my studies. Without them, I would not have come this far.

To all readers of this thesis, thank you for your interest. I hope you enjoy reading it and that it provides valuable insights and learning opportunities.

Contents

List of Figures

List of Tables

Acronyms

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

DSRM Design Science Research Methodology

ECLI European Case Law Identifier

EU European Union

KOOP Kennis- en Exploitatiecentrum voor Officiële Overheidspublicaties

legaltech Legal technology

LLM Large Language Model

ML Machine Learning

MLM Masked Language Modelling

NLP Natural Language Processing

NSP Next Sentence Prediction

OP Publications Office of the European Union

SLR Systematic Literature Review

SOP Sentence Order Prediction

WWM Whole Word Masking

RechtBERT: Training a Dutch Legal BERT Model to Enhance LegalTech

Maarten S. Looijenga

December 18, 2024

Abstract

The introduction of generative AI has got massive user interest and sparked discussions about AI adoption in the legal domain. Many AI applications in the legal field are built upon the BERT transformer architecture, but little research has been conducted on adapting such models to the Dutch language. This research aims to design a domain-specific legal Dutch BERT model that outperforms generic Dutch BERT models, enabling legal professionals to perform tasks more efficiently and advance legal tech through NLP applications. It introduces a set of domain-specific legal Dutch BERT models called RechtBERT. We conclude that further pre-training existing Dutch BERT models does not yield better performance on legal NLP tasks than using the Dutch BERT models out of the box. This research addresses a gap in literature regarding the development and use of domain-specific legal Dutch BERT models.

Chapter 1

Introduction

1.1 Problem Identification & Research Goal

Artificial Intelligence (AI) has experienced an exponential rise in recent years. Where in 2016 the commercial and private use of AI was marginal, adaptation by organisations doubled between 2017 and 2022 [27]. A breakout was achieved by the introduction of a publicly available generative AI tool called "ChatGPT" in November 2022. This tool could chat with users and generate text based on users' text prompts [37, 84]. Through this tool, massive interest in AI arose, and the tool gathered a user base of 100 million users within two months after launch [84].

Discussions about AI adaptation in the legal industry increased significantly following the exponential growth. Implementing AI in this sector is met with both enthusiasm for its possibilities and a heightened awareness of the potential risks. The integration of AI, while linked to an increased chance of data threads, leakage of company information [1, 12, 51], and elevated risk of cyberattacks [12], also promises to make legal workers more efficient [4, 11, 28, 44, 48, 64, 111, 114, 117, 118] and deepen and broaden their areas of expertise [4, 48].

Many AI applications in the legal domain are built upon the BERT transformer architecture introduced by Devlin et al. [38]. The impressive performance of the original BERT has inspired numerous researchers to refine and extend this transformer model [2, 21, 33, 36, 78]. This research introduces a new set of domain-specific legal BERT transformers for use in the Dutch legal domain. By further pre-training existing generic Dutch BERT models on Dutch legal data, these models can be used for downstream legal tasks. The set of models is evaluated by performing legal topic classification on a large EU dataset. While research has already been conducted on transcoders specifically for the legal domain in one language [2, 21, 41, 78] and transformers for generic Dutch tasks [33, 36, 38], the combination has not yet been explored. This research answers the main research question:

> ***How to*** *design a domain-specific legal Dutch BERT model* ***that*** *outperforms generic Dutch BERT models* ***so that*** *legal professionals can perform tasks more efficiently* ***in*** *the advancement of legaltech through NLP applications?*

1.2 Research Problem & Questions

The stated main research question is a technical research problem formulated according to the design problem template of Wieringa [124]. To solve this problem, additional research

questions have been formulated:

- **RQ1:** What is the current state of research on the use of AI in legaltech?
- **RQ2:** Which AI integration opportunities address the challenges faced by legal professionals?
- **RQ3:** What insights from existing BERT-based models can guide the design and training of RechtBERT for optimal performance in legal NLP tasks?
- **RQ4:** How can the identified characteristics of BERT models guide the development of a legal Dutch BERT model for NLP tasks?
- **RQ5:** What NLP applications can be used to validate the performance of domain-specific legal Dutch BERT models?
- **RQ6:** How does the domain-specific legal Dutch BERT model's performance compare to that of generic Dutch BERT models on previously selected tasks?

1.3 Research Methodology

To answer these questions, research is conducted following the *Design Science Research Methodology* (DSRM) of Peffers et al. [102], as this methodology is specifically designed for design research in information systems. The methodology consists of six phases: Identify Problem & Motivate, Define Objectives of a Solution, Design & Development, Demonstration, Evaluation and Communication. Peffers et al. [102] also defined four different entry points from which the research can be started. The DSRM process flow is shown in Figure 1.1. Next, we discuss the starting point of this research and each phase of the DSRM.

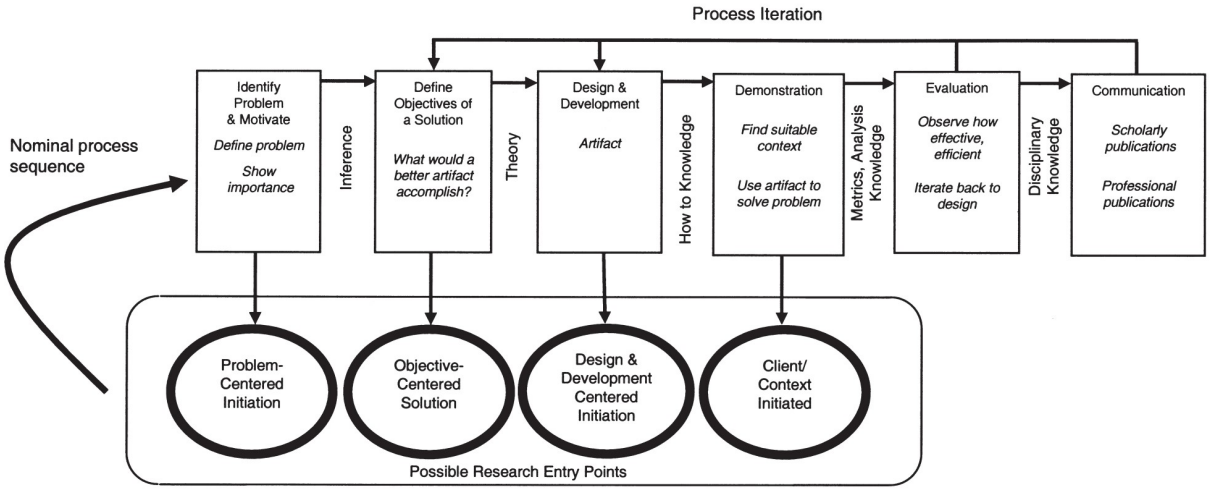

FIGURE 1.1: DSRM Process Flow by Peffers [102]

Research Entry Point The DSRM is structured sequentially but allows iteration to improve the artefact based on feedback in later phases. Additionally, the process can begin at any step and expand outward. Peffers et al. [102] identify four starting points for initiating the sequence.

This research starts at activity 2. Peffers et al. describe that "an objective-centred solution, starting with activity 2, could be triggered by an industry or research needs that can be addressed by developing an artefact" [102]. This study was prompted by an industry question regarding the potential effects of AI in their field. A literature review revealed a need for further research into AI applications and their feasibility. Consequently, this design problem is addressed by developing an artefact.

DSRM 1 - Identify Problem & Motivate The first phase of the DSRM defines the research problem and justifies the value of the developed solution. The problem needs to be conceptualized so that the solution can capture its complexity. The solution's value needs to be justified to motivate the researchers and the audience to pursue a solution and accept the results. It also helps to understand the reasoning associated with the researcher's understanding of the problem.

This research starts with problem identification in the first chapter, formulating the research problem and questions to develop an artefact to solve the stated problem.

DSRM 2 - Define Objectives of a Solution The next step is to determine the performance objectives of the solution. Peffers et al. [102] state that these objectives should be rationally inferred from the problem specification and require knowledge of the current problems, solutions, and their efficacy.

This research presents the results of RQ1 by a literature review that aims to investigate the state of problems concerning the use of AI in legaltech and current implementations that have already been researched. The review is conducted to gain knowledge of the state of problems and current solutions. Next, through expert reviews, the results of this literature review are evaluated, and possible AI applications are reviewed by experts. Besides, a literature review has been performed on relevant BERT models, answering RQ3.

DSRM 3 - Design & Development The third phase of design science focuses on designing and developing the artefact. Desired functionality and architecture are determined, followed by artefact creation. Requirements are derived from insights from the previous phase.

This phase explores various dataset options and training techniques to develop the artefact. Training metrics are analysed to assess performance, answering RQ4.

DSRM 4 - Demonstration In this phase, the artefact is demonstrated through experiments, case studies, or simulations, ensuring it addresses at least one defined problem and functions as intended to solve it. A few options for using our legal BERT model are explored, and a legal topic classification task has been performed, answering both RQ5 and RQ6.

DSRM 5 - Evaluation The evaluation phase automatically follows the demonstration phase. The results of the demonstration phase are evaluated. In combination with the previous phase, this section answers RQ6.

DSRM 6 - Communication When the artefact is developed and tested, the results must be communicated. Everything relevant to the audience needs to be communicated. This will be done using presentations and this report. Presentations will be given to the company of which experts are used to create the artefact. Since this thesis is part of a graduation project, a public colloquium will also be held.

1.4 Thesis Structure

The thesis follows the framework of Peffers' DSRM [102], as illustrated in Figure 1.1. Chapter 1 serves as an introduction, identifying the problem, explaining the research goals and motivations, and stating the research questions and methodology. Chapter 2 tells the foundational aspects of the thesis, offering insights into AI and legal technology and conducting a literature review on AI's current applications in the legal domain. Besides, it shows the results of an expert review and further in-depth literature review into BERT models. Chapter 3 goes into the design criteria of the artefact, elaborating on the most suitable transcoder techniques to use within the legal domain and stating the data utilized for training. In Chapter 4, the developed models are evaluated on a legal topic classification task. Besides, other validation options are discussed. Chapter 5 draws conclusions and states options for future research.

Chapter	**DSRM Activity**	**Research question**
Chapter 1: Introduction	DSRM 1	-
Chapter 2: Background	DSRM 2	RQ1 & RQ2 & RQ3
Chapter 3: Modelling RechtBERT	DSRM 3	RQ4
Chapter 4: Experiment & Validation	DSRM 4 & DSRM 5	RQ5 & RQ6
Chapter 5: Conclusion & Future Work	DSRM 6	ALL

TABLE 1.1: Thesis structure mapped to DSRM activities of Peffers [102]

Chapter 2

State of the Art

This chapter provides an overview of the state-of-the-art research on the application of artificial intelligence in the legal sector. A *Systematic Literature Review* (SLR) was conducted, of which the results relevant to our research objectives are discussed in Section 2.1, including relevant key concepts. In the subsequent Section 2.2, expert reviews are presented to evaluate the implications of these findings in the real-world context of a law firm. Finally, Section 2.3 elaborates on related work involving BERT transformers, a technology identified in the literature as promising for using AI in the legal field and used in this research.

2.1 Findings Systematic Literature Review

A SLR was conducted to examine the use of AI in the legal sector. The review has resulted in a qualitative investigation of 94 selected papers about the use of AI in legaltech. The research followed the guidelines for systematic literature reviews proposed by Kitchenham & Charters [61]. The review focused on publication trends, AI applications in legal domains, AI technologies, and the associated benefits, challenges, and limitations. The SLR shows a growing interest in the use of AI in legaltech. The findings of this research can serve as a foundational source for future investigations into the use of AI within the legal domain [1].

2.1.1 Search methodology

Search library Scopus was chosen as the primary digital library due to its comprehensive collection of peer-reviewed journals and conference proceedings, as well as its user-friendly interface [49, 88]. Additional searches were executed in other libraries to include literature that may not have been covered in Scopus, being Web of Science, ScienceDirect, ACM, IEEE Xplore, JSTOR, and LegalIntelligence.

Search query The most common keywords used in research about using AI in the legal sector are identified to construct the search string. For AI, both the abbreviation and the fully written word were used. Furthermore, it was noted that AI-related technologies as *Machine Learning* (ML) [35, 82] and *Natural Language Processing* (NLP) [30, 89, 91]

[1] The SLR is submitted for publication as a separate paper titled *"A Systematic Literature Review on the Evolution of Artificial Intelligence within Legaltech."* Since the publication process is ongoing, no citation is provided.

are used in research about AI in the legal sector without explicitly mentioning AI in the abstract, keywords or title, which was resolved by adding these terms to the search query.

For the legal sector, the term "legaltech" is identified as the term used for technology that performs tasks generally handled by legal professionals [107], with variations being "legal tech", "legal technology" and "lawtech" [107]. There is no clear distinction between using these terms besides preference based on geography [107].

All terms were used in both their total and abbreviated forms. Wildcard characters have been used to search for both "tech" and "technology". The final search string which is used for data extraction is as follows:

> *("legal tech*" OR "legaltech*" OR "law tech*" OR "lawtech*" OR "legal profession" OR "legal practise") AND ("artificial intelligence" OR ai OR "natural language processing" OR nlp OR "machine learning" OR ml)*

Inclusion criteria Inclusion criteria are defined to only take into account studies that are relevant to the review's objectives. The inclusion criteria are:

1. The paper is directly relevant to the scope of our review, focusing exclusively on the application of artificial intelligence to enhance legal services.
2. The paper does discuss legal technology and artificial intelligence as its main topic.
3. The paper is published in a peer-reviewed journal or conference proceeding.
4. The paper is written in English.
5. The paper is available for download.

The inclusion criteria have been applied via automatic filtering options within the various search engines. Then, all remaining literature is manually analysed to use the first two inclusion criteria. The selection process only looks at a paper's title, keywords, and abstract, ignoring the remainder of its text.

2.1.2 Search execution

On October 6th, 2023, the final search was conducted on Scopus and the other libraries, discovering 380 papers. Figure 2.1 gives a detailed overview of our execution of the data extraction process. We applied the inclusion criteria, excluding 38 non-English papers and 78 non-peer-reviewed papers, resulting in 264 papers. Papers from various libraries were combined, duplicate papers removed, and any results without a full-text version available were removed. During this process, it was discovered that four papers were not written entirely in English, so these papers were also excluded, leaving a total of 148 papers. The specific number of documents at each step for each library can be found in Table A.1 in Appendix A.1.1.

The remaining papers have been subject to a qualitative content analysis. This analysis checked all titles, abstracts, and keywords to see if the papers are directly relevant to the scope of the SLR and if their main topic discusses legaltech and AI. After this last selection, 91 papers remained.

Two significant categories of papers have been excluded from further review, although they were notably present in the subset of papers selected for the SLR. First, we excluded eight papers that addressed the education of law students. Although these papers examined how IT might change the way legal professionals practice, they were excluded from our

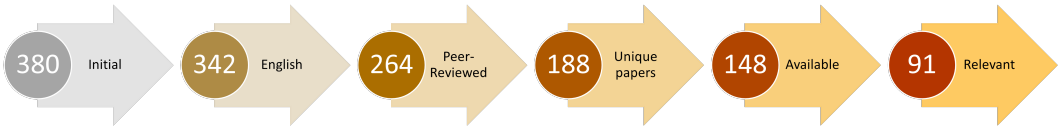

FIGURE 2.1: Data Execution Process

study because they were deemed not directly relevant. Papers concerning the ethics of using AI systems in legal practice have also been excluded since these papers do not offer any new information regarding the use of AI in legaltech.

2.1.3 Key concepts and definitions

Legaltech *Legal technology* (legaltech) refers to the application of innovative technologies within the legal field. These technologies typically perform tasks traditionally carried out by lawyers and other legal professionals [107]. The primary aim of legaltech is to improve the delivery of legal services through technical solutions. These solutions aim to enhance efficiency, productivity, cost-effectiveness, and overall client outcomes [107]. More information on the history and evolution of legaltech can be found in Appendix A.1.2.

Artificial Intelligence *Artificial Intelligence* (AI) is the simulation of human intelligence in machines programmed to think and learn. According to Oxford English Dictionary [100], AI is defined as "the capacity of computers or other machines to exhibit or simulate intelligent behaviour". AI systems simulate human cognitive processes like representation, learning, rules and search to address real-world challenges [26]. AI concepts and theories are widely applied across engineering disciplines, including automation, production, optimization, and planning [26]. AI serves as an overarching discipline with multiple specialized subfields, such as *Machine Learning* (ML) and *Natural Language Processing* (NLP).

Machine Learning *Machine Learning* (ML) is a subset of AI that focuses on creating algorithms that allow systems to learn and improve from experience without explicit programming. These algorithms can learn from data and make predictions based on it [62]. ML relies on statistical methods to identify patterns and make decisions or predictions based on available information.

The different algorithms are distinguished by their approaches to learning, pattern recognition, and the outcomes they produce, which align with user expectations. There is no unique algorithm for detecting valuable patterns. The choice of algorithm depends on the specific problem a developer is trying to solve and the data available for analysis [74]. More information on various fields of ML can be found in Appendix A.1.3.

Natural Language Processing Humans are considered to be intelligent beings. They present and update their knowledge, reasoning, and inferences through natural languages, like English or Dutch [26]. As society increasingly digitises, there is a growing abundance of documents written in natural language. The sheer volume of these documents makes

extracting knowledge from them progressively challenging, particularly within specified time constraints [26].

The aim of automating *Natural Language Processing* (NLP) is to efficiently and accurately handle this task. However, determining the meaning of each word in a sentence through automation is challenging, mainly because it requires understanding the associations and contextual knowledge associated with each word. NLP stands as a subfield of AI, primarily employed in tasks such as information retrieval, machine translation, question-and-answer systems, and summarization. NLP uses a set of computational techniques designed for the automatic analysis and representation of human languages, driven by theoretical foundations [26].

One significant advancement in NLP development is the introduction of transformers, in particular, the by Google developed BERT [38]. This transformer is designed "to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers" [38]. Pre-trained language models effectively improve many NLP tasks, including sentence-level tasks, paraphrasing and token-level tasks. Technologies before BERT restricted the power of pre-trained representations, especially for fine-tuning approaches. However, a BERT model can be fine-tuned with just one additional output layer to create state-of-the-art models for various tasks. BERT is elaborated on more in Section 2.3

2.1.4 Publication trend

Tracing the historical development of AI in legaltech, the first paper on this topic was published in 1988. From then until 2016, occasional publications marked the field's early stages. However, a significant surge in interest and research can be observed from 2016 onwards, indicating a growing interest in AI in legaltech. Figure 2.2 illustrates the distribution of papers according to their publication year.

The collection of research papers (91) is almost evenly split between articles published in journals (50) and conference proceedings (41), reflecting a trend towards peer-reviewed research in this relatively new field. A classification system has been applied, categorising papers as descriptive or non-descriptive.

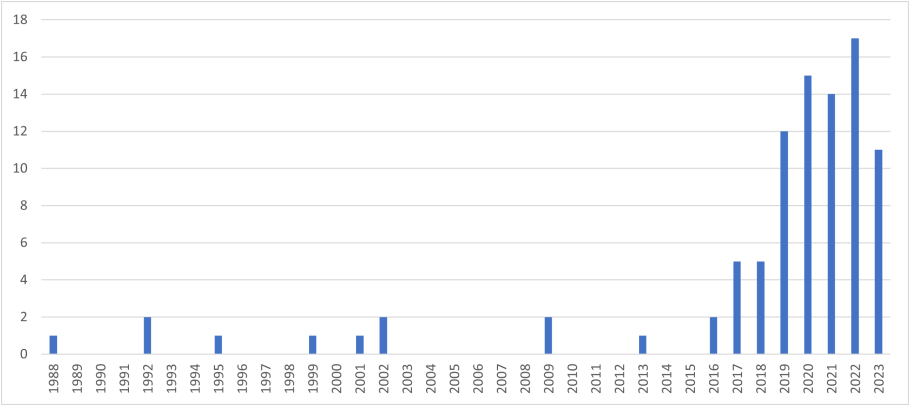

FIGURE 2.2: Number of Publications per Year

When examining the publication trend, it becomes clear that papers addressing the application of AI in the legal domain have been available for a considerable time. However, a substantial rise was detected around 2016. Since then, many papers have been released about the topic, indicating a significant interest in AI within the legal domain, with a

considerable volume of scientific research already conducted. Furthermore, it is noteworthy that a substantial portion of this research undergoes rigorous peer review, indicating a commitment to thoroughly investigating this topic.

2.1.5 Legal domains

The review examined which legal domains AI are most widely adopted. While many papers offer a general perspective on the legal field without specifying a particular area of expertise, most papers focus on a single legal domain, concentrating on lawyers and law firms. Another noteworthy category includes papers discussing the field of justice, where the focus is on the augmentation or replacement of judgment and dispute resolution within the court system. Additionally, some papers explore how AI can support clients, government & lawmakers, and law enforcement. The various domains and the corresponding number of papers explicitly applicable to each domain are detailed in Figure 2.3. In contrast, the exact studies per legal domain are shown in Table A.2, which is included in Appendix A.1.4.

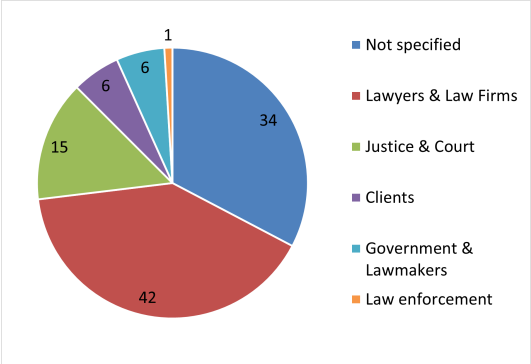

FIGURE 2.3: Presence of Legal Domains in Research Papers

It is interesting to see that many papers are focusing on the domain of lawyers and law firms. This observation leads to the conclusion that current research efforts by scientists are mainly concentrated on developing and implementing AI solutions within this specific legal field.

2.1.6 AI applications

Several applications of AI within legaltech have been identified, which are shown in Figure 2.4. Of all the papers, the most discussed application is eDiscovery, in which AI categorises documents as relevant or irrelevant. Litigation & Prediction analysis is often mentioned as a promising application to predict the outcome of litigation of court decisions. Legal search involves studying legal documents by AI in large databases to get better results and more of what the searchers seek. Document generation uses AI to create documents, while contract analysis encompasses examining contracts and identifying crucial information that may need additions, alterations, or deletions within the contractual framework. The exact studies per application are listed in Table A.3, which is included in Appendix A.1.5.

Additionally, specific applications are occasionally referenced in the papers. One such application is due diligence reviews, employed for examining a wide array of corporate documents, particularly for assessment in the context of acquisitions or mergers. Predictive billing is another application used to optimise the allocation of fee earners and provide informed estimates of the costs associated with legal services. Moreover, the potential for

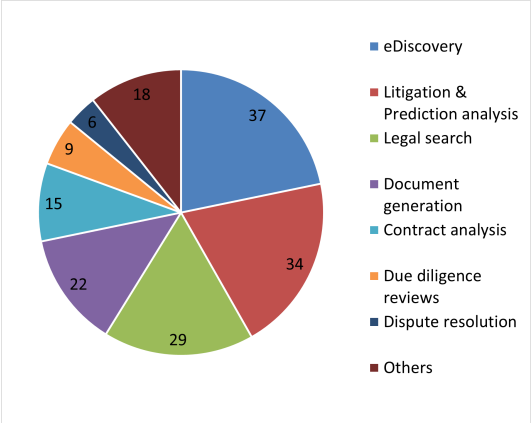

FIGURE 2.4: Number of AI Applications in Research Papers

dispute resolution through E-justice is acknowledged, where online, digital, or alternative dispute resolution mechanisms are leveraged to automate and optimise decision-making procedures.

This section concludes that specific fields within AI applications are interesting for further research. The identification of promising applications suggests some opportunities for future investigation. This insight provides developers and researchers with diverse, prospective options and underscores the high potential for future work in AI applications.

2.1.7 AI technologies

All sample papers have been analysed to identify used AI technologies. AI is a vast field with many branches. Two categories were determined to be overrepresented in the adoption of discussion on AI in legaltech: *Machine Learning* (ML) and *Natural Language Processing* (NLP). While machine learning can sometimes be used to complete NLP tasks, the technologies are discussed separately for this review.

ML explores the study and construction of algorithms, which learn from and make predictions on data [62]. There is no universal algorithm for detecting valuable patterns. The choice of algorithms depends on the specific problem the developer is solving and the available data [74]. Supervised machine learning techniques are primarily represented when looking at ML. Also, deep learning is a widespread technique in the field of ML with legaltech. Tables listing various algorithms and the studies in which they are represented for supervised learning (A.4), unsupervised learning (A.5), and deep learning (A.6) are provided in Appendix A.1.6.

NLP is a branch of artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language, aiming to efficiently and accurately handle automatic analysis and representation of human languages [26]. A clear trend emerges in the NLP studies, with the BERT language model being the most extensively employed model. BERT manifests in various iterations tailored to specific purposes within the legaltech domain. Identified families of language models and various types of BERT mentioned in the literature, along with their corresponding studies, are presented as Table A.7 and Table A.8 in Appendix A.1.7.

2.1.8 Benefits, challenges and limitations

Many benefits, challenges, and limitations are mentioned and discussed in the reviewed papers. We see multiple benefits for the use of AI in legal technology. AI can complement the work of lawyers and is considered a potential substantial improvement in efficiency [4, 11, 28, 44, 48, 64, 111, 114, 117, 118]. Alongside improving efficiency, AI could benefit from improved accuracy [1, 28, 44, 117], reduced costs [1, 44, 45, 48, 128], time-saving [128], increased revenue [11, 45, 48], and improved quality of work [64, 111].

Furthermore, AI can decrease the gap between large and small law firms [4, 48]. Moreover, a notable benefit of AI algorithms is that, unlike humans, they do not experience fatigue, health issues, or fluctuations in productivity [48, 60, 29]. Besides, AI can help analyse large volumes of data much better [119, 128]. AI algorithms can not bias colour, gender, or other visible characteristics [29, 32, 34]. A possible effect of using AI in the legal sector is the opportunity to change the work that lawyers can perform, allowing lawyers to broaden or further specialise their work [4, 48, 64, 86, 114].

There are also some challenges and limitations to consider when discussing the use of AI within legal technology. One of the most commonly cited challenges in implementing AI within the legal sector is liability and the potential risk of privacy infringement [24]. Firms are worried about the possible leakage of sensitive information [1, 51]. The integration of AI is often linked to an increased chance of data threats, leakage of important company information and mismanagement of essential information [1, 12]. Moreover, using AI is frequently linked to an elevated risk of cyberattacks [12].

Besides the challenges connected to privacy, litigation, and risk, additional challenges and limitations must be considered. Firstly, AI is considered complex, making it challenging for most people to understand and trust its functioning [60, 114, 128]. Additionally, the field of law is very complex, making it difficult for IT specialists to develop AI programs for this specific sector [98, 128]. The cost of implementing and maintaining an AI system and how money is earned in a law firm are also considered limitations [4, 5, 51]. Using billable hours and the partner structure that most law firms have further add to these limitations [6, 7, 30, 48, 64]. Moreover, the current capabilities of AI mean that lawyers cannot be replaced, so human input and validation are still required [3, 4, 5, 34, 48, 60].

2.2 Expert Reviews

The SLR shows a detailed state of the art on using AI in legaltech, including possible AI applications and their benefits and limitations. Input from experts has been gathered to adapt theoretical findings to real-world applications. Experienced legal professionals from one major Dutch law firm have been interviewed, focusing on identifying repetitive and labour-intensive workflows and the possible use of AI in these tasks.

2.2.1 Methodology

In December 2023, four fee-earners from a Dutch law firm participated in face-to-face interviews, each lasting between 30 and 60 minutes. These participants volunteered without compensation and were selected to represent diverse legal expertise, including corporate law, commercial law, employment law, contract law, litigation and liability law, and civil law notary services. The group included a candidate lawyer and three lawyers, one of whom is a partner on the firm's board. Some participants have expertise in multiple fields. A structured interview guide, detailed in Appendix A.2, guided the conversation, allowing for open, in-depth discussion.

2.2.2 Results

Topic	Lawyer 1	Lawyer 2	Lawyer 3	Lawyer 4
Repetitive Nature of Documents	Agrees – documents 85-90% similar	Agrees – recognizes documents overlap	Agrees	Disagrees – work not repetitive, but has overlap
Use of Templates and Models	Disagrees – no models, lack of investment	Disagrees – colleagues hesitate to share templates	Agrees – everyone uses own templates	Agrees – department uses automated models heavily
Access to Legal Resources (e.g., LegalIntelligence)	Agrees – values articles and blogs via LegalIntelligence	Agrees – uses it but finds interface and search algorithms suboptimal	Agrees – combines it with Google searches for better results	Not stated
Efficiency of Document Search	Not stated	Disagrees – firm's IT system doesn't allow full document search	Disagrees – inefficient and time-consuming	Disagrees – supports claim that current search method is inefficient
Translation Work for International Clients	Not stated	Agrees – translation is repetitive, often uses DeepL, questions billing method	Agrees – repetitive; ChatGPT for consistency in translations	Agrees – ChatGPT over DeepL for fewer synonyms and better consistency
View on AI in Law Practice	Skeptical – doesn't understand the hype around AI	Concerned – hesitant due to privacy issues and previous IT failures	Optimistic – believes firm should adopt AI immediately, but start small	Not stated

TABLE 2.1: Summary of Lawyer Perspectives on Key Topics

All insights from the expert review have been summarized and organized by topic in Table 2.1. This table provides an overview of each lawyer's perspective on key themes. A more detailed report, including specific statements from each lawyer, is attached as

Appendix A.3.

The lawyers' feedback from the interviews shows the challenges in integrating AI and automation into the legal domain. While most acknowledge that tasks like document drafting and translation could benefit from automation, individual lawyers have varying opinions about technology. Challenges include IT infrastructure, privacy concerns, and an overall reluctance to use new tools, resulting in some departments falling behind in automation.

The potential of AI in legal research, document generation, and translation is recognized. However, doubts persist about its immediate benefits and the feasibility of implementing new IT within the firm. Successful integration of AI into legal workflows will require a measured small-scale approach with projects that demonstrate clear value. Additionally, promoting a more collaborative culture around shared resources such as templates and improving internal search tools could help some of the inefficiencies currently faced by lawyers without immediately needing new large IT systems.

2.3 BERT

From the SLR, we can conclude that many AI applications in the legal domain are built upon the BERT transformer architecture introduced by Devlin et al. [38]. The impressive performance of the original BERT has inspired numerous researchers to refine and extend this transformer model [2]. Section 2.3.1 provides a detailed comparison of the work conducted by various researchers in developing BERT-based models specifically tuned for the legal domain. These outputs form the design choices for our model, including the training techniques and parameters used. Additionally, Section 2.3.2 discusses generic BERT models trained for the Dutch language, which are further pre-trained on legal datasets to enhance their applicability in Dutch legal contexts. All training settings are summarized in Table 2.2.

Name	**Task**	**Steps**/ Epochs	**Optimizer**	**Learning** rate	**Batch** size	**Sequence** length	**GPU**
Legal-BERT	MLM, NSP	1M steps 40 epochs	AdamW	1e-4	256	512	Google v3 TPUs
JuriBERT	MLM	6M-110M parameters	AdamW	1e-4	4, 8	512	GTX 1080Ti
jurBERT	MLM NSP	40 epochs	Slanted traingular	2e-5	- -	128 (90%) 512 (10%)	v3-8 TPU
AraLegal -BERT	MLM	50 epochs	-	1e-5- 5e-5	512	512	8x
BERTje	MLM SOP	1M iterations	-	-	-	-	Google TPUs
RobBERT	MLM	2 epochs	AdamW	1e-6	-	-	4x P100
mBERT	MLM NSP	-	-	-	-	-	Google TPUs

TABLE 2.2: Training settings of BERT models

The original BERT model of Devlin et al. is pre-trained using two unsupervised tasks: *Masked Language Modelling* (MLM) and *Next Sentence Prediction* (NSP) [38]. However, it has been shown that this approach for pre-training does not work well when the transcoder

is used within a domain with particular terminologies, such as legal, science or medicine [2, 13, 21, 68]. Two different strategies have been investigated to overcome this limitation: removing the NSP task from the pre-training process [72] or replacing the NSP task with another training task [66].

2.3.1 Domain-specific Legal BERT models

Many researchers tried to improve the BERT model of Devlin [38]. BERT has been reported to underperform in specific domains, including the legal domain [13, 21, 68]. Some research has been performed on developing a domain-specific legal BERT model for a specific language.

Chalkidis et al. [21] identified three different alternatives when employing BERT for NLP tasks in specialised domains:

1. Use BERT out of the box
2. Further pre-train BERT on domain-specific corpora
3. Pre-train BERT from scratch on domain-specific corpora

All models compared in this section use the same procedure of Chalkidis to create a legal-specific BERT model that outperforms the generic BERT model of Devlin et al. [38]. The models that are compared are LegalBERT (English) [21], JuriBERT (French) [41], jurBERT (Romanian) [78], and AraLegalBERT (Arabic) [2].

Legal-BERT - English Chalkidis et al. [21] explored several approaches for applying BERT models to downstream legal tasks evaluated on multiple datasets. They released Legal-BERT, a family of English BERT models intended to assist legal NLP research, computational law and legaltech applications. They also proposed new strategies for adapting BERT in specialised domains. Their research finds that the best strategy to port BERT to a new domain may vary, using either further pre-training on an existing model or pre-training from scratch [21].

For training, Chalkidis et al. [21]. scraped 12 GB of legal English data from various sources, containing legislation, court cases and contracts, gathered from public sources from the European Union, United Kingdom and United States [21].

The Legal-BERT models have the same architecture as BERT-Base from Devlin et al. with 12 layers, 768 hidden units, and 12 attention heads (110M parameters). They also experimented with a smaller Legal-BERT model with six layers, being 32% of the size of the size of BERT-Base [21].

Legal-BERT is trained for 1M steps, comprising almost 40 epochs over all their training corpora. The data is trained in batches of 256 samples, including up to 512 sentence-piece tokens. An AdamW optimiser with a learning rate of 1e-4 is used. The models are trained using the official BERT code using v3 TPUs from Google Cloud Compute Services.

For fine-tuning, Chalkidis et al. [21] did not follow Devlin et al.'s suggestions, which are often blindly followed in the literature, but considered an additional lower learning rate of 1e-5 and an additional higher drop-out rate of 0.2. They also used early stopping based on validation loss without a fixed number of training epochs.

Chalkidis et al. [21] do not mention using different training techniques to develop their models. Since they state that they have trained all their models using the official BERT code, they are assumed to be trained on the MLM and NSP tasks.

JuriBERT - French Douka et al. [41] explored using smaller architectures in domain-specific sub-languages and their benefits for the French language. They investigated creating a language model adapted to French legal text to help law professionals, releasing JuriBERT, a set of BERT models adapted to the French legal domain [41].

Douka et al. [41] used two different French legal text datasets, with a combined size of 6.3 GB, for training. The first set contains data from legal French documents scraped from the Légifrance website containing raw French legal text. The second set consists of 123.361 documents with French court decisions.

In their research, they trained a BERT model from scratch and further pre-trained a general French BERT model [41]. They further pre-trained CamemBERT Base from Martin et al. [77]. The pre-trained from scratch models had four different architectures to compare the difference, comprising training on 6M and 110M parameters.

The dataset is tokenised using a RobertaTokenizer with a maximum length of 512 per token and a minimum token frequency of 2. The total vocabulary is restricted to 32.000 tokens. The models were trained on the MLM training technique.

All models were pre-trained using a learning rate of 1e-4 along with an AdamW optimiser with $\beta_1 = 0.9, \beta_2 = 0.999$, and a weight decay of 0.1. They used a linear scheduler with 10,000 warm-up steps. The three smallest models were trained with a batch size of 8. Due to hardware limitations, the Base model and the further pre-trained CamemBERT model were trained with a batch size of 4. All models were trained on an Nvidia GTX 1080Ti GPU.

Douka et al. [41] found that the further pre-trained model outperformed the models developed from scratch. According to them, further pre-training of a general-purpose language model can have better results than training from scratch. The general purpose model outperformed the further pre-trained JuriBERT model on specific validation tasks, concluding that JuriBERT needs more pre-training corpora to perform better. The researchers state that acquiring large-scale legal corpora, especially for languages other than English, has proven challenging due to their confidential nature [41].

jurBERT - Romanian Masala et al. [78] employed the first study on the applicability of state-of-the-art NLP methods for Romanian legal judgment prediction. They introduced a Romanian BERT model pre-trained on a large specialized legal corpus, outperforming several strong baselines for legal judgment prediction on Romanian trial cases of banks in Romania [78].

Their dataset, provided by a Romanian bank, comprises original lawsuit documents with a raw size of 160 GB. Due to the larger pre-training corpus, they opted for full pre-training from scratch. They trained two variants, jurBERT-base and jurBERT-large, for which they stuck to the same model architecture and training procedure as suggested by Devlin [38].

The models are trained with a vocabulary of 33K tokens, with sequence lengths of 128 for 90% and 512 for the last 10%. Words were tokenized following the *Whole Word Masking* (WWM) principle. The models were trained using 40 epochs on the tasks of both MLM and NSP. For the learning rate, they found the best learning strategy to be a slanted triangular learning rate with a maximum learning rate of 2e-5 with a cutout of 0.1 and a ratio of 32. The training was performed on a v3-8 TPU provided by Tensorflow Research Cloud.

The models were compared to RoBERT, a general Romanian BERT model, and two standard CNN and BI-LSTM models with an attention mechanism. They found that their model outperformed the considered baselines [78]. However, on another task, jurBERT

is only slightly better than much simpler models as it struggles to handle long texts. According to Masala et al., [78], the limitations of BERT-like models regarding maximum input size significantly hampers their performance.

AraLegalBERT - Arabic Al-Qurishi et al. [2] examined how BERT can be used in the Arabic legal domain. They used several domain-relevant training and testing datasets to train a Legal Arabic BERT model from scratch called AraLegal-BERT. They evaluated their model against three BERT model variations for the Arabic language. They found that their developed model achieved better accuracy than the general and original BERT models over legal text.

They manually collected data from several sources and included many regional variations, as publicly available large-scale resources in Arabic legal text are scarce. They used legislative documents, judicial documents, contracts and legal agreements, Islamic rules, and the Islamic jurisprudence Fiqh, comprising a dataset of 4.5GB and 13.7 million sentences [2].

AraLegal-BERT was trained on 50 epochs and followed Devlin et al.'s original BERT pre-training procedure [38]. They used a batch size of 8, distributed in parallel over 8 GPUs, coming to a total batch size of 512. The maximum sequence length was set to 512, and the used learning rate ranged from 1e-5 to 5e-5. The model was trained on the MLM training procedure.

Al-Qurishi et al. [2] found that pre-training from scratch for a specific domain in the Arabic language is better than general models when developed for a specific task. The tested model of AraLegal-BERT was said to be a base, cost-efficient version suitable for a broad range of Arabic legal textual applications [2].

2.3.2 General Dutch BERT transformers

For our research, we use option 2 of the strategy introduced by Chalkidis et al. [21] to employ BERT for tasks in the legal domain: "Further pre-train BERT on domain-specific corpora" [21]. This section discusses three generic BERT models that can be used in Dutch: BERTje, RobBERT and mBERT. These models will be further pre-trained to create our family of legal-Dutch BERT models RechtBERT and serve as a foundation for our models.

BERTje De Vries et al. [33] explored using a monolingual Dutch BERT model to outperform an equally sized multilingual BERT model on downstream NLP tasks. They introduced BERTje, a generic Dutch monolingual BERT model. Their research finds that their model consistently outperforms the equally sized multilingual mBERT released by Devlin et al. [38] on Dutch NLP tasks.

For training, de Vries et al. [33] tried to train a Dutch model architecturally equivalent to the BERT model of Devlin et al. [38], trained with a dataset of similar size and diversity as used for the English BERT model. They combined several corpora, including a collection of fiction novels, Dutch news reports, Dutch Wikipedia pages and a multi-genre reference corpus, combining 12 GB of uncompressed text after cleaning.

BERTje was pre-trained on two objectives: MLM and *Sentence Order Prediction* (SOP). De Vries et al. recognized the shortcomings of the NSP task and chose to adopt SOP instead, in correspondence with the findings of Lan et al. [66]. They also applied a different strategy for the MLM objective. Instead of randomly masking single word pieces, they masked consecutive word pieces that belong to the same word.

The model has the same architecture as the original BERT model with 12 transformer blocks and is pre-trained on 1 million iterations. Fine-tuning was also evaluated at 850K

iterations. All models were fine-tuned for four epochs on the training data. Other training parameters are not disclosed.

RobBERT Delobelle et al. [36] also introduced a Dutch monolingual BERT model. They recognized the performance of BERTje by de Vries et al. [33], but tried a different architecture and training approach. Delobelle et al. [36] introduced RobBERT, a RoBERTa-based Dutch monolingual language model.

With the research, they evaluated their introduced RobBERT to BERTje. They also proposed several new tasks for testing the model's zero-shot ability and assessing its performance on smaller datasets and language-specific tokenizers. They found their model outperformed earlier models on several complex Dutch language tasks, significantly outperforming when dealing with smaller datasets. They suggest their model is a valuable resource for many application domains and can serve as a base for fine-tuning other tasks [36].

The data used for training is gathered from the Dutch section of the OSCAR corpus, a large multilingual corpus [36]. This corpus is 39 GB large, extensively larger than the corpora used for BERTje by de Vries et al. [33].

As the model is based on the RoBERTa architecture by [72], the model is only trained on the MLM training technique, disregarding Devlin et al. [38] suggested NSP technique, not replacing it with another training technique. The training process further uses the Adam optimizer with polynomial decay of learning rate 1e-6 and a ramp-up period of 1000 iterations, using $\beta_1 = 0.9, \beta_2 = 0.98$ with a weight decay of 0.1 and a dropout rate of 0.1 to prevent overfitting. The model was trained on a computing cluster, using at most 20 nodes of 4 Nvidea P100 GPUs per node, with a median node use of 5. The model trained for 2 epochs, equalling 16K batches in total.

mBERT The most prominent pre-trained language model is Bidirectional Encoder Representations from Transformers, which was released for the English language by Devlin et al. [38]. They also released a multilingual version of BERT called mBERT. This model is trained to understand 104 languages, including Dutch. Although mBERT performs well on many tasks, recent studies show that BERT models trained on a single language significantly outperform the multilingual version [36].

The model mBERT is solely trained on Wikipedia data and supports multiple languages simultaneously [38]. The model has the same architecture as the original BERT model, with 12 layers (Transformer blocks), 12 attention heads, a hidden size of 768, and total parameters of around 110 million. The model is trained on the MLM and NSP tasks.

Chapter 3

Modelling RechtBERT

This chapter describes the process of designing the legal Dutch BERT models RechtBERT. Three BERT models are further pre-trained on domain-specific corpora, following the suggestion of Chalkidis et al. [21] to improve BERT models for legal tasks. These three further pre-trained models are collectively called *"RechtBERT"*. The original models and the content of the family of models called RechtBERT are depicted in Figure 3.1.

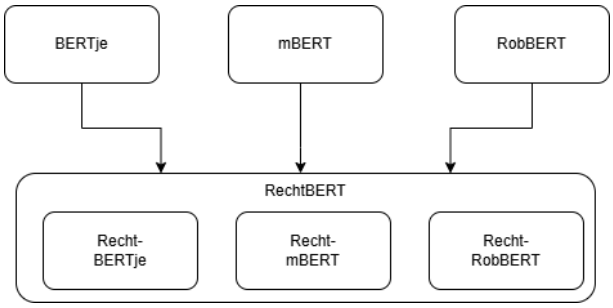

FIGURE 3.1: Family of RechtBERT models

In this section, the training data is described, and the cleaning process is explained, elaborating on the preparation of the data from retrieval to raw training data. Thereafter, the results of a statistical analysis of the sentence length of the data will be discussed.

In the second part of this chapter, the procedure of training is discussed. First, the training procedure and parameters of other legal BERT transcoders are discussed, whereafter, the training technique used for developing the legal Dutch BERT transcoders is disclosed, explaining the use of MLM and SOP. The tokenization process is described hereafter, and the training parameters are disclosed.

3.1 Data for Training

3.1.1 Data description

We gathered and combined two Dutch corpora with a legal background for training. First, we used the Basiswettenbestand, a Dutch database with 33.000 regulations on the Dutch legislature, comprising all valid Dutch laws enacted since the adoption of the Dutch Civil Code (Nederlands Burgelijk Wetboek) in 1838. Second, we retrieved court rulings from Dutch courts, comprising 452.771 full-text rulings.

Both corpora are part of the public domain. According to the Dutch copyright law (Auteurswet), "There is no copyright on laws, decrees, and regulations issued by public

authorities, nor on judicial decisions and administrative rulings." [85]. Therefore, the collected data can be freely copied and used for scientific and commercial usage, including training AI algorithms.

Dutch Legislature All of the Netherlands' legislation is gathered in a database called Basiswettenbestand. It comprises all regulations, including the Netherlands' treaties, laws, and ministerial regulations, excluding the "Caribbean Netherlands", a term mainly used to refer to the public entities of Bonaire, Sint Eustatius and Saba [59]. In total, 30.000 unique regulations are stored in this database. When changes are made to regulations, the new version is added to the dataset as a separate entry rather than replacing the existing version.

The dataset is sourced from *Kennis- en Exploitatiecentrum voor Officiële Overheidspublicaties* (KOOP), a Dutch governmental organization responsible for publishing all Dutch governmental data. KOOP operates under the Dutch Ministry of the Interior and Kingdom Relations (Ministerie van Binnenlandse Zaken en Koninkrijksrelaties). The dataset can be requested by email. The dataset contains approximately 85.000 versions of regulations, totalling around 150 GB of uncompressed storage space before cleaning.

Text of Dutch legislation is stored in XML files. In the dataset, not every file is an XML containing regulations, other files are also included. Additional files are sometimes necessary to clarify specific articles in the law. These supplementary files are associated with individual law articles and are listed by file type in Table 3.1.

Type	Number	Size
.PNG	297.081	95,26 GB
.XML	171.755	58,29 GB
.WTI	42.613	4,24 GB
.DB	27	815 KB
.XSD	2	5 KB

TABLE 3.1: Number of files in BWB dataset per extension

The dataset includes many PNG files representing images of formulas not expressed in plain text. Additionally, standardised forms required by law are stored as PNG files. Examples include diplomas with grading lists for high school students and application forms for immigrants. PNG files are also used for graphical representations, such as law requirements for house and boat measurements. Regulations also involve WTI files, custom files containing technical information about the law (wettechnische informatiebestanden).

Furthermore, XSD files are present, which are scheme definition files for XML that define the structure, constraints, data types, and relationships within the XML dataset. XSD files ensure data adherence to rules and consistency across instances. Thumbs.db files are system files created by Windows to cache thumbnail images. Lastly, not all XML files contain legislation. The dataset also includes additional XML files with manifest data about each regulation. Although useful for cleaning, these manifest files are discarded as they do not contain plain legislative text. In total, 42.613 XML files of 125,35 MB were XML manifest files.

Dutch Court Rulings Dutch court rulings and proceedings are published on the website of the Dutch court organisation Rechtspraak.nl. More than 1,4 million court rulings and proceedings are published on this website. Not all rulings are published with complete text; sometimes, only the information on a ruling is provided.

The complete dataset can not be requested or downloaded. Rechtspraak provides a RESTful web service to request information, which is provided in XML format. Two steps are required to extract full-text information. First, based on input criteria, IDs of court rulings can be retrieved. With these IDs, full-text versions of rulings can be requested. The Dutch court system uses the *European Case Law Identifier* (ECLI) for IDs of court rulings. ECLI serves as an identifier for court decisions across Europe and is used by national courts in European countries, as well as by the European Union, the Council of Europe, and the European Patent Office [43].

For our research, we retrieved 1.288.438 court rulings, of which 452.771 full-text versions are present. These are all rulings between the 1st of January, 2014 and the 31st of December, 2023. The collected data contains around 7.13 GB of uncompressed storage space before cleaning. The total number of ECLI and full-text content for each year is listed in Table 3.2.

Year	**# of ECLI**	**# of full text**
2014	154.453	29.159
2015	151.774	31.089
2016	145.083	31.519
2017	146.401	34.723
2018	141.134	37.170
2019	140.004	38.523
2020	133.911	46.921
2021	142.286	53.350
2022	138.073	55.953
2023	145.403	64.284
Total	**1.288.438**	**452.771**

TABLE 3.2: Number of ECLI and available full text

3.1.2 Data cleaning process

Before data can be used for training, it needs to be cleaned. This section describes how the datasets are transformed from raw collected data to one training set. [add methodology] The following steps have been performed in the data-cleaning process:

1. Removing unnecessary files
2. Remove non-Dutch texts
3. Extract and merge sentences
4. Clean sentences

The first three steps are performed for the two datasets separately for this list. Step four is equal for the legislation and the court rulings. This section will describe the steps in more detail.

Removing unnecessary files When collecting data, information that is not necessary to achieve the goal of the data collection is sometimes collected. In our research, we collect data to train an *Large Language Model* (LLM). In both datasets, some data is not needed

to achieve our goal. We want to clean all unnecessary files for our data, extract the raw texts of legislation and rulings and remove all other information.

As mentioned in the data description section, the legal legislation dataset also provides files to clarify specific articles in the law. Table 3.1 lists all types of files in the Basiswettenbestand dataset. The legislation text is stored in XML files. In total, 129.142 files comprising 58,17 GB of disk space contained legislation. All other files are discarded.

Within the remaining legislation files, there are many duplicate laws in the dataset with minor differences. Laws change over time. When a law is modified, a new version is added to the dataset while the old version remains intact. Only the most recent version of each law is kept to enhance training efficiency, and all previous versions have been removed.

Remove non-Dutch texts Not all legislature is written in Dutch. While almost all legislatures are written in Dutch, the Dutch government has also made some treaties with other countries. These treaties are not usually written in multiple languages. Luckily, a Dutch translation is often given when such a document is created in another language. However, they are both stored in the same XML file. Based on the manifest files, it can be determined if the document contained language other than Dutch. Non-Dutch texts were discarded, keeping the translated Dutch versions of the contracts. The treaty is removed from the dataset when no Dutch version is present.

All court rulings published by Rechtspraak are written in Dutch. No language cleaning was necessary for court rulings.

Extract and merge sentences We need stand-alone sentences for using text to train LLMs. Therefore, we collected all text data from the gathered files and combined them into one text file. Since most training procedures use TXT files for training, we also put all the data in this file format.

All laws, regulations, and rulings are stored in TXT files. The text is delimited by dots, marking the end of a sentence. After every law or ruling, a white line separates the different proceedings.

After extracting all the files, they were combined and ready for cleaning and training.

Clean sentences The following steps have been taken to clean the textfile to reduce noise for training purposes: This paragraph describes the cleaning process to clean the created dataset to reduce noise for training purposes. First, some cleaning steps have been performed on every sentence in the dataset:

- Removed all parentheses, brackets, and other non-alphabetic/numeric characters except space and punctuation.
- Replaced multiple whitespaces with a single whitespace.
- Removed all citation characters.
- Replaced semicolons with commas.

Besides cleaning on complete sentences, there also have been conducted some cleaning steps on the individual words:

- Converted words fully written in capitals to lowercase.
- Removed words not containing normal characters or numbers with commas.

- Removed words containing more than three numbers.
- Removed words that are two digits with one being a '.'.
- Removed [] from the beginning and end of words.
- Removed titles and names.

It was found that the titles of law employees caused many linebreaks while their names were not at the end of the line. Therefore, their names and titles were removed from the dataset.

3.1.3 Data analysis

A statistical analysis is performed on our merged dataset. Analysis is performed on the number of words in one sentence. The results of this analysis are used to determine the maximum tokenization length for the tokenization of the data, which is elaborated on in Section 3.2.2.

Our dataset contains 44.412.162 lines, of which 43.947.851 are sentences and 464.311 are white lines. The white lines highlight the number of files in our set, as they are used to separate the different laws and rulings. The length of every sentence is graphically shown in Figure 3.2, while the key descriptive statistics are shown in Table 3.3. The results do not contain sentences of one word, as these were cleaned during the cleaning process.

There are some extreme outliers. To limit the effect on statistics due to these outliers, sentences with a length larger than 100 words were capped at 101. Table 3.3 shows the statistical results for both the capped and the uncapped results. Figure 3.2 shows the frequency of the sentence length with length caption at 100, giving a sentence longer than 100 the value 101.

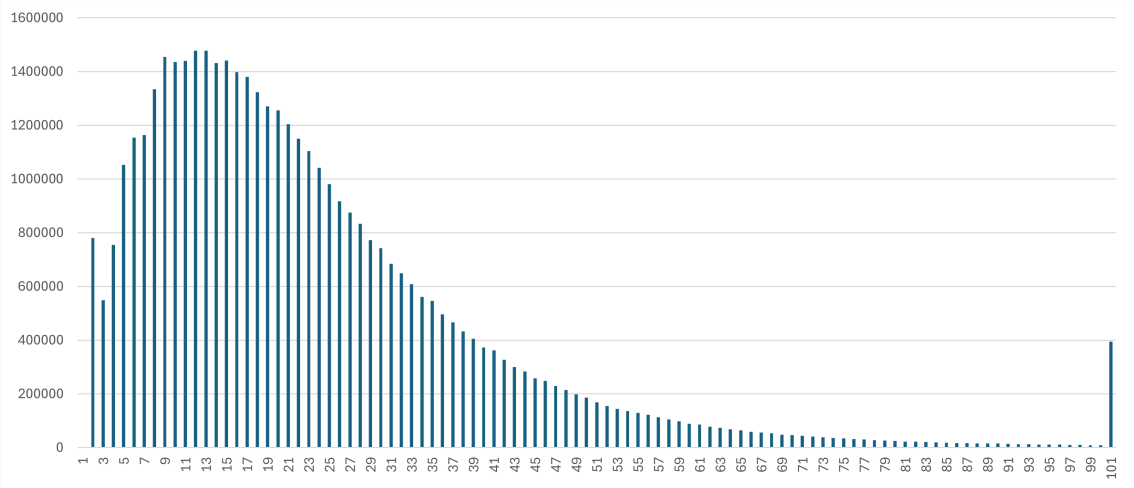

FIGURE 3.2: Frequency of sentence length within dataset

In both groups, the most common sentence length is 12 words, as shown by the mode. The average sentence length is consistent at 19 words, indicating that, on average, sentences are relatively short. This similarity suggests that including longer, unfiltered sentences has minimal impact on the overall average length. The median sentence length is 23,2 words in the capped group and 24,19 words in the uncapped group, further demonstrating that

	≤ 100	all
Median	23,2	24,19
Mean	19	19
Mode	12	12
St. Dev.	17,04	31,08
Minimum	2	2
Maximum	101	29.839
25th Percentile	11	11
Median	19	19
75th Percentile	30	30
95th Percentile	55	55

TABLE 3.3: Statistical analysis of sentence length of dataset

typical sentence length is short in both cases, with half of the sentences being under 24 words.

The distribution of sentence lengths in both groups skews towards shorter sentences, indicating that most sentences in the dataset are concise. The 25th percentile (11), 75th percentile (30), and 95th percentile (55) are the same in both groups, suggesting that the majority of the dataset is unaffected by the capping. The main difference arises in the longer sentences in the uncapped group, which are found beyond the 95th percentile. These identical percentile values highlight that the capping has minimal effect on most of the dataset, with differences primarily occurring in the upper extremes of the distribution.

However, filtering does affect the variability. The uncapped set has a standard deviation of 31,08, nearly double that of the filtered set, with a standard deviation of 17,04. As expected, the range differs significantly between the groups: in the capped group, the maximum is restricted to 101, while in the uncapped set, the maximum reaches 29.839. The extreme maximum length in the uncapped group indicates the presence of outliers, very long sentences that substantially increase both the range and standard deviation.

3.2 Procedure for Training

3.2.1 Training techniques

The original BERT model of Devlin et al. [38] was developed with two training objectives: MLM and NSP. These objectives force the model to learn semantic information within and between sentences [38]. The MLM randomly masks some of the tokens from the input. The objective is to predict the original, masked word-based based only on the context. The NSP task predicts if two sentences occur subsequent in the data corpus or if the second sentence is an entirely random sentence from the data corpus. With NSP the model is forced to learn semantic coherence between sentences.

After the initial release of the model, it was found that differentiations in this training procedure lead to better results [66, 72]. The NSP was intended to learn inter-sentence coherence, but BERT actually learned topic similarity [72]. If the next sentence is random, it is not just a matter of coherence, often the topic is likely different. That Devlin et al. [38] trained a better model when using NSP than without NSP is likely due to the model learning long-range dependencies that were longer than when just using single sentences [36].

Two strategies have been identified to overcome the NSP training problem. Liu et

al. [72] removed the NSP task from the pre-training process while creating RoBERTa, training only on the MLM task and using only full sentences in every input. Lan et al. [66] replaced the NSP task with the self-developed SOP task while creating ALBERT. In SOP, two sentences are either consecutive or swapped, determining whether two consecutive sentences appear in their correct order [33]. Using SOP should improve the model's ability to model sentence-level coherence and relationships, leading to better performance on downstream NLP tasks [66].

3.2.2 Training parameters

For our research, we have chosen to further pre-train existing BERT models, in line with possible strategies of Chalkidis [21]. As general BERT models for the Dutch language, we identified two models in literature: BERTje and RobBERT. We also further pre-trained mBERT of Devlin et al. [38]. This model contains many languages and is trained to understand Dutch. Also, this model was created by the original developer of BERT, giving it an ideal comparison of the workings of our models. This model is, therefore, also used in other research where BERT models have been developed.

We used the original libraries, which were used by the original models, to tokenize the data. For BERTje and mBERT, the 'BertTokenizer' and 'BatchEncoding' libraries from HuggingFace were used. Both models are further pre-trained on the MLM and SOP training techniques. For RobBERT, the 'RobertaTokenizer' and 'BatchEncoding' libraries from HuggingFace were used. Since RoBERTa-based models are only trained on MLM, we only transcode the data using the MLM technique for this model.

The data is tokenized per two consecutive sentences extracted from the data, with in between a token separator. For the SOP technique, the sentences are randomly swapped from position with a probability of 50% to be in the correct order or not. The two combined lines were capped at a maximum length of 100 tokens to optimize training time for limited training resources based on the average sentence length in the dataset. For the RobBERT model, two lines are combined but always kept in the correct order. 15% of the words were masked for the MLM training technique.

All models were trained with the same training arguments. They were trained in a batch size of 32 and saved in a checkpoint after every 2M sentences, giving 24 checkpoints. We used an Adam learning rate of 1e-5 with a weight decay of 0.01. All data was trained in 2 epochs. The models are trained on a single NVIDIA A40 (48GB) running in a High-Performance Cluster. The University of Twente made the hardware available without charge.

3.3 Training results

For training, three different models have been further pre-trained. While training, the cross-entropy loss for every batch was computed. This loss function is widely used in machine learning applications [76]. The loss has been calculated for every step and averaged for every epoch. For every epoch, these values are put in tables, which can be found in Appendix A.4. Graphs as summarizations of these results are depicted in this section. For BERTje, the results are shown in Table A.9 and Figure 3.3. For mBERT, the results are shown in Table A.10 and Figure 3.4. For RobBERT, the results are shown in Table A.11 and Figure 3.5.

The cross-entropy loss shows a trajectory that generally follows the expected pattern, with three noteworthy anomalies deviating from the typical curve.

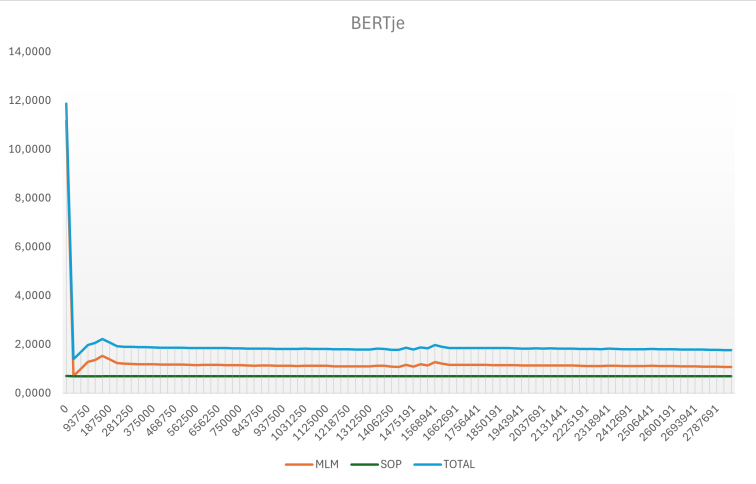

FIGURE 3.3: Cross-Entropy Loss BERTje

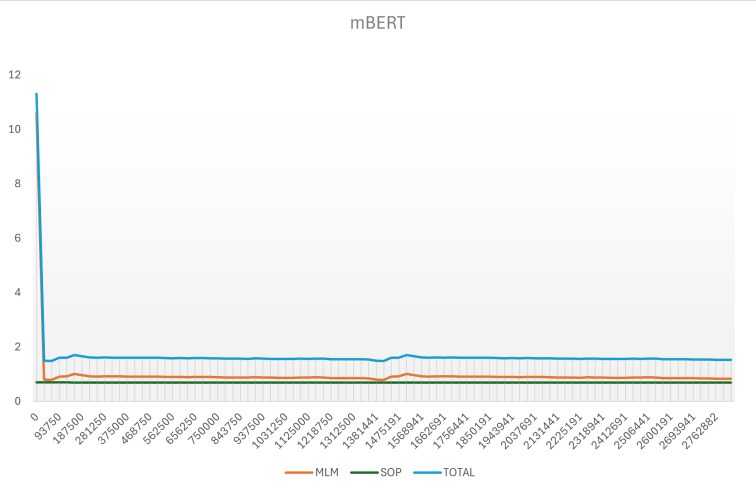

FIGURE 3.4: Cross-Entropy Loss mBERT

The first anomaly occurs at the start of training, where the loss for the MLM task drops steeply. For BERTje and mBERT, the loss falls below the lowest loss achieved after this initial phase. This behaviour can be attributed to the training configuration. Due to hardware instability, the models were set to automatically resume from the last checkpoint in case of a system crash, with checkpoints saved after every 2 million training sentences. As a result, no warmup steps were defined in the optimiser, leading to overfitting at the beginning of the training process. The loss subsequently increased again as more diverse data was introduced into the training pipeline.

The second anomaly is observed at the 1.412.691 training step for the MLM task for all three models. This behaviour can also be attributed to the training configuration. The models are trained for four epochs in two runs. Every batch of data is trained for two epochs, and the whole process is repeated after completion. The training restarts at the said training step, beginning at the start of the dataset. As the data is not shuffled, the model suddenly re-encounters regulations after a long period of only court rulings, which could explain the sudden extra loss.

Notably, the SOP task shows a remarkably stable loss quickly after the initialisation.

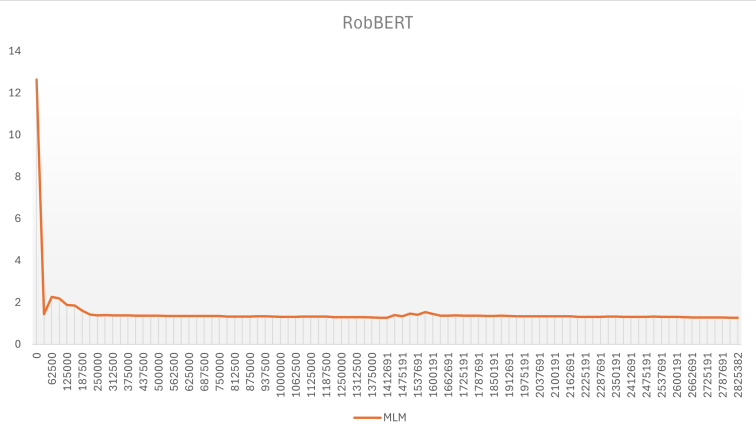

FIGURE 3.5: Cross-Entropy Loss RobBERT

This task is only performed for the BERTje and mBERT models. This indicates that the model calculates SOP very consistently, with the AdamW optimiser effectively preventing overfitting during this task, preventing the loss from growing. The validation phase will assess the impact of including this training task, as the RobBERT model does not use it.

Finally, it is essential to highlight that all three models have the same anomalies, suggesting that the observed anomalies came through the training configuration or used dataset rather than architectural differences. No significant architectural variations between RobBERT and the BERT-based models (BERTje, mBERT) were identified in these anomalies.

Chapter 4

Experiment & Validation

This chapter elaborates on the experiment to validate the performance of the three developed Dutch legal BERT models, collectively referred to as RechtBERT. First, it discusses the various validation tasks that can be performed to evaluate the models and the motivation behind the chosen approach. Next, it describes the experimental setup in detail. Finally, the chapter presents and discusses the results.

4.1 Validation Tasks

Multiple options were considered when selecting a suitable validation task to test our developed models. Unfortunately, not all tasks turned out to be feasible or implementable within the scope of this research. This section discusses two validation tasks identified as promising for enhancing legaltech. However, it could not be realized, whereafter the implemented validation option is explained.

4.1.1 Contract analysis & Document generation

Initially, this research aimed to implement an AI system capable of analyzing and drafting legal contracts. As identified in the literature review (Section 2.1), such AI applications are considered promising for advancing legaltech. However, this implementation was hindered by the unavailability of a sufficiently large and representative dataset.

Consistent with the findings of Douka et al. [41], obtaining legal datasets suitable for machine learning is challenging, especially for non-English languages. It was found that legal data can generally be categorized into three types:

- **Private data:** Documents generated by law firms and companies as part of their legal operations.
- **Published data:** Content provided by publishers, accessible to companies through subscriptions.
- **Public data:** Publicly available data, typically by government organizations or research institutions.

Private Data Private data are files generated by law firms or other legal businesses containing sensitive and confidential information. Due to privacy laws and the economic value of these documents, companies are reluctant to share such data. For this research, access to data from a major Dutch law firm was given for scientific use. Unfortunately,

the data was insufficient to train an AI model for generating contracts, highlighting that developing such an application solely for individual law firms is not feasible.

Published Data Published data are resources law firm employees use to access legal information, such as articles on specific laws, publications in law journals, and analyses of court decisions and new regulations. Law firms typically have subscriptions to these resources. In the Netherlands, a major publisher of such data is LegalIntelligence, which is accessible by subscription.

Unfortunately, the law firm's or university's subscription to LegalIntelligence did not allow using their data for this research. The provider, Wolters Kluwer, explicitly prohibits "automated processing" of their data in their user agreement, making it unavailable for training AI-powered systems [125].

Public Data Public data in Dutch is scarce. For this research, three datasets were identified as suitable. Two datasets published by the Dutch government consist of court rulings and Dutch regulations. Both were used for training RechtBERT (3.1). The third dataset, sourced from the *European Union* (EU), includes European regulations and treaties used for our classification validation task (Section 4.2.1). However, no public datasets containing contracts were found.

4.1.2 Legal search

Another option explored was the development of a legal search system. Legal search is widely regarded as one of the most promising applications of AI in the legal field. It is extensively used in research, as identified by our literature review (Section 2.1). The conducted expert reviews (Section 2.2) also highlighted this as a valuable direction, stating the current search functionality within the law firm's file system as "dreadful" or even "absent". While this option was deemed feasible, it was discarded due to constraints related to time, hardware availability, and strict limitations on data usage.

Discussions were held with a major Dutch law firm to validate this approach by developing a legal search system that their employees could use to assess its effectiveness. However, imposed restrictions on their use of private data resulted in significant challenges. For instance, the use of cloud services for temporary data storage was prohibited, particularly for non-EU-based services, due to security concerns. Furthermore, data access or storage by third parties, including the University of Twente, required a data processing agreement, even when the data was only stored temporarily for access by training hardware.

On-site training of the model was also not feasible due to the lack of available hardware or insufficient funding to hire needed hardware. Given these constraints and the limited timeframe, this option was set aside, leaving a potential gap for future research.

4.1.3 Legal Topic Identification

A classification task is often used to test the performance of BERT models [2, 23, 21, 33, 36, 41]. For all the discussed domain-specific legal BERT models of Section 2.3.1, only jurBERT (Romanian) was tested on a different task: judgement prediction [78]. The task that was used as a validation experiment is legal topic classification. A task performed by Chalkidis et al. [23, 21] to evaluate the performance of legalBERT [21] and multilingual domain-specific legal BERT models [23]. The task is further explained in the next section.

4.2 Experimental Setup

4.2.1 Validation Data

For the evaluation phase of our research, we used the MULTI-EURLEX dataset of Chalkidis et al. [23], a multi-lingual dataset for topic classification of legal documents. The dataset comprises 65.000 laws of the EU in 23 official EU languages, as EU laws are published in all official EU languages, except Irish and including English [20, 23]. Irish is not included as it was not made available by the EU at the moment of data gathering due to resource-related reasons [23]. English remains an official and working language of EU institutions, even though only 1% of EU citizens are native English speakers after the United Kingdom left the EU in 2020 [39].

Each law has been annotated with EuroVoc concepts (labels). EuroVoc is a multi-lingual, multidisciplinary thesaurus covering the activities of the EU, containing terms in the 24 EU languages [94]. These concepts are assigned by the *Publications Office of the European Union* (OP). EuroVoc has eight levels of concepts, where every document is assigned one or more labels. The original documents have been assigned with a label of levels 3 to 8. Chalkidis et al. [23] created three alternative sets of labels per document by replacing each assigned concept by its ancestor from level 1, 2, or 3, respectively.

The dataset has not undergone an extensive cleaning process. The text of all articles has been left in its original form without any preprocessing. However, all documents in 22 non-Dutch languages were removed, resulting in 65,000 documents. These documents were divided by Chalkidis et al. [23] into training, test, and validation sets of 55.000, 5.000, and 5.000 documents, respectively.

Our classification task focused only on the ten most frequent labels from the Level 1 label set. The Level 1 labels, corresponding descriptions, and frequencies in the Dutch dataset are provided in Appendix Table A.12. This approach was chosen because test runs showed significant challenges in predicting less frequent labels due to the highly imbalanced availability of each label, which severely impacted model performance.

4.2.2 Experiment approach

With the introduction of BERT, Devlin et al. [38] proposed an approach for determining the most optimal parameters for fine-tuning based on a search within a limited range. The learning rate, length of training, size of training batch and dropout range are all fixed or can take one of a few possible ranges [2, 38]. While no particular reason has been given for this approach, it has been widely replicated in studies with BERT derivatives [2]. In our experiments, we adhere to the guidelines of Devlin et al. [38]. The strategy is determined based on these guidelines, hardware limitations and training findings of other legal BERT derivatives [2, 23, 21, 41, 78].

Our experiment is designed to train, validate, and evaluate a multi-label classification model using six pre-trained BERT models: BERTje, mBERT, and RobBERT, along with their further pre-trained counterparts Recht-BERTje, Recht-mBERT and Recht-RobBERT, respectively. The text data is tokenized using the 'BertTokenizer' for BERTje, mBERT and their counterparts and 'RoBERTaTokenizer' for RobBERT and its counterpart. The sequence length is set to 512, which is the maximum the models support. The labels are converted to a multi-hot encoding format to support multi-label classification with dynamically determined optimal thresholds.

The model was trained and validated for five epochs using binary cross-entropy with logits as the loss function, a learning rate 2e-5, and the AdamW optimizer to prevent

overfitting. Additionally, a warmup phase of 5,000 steps was applied during training. These settings were inspired by the validation task of Douka et al. [41], who evaluated their models on a multi-class classification task with eight highly imbalanced classes. Their further pre-trained model demonstrated the best performance with a learning rate of 2e-5. However, they noted only minor differences across other selected learning rates within the range of the strategy of Devlin et al. [38, 41].

We used the AdamW optimizer for training to prevent overfitting as all legal BERT implementations consistently used the AdamW optimizer [2, 21, 41, 78]. We opted for five training epochs because, during preliminary testing, the loss on the fifth epoch was nearly identical to that of the fourth. In contrast, a significant improvement was observed between the third and fourth epochs.

To determine which models are best, we looked at the weighted F1-Score. The F1-Score is the harmonic mean of the precision and recall rates. Precision is the ratio of true positives to all positives, which minimizes false positives. At the same time, recall is the ratio of true positives to all correctly classified messages, which reduces false negatives. The F1-score balances these metrics to evaluate overall performance.

To improve the metrics for all the models, we implemented threshold optimization. Instead of relying on a fixed decision threshold, the optimal threshold is dynamically determined for each label using the validation dataset after each epoch. It evaluates a range of possible thresholds (0.25 to 0.75, with a step of 0.05) and selects the threshold that maximizes the F1 score for each class. This approach balances precision and recall for all labels individually, ensuring that the model's predictions align more closely with expected values.

The same hardware was used during the training phase of this research for this experiment.

4.3 Experiment Results Analysis

In this experiment, we conducted a multi-label classification task using six different variants of Dutch BERT models to classify legal topics in EU documents. The primary goal was to compare the newly developed RechtBERT family of models with their original general counterparts, assessing whether the further pre-trained models, trained explicitly on legal data, achieve higher accuracy in classifying topics within EU documents. The results of this experiment are presented in Table 4.1. This section analyzes the results of the performance of the RechtBERT models compared to the original models.

	BERTje (Scratch)	Recht-BERTje	mBERT (Scratch)	Recht-mBERT	RobBERT (Scratch)	Recht-RobBERT
wPrecision	0,8358	0,6022	0,8307	0,6535	0,8514	0,7210
wRecall	0,8502	0,7573	0,8484	0,8463	0,8418	0,8099
microF1	0,8374	0,6780	0,8345	0,6923	0,8403	0,7410
macroF1	0,8089	0,5930	0,8064	0,6625	0,8148	0,6970
weightedF1	0,8404	0,6580	0,8364	0,7167	0,8424	0,7537

Table 4.1: Experiment Results: Precision, Recall and F1-Scores

The results presented in Table 4.1 have been analyzed, leading to several key observations. Regarding performance metrics, legal topic classification can be effectively performed using Dutch BERT models. The original RobBERT model stands out as the most suitable

option for this task, achieving a weighted F1 score of 0,8424, closely followed by the original BERTje model, with a weighted F1 score of 0,8404.

Additionally, it can be concluded that the original models consistently outperform their further pre-trained counterparts, collectively called the RechtBERT models. The difference is most shown between BERTje and Recht-BERTje, with BERTje having a weighted F1 score of 0,8404 compared to Recht-BERTje's 0,6580. The domain-specific fine-tuned models exhibit lower weighted precision scores, indicating that the RechtBERT models often introduce numerous false positives.

In analyzing model-specific trends, it is clear that the original RobBERT outperforms both the original mBERT and BERTje models. Among the RechtBERT family, Recht-RobBERT exhibits the best performance. However, other models excel in weighted recall, with BERTje outperforming RobBERT and Recht-mBERT, surpassing Recht-RobBERT.

The results also show a notable trade-off between precision and recall in the RechtBERT models. The models show a decent weighted recall, with Recht-mBERT outperforming the original RobBERT. However, this recall score was accompanied by a significant drop in weighted precision. This indicates that the models are lenient in classifying something as positive, capturing more true positives, and often classifying entries incorrectly.

4.4 Discussion

In this study, we researched the use of AI in legaltech by creating domain-specific Dutch BERT models for use in the legal domain. We introduced a family of Dutch BERT models called RechtBERT. These models were created by further pre-training three existing generic BERT models already trained to understand the Dutch language. The pre-training followed one of the recommended strategies suggested by Chalkidis et al. to enhance BERT models before fine-tuning them for specific tasks [21]. The training dataset consisted of Dutch laws and court rulings.

An experiment was conducted to evaluate the performance of the RechtBERT models compared to the original model from which they were trained. This experiment involved a multi-label classification task to classify legal topics within a EU dataset. The results of this task were analyzed, and the findings are presented in Section 4.3. The RechtBERT models do not show a higher performance than the original models. This section discusses the implications of this finding.

The RechtBERT models are created to enhance legaltech by using AI. A conducted literature review showed that many AI applications are built upon the BERT transformer architecture introduced by Devlin et al. It is claimed that generic BERT models underperform in specific domains, including legal. However, the experiments show that generic Dutch BERT models do not underperform, showing robustness when applied to complex legal classification tasks. Therefore, it can be debated whether the claim of underperforming is also valid for the Dutch language.

When discussing why the RechtBERT models underperform in comparison to using generic BERT models out of the box, we need to look at the design choices made. When comparing the training of our model to other legal BERT models, which have been discussed in Section 2.3.1, some design choices have been made different to the original procedure as suggested by Devlin et al. [38] and the results of designing of other monolingual legal BERT models.

One design choice that distinguishes RechtBERT from other legal BERT models is its sequence length. Unlike other models, where the architecture's capacity often determines sequence length, RechtBERT's sequence length is based on the length of two sentences in

the dataset. A minimum of two consecutive sentences is required for the SOP training task. As analyzed in Section 3.3, the average sentence length in the dataset led us to set the training sequence length to 100 tokens. This is substantially shorter than the sequence lengths typically used by other models, which are often 512 or 256 tokens. However, the shorter sequence length of training tokens limits the model's ability to train on broader contextual information, as fewer words are available for interpretation.

Another important factor to consider is the training data used for RechtBERT. Unlike other models trained on datasets specifically designed for their respective languages, RechtBERT was trained on a combination of court rulings and legal texts. However, the dataset is imbalanced, with court rulings significantly outnumbering legislative laws. While court rulings address legal matters, they do not represent formal legal texts. This discrepancy may affect the model's understanding of legal language. This imbalance could explain why RechtBERT may be less suitable or effective for specific legal classification tasks.

The experimental setup may have contributed to the observed differences in performance. The evaluation was conducted using a new EU dataset, distinct from the dataset used for training. While this approach allows for a fairer comparison with general-purpose models, it could put RechtBERT at a disadvantage. The dataset comprises EU legislative documents, many of which are translations, and it may have a different formulation style compared to traditional legal texts.

Additionally, the experiment followed the methodology established by Chalkidis [22], using a sequence length of 512 tokens. Generic models were trained with this sequence length, but RechtBERT was not, likely negatively impacting its performance. Furthermore, RechtBERT aims for comprehensive coverage by including all relevant categories. While this results in more accurate identification of positive labels, it also leads to a higher rate of false positives. This trade-off may be beneficial for legal classification tasks, where the priority is to avoid missing critical information.

Chapter 5

Conclusion & Future work

This chapter presents the general conclusions of the thesis, starting with discussing the research questions and their answers in Section 5.1. Subsequently in Section 5.3, we discuss the limitations of our research and recommendations for future work.

5.1 General Conclusions

This thesis introduces RechtBERT, a set of domain-specific legal BERT models to enhance legaltech. Three generic Dutch BERT models (BERTje [33], RobBERT [36], mBERT [38]) are further pre-trained on Dutch legal data. The created models are validated by performing a multi-label legal topic classification task, whereafter the performance of the models is compared with their original counterparts. The methodology used for this design research is the Design Science Research Methodology of Peffers et al. [102]. To develop RechtBERT, we answered six research questions, whereafter the main research problem can be answered. These questions are listed and discussed in this section.

RQ1: *What is the current state of research on the use of AI in legaltech?*

A SLR has been conducted to gain a deep understanding of the use of AI in the legal domain. In total, 94 papers were found and analysed. The research shows a growing interest in the use of AI in legaltech. Five distinct categories were identified: publication trends, legal domains, AI applications, AI technologies and benefits, challenges and limitations.

The SLR provides valuable insights into the application of AI in legaltech, offering a strong foundation for future research in this domain. Beyond establishing a background, several key findings shaped the direction of our study.

Analysis of identified AI technologies revealed a clear trend in NLP, with the BERT language model standing out as the most widely used model. It appears in various derivations, each adapted to specific purposes within the legaltech domain. Moreover, the review highlighted significant potential for AI applications in areas such as legal search and eDiscovery, document generation and contract analysis.

Regarding the legal domains explored, many studies offered a general perspective on the legal field, not specifying specialization in a particular domain. However, most papers focus on a single domain, concentrating on lawyers and law firms. These findings collectively formed the basis of our research trajectory.

All findings from the SLR are relevant to answering **RQ1** and are detailed in Section 2.1.

RQ2: *Which AI integration opportunities address the challenges faced by legal professionals?*

Experts have gathered input to adapt theoretical findings from the SLR to real-world applications and legal professionals from a major Dutch law firm have been interviewed. All insights have been summarized by topics and are represented in Table 2.1, which is discussed in Section 2.2.

The table shows lawyers identifying that many documents they create have overlap. However, not all departments use standardized models, as some do not invest in making these models, while others are reluctant to share their models with colleagues, leaving an opportunity for AI-powered document generation.

Lawyers recognize the value of published data and can effectively find needed information. However, search capability is insufficient when searching through the firm's private data, leaving an opportunity for AI-powered legal search.

RQ3: *What insights from existing BERT-based models can guide the design and training of RechtBERT for optimal performance in legal NLP tasks?*

The SLR identified the BERT language model and its derivatives as the most widely used models in legal technology. Section 2.3 further discusses BERT models, providing elaboration and comparisons of domain-specific legal BERT models alongside generic Dutch BERT models. The results of this analysis form the foundation for our design choices in developing the RechtBERT model, including the training techniques and parameters used, answering **RQ3**.

Chalkidis et al. [21] outlined three strategies for using BERT in NLP tasks within the legal domain: (a) using BERT as-is, (b) further pre-training BERT on domain-specific corpora, and (c) pre-training BERT from scratch using domain-specific corpora. Four domain-specific legal BERT models have been identified: Legal-BERT for English [21], JuriBERT for French [41], jurBERT for Romanian [78], and AraLegalBERT for Arabic [2]. Both strategies (b) and (c) for English and Arabic demonstrate significant performance improvements. However, for French and Romanian, the performance gains are less pronounced. The Romanian model improves slightly, whereas further pre-training for French does not outperform the original French CamemBERT implementation.

Three generic Dutch BERT models are reviewed: BERTje, RobBERT, and mBERT. Among these, BERTje and RobBERT are entirely monolingual models designed explicitly for Dutch. At the same time, mBERT is a multilingual model capable of understanding Dutch. The models differ in their training objectives: RobBERT is trained solely with the MLM technique, whereas mBERT employs both MLM and NSP, and BERTje uses SOP, a successor to NSP.

Besides, the training strategies for all identified models have been listed and used to determine the training strategy for the RechtBERT models.

RQ4: *How can the identified characteristics of BERT models guide the development of a legal Dutch BERT model for NLP tasks?*

In Chapter 3, the modelling process for the Dutch legal BERT models RechtBERT is described. Two datasets were used for training: legislative texts from the "Basiswettenbestand" dataset provided by KOOP and court rulings from Rechtspraak.nl. After cleaning, the combined dataset comprised 43.947.851 sentences. An analysis of sentence length was conducted to determine the optimal sequence length for training.

The chapter also discusses the training techniques employed, elaborating on MLM and SOP. Training parameters were chosen based on insights from other legal BERT models. The training results are analyzed, focusing on the cross-entropy loss, which revealed three anomalies: a sharp drop in loss shortly after training initiation, an additional loss spike when restarting training at the beginning of the unshuffled dataset, and a stable loss trajectory for the SOP technique. Since all models show the same anomalies, it is concluded that these issues came from the training configuration or dataset used rather than architectural differences between the models.

> ***RQ5:*** *What NLP applications can be used to validate the performance of domain-specific legal Dutch BERT models?*

Based on the findings from **RQ1** and **RQ2**, two potential validation tasks were identified: contract analysis & document generation and legal search. However, these tasks could not be executed during this research. Document generation was deemed unfeasible due to a lack of sufficient data. Publicly available data, particularly contracts, is minimal, and user agreements often prohibit the automated processing of published documents. Additionally, the dataset from a single large law firm was insufficient to develop a document generation application tailored exclusively for that firm. Legal search on private data was also deemed not feasible due to hardware and data processing limitations.

It was found that other researchers often use classification tasks to validate BERT models. Therefore, we conducted a legal topic classification task on a mult-label dataset by Chalkidis et al. [23]. This dataset contains 65.000 legal documents of the EU. Section 4.1 elaborates the discussion on different validation tasks.

> ***RQ6:*** *How does the domain-specific legal Dutch BERT model's performance compare to that of generic Dutch BERT models on previously selected tasks?*

All our developed models underperform compared to their original counterparts. Second, the original Dutch BERT models demonstrate strong performance on this legal classification task. Thirdly, the RechtBERT models have low weighted precision scores, indicating that the models frequently misclassify texts as belonging to certain legal topics when they do not.

> **Main research problem:** ***How to*** *design a domain-specific legal Dutch BERT model* ***that*** *outperforms generic Dutch BERT models* ***so that*** *legal professionals can perform tasks more efficiently* ***in*** *the advancement of legaltech through NLP applications?*

Unfortunately, we were unable to design a model that outperformed existing ones. Developing such a model is challenging due to several limitations, including the availability of data and adequate hardware resources for training. Additionally, the impressive performance of existing general Dutch BERT models is worth noting.

The results from the validation phase suggest that legal Dutch language may not be as distinct from general Dutch as initially assumed. Given the high performance of general models, the linguistic differences between the two appear to have less impact than expected.

These findings align with the results of jurBERT [78], where the specialized legal model showed minimal improvement over the generic language model. In the case of JuriBERT [41], further pre-training of the BERT model even resulted in worse performance for some tasks compared to the generic model. However, the JuriBERT model trained from scratch demonstrated better performance, raising whether similar improvements could be achieved if RechtBERT were trained from scratch using the same dataset, leaving an opportunity for future work.

5.2 Research Contribution

Contribution to science This research contributes to science by addressing a gap in the literature on the development and use of domain-specific legal Dutch BERT models. No legal Dutch BERT model has been found in the literature. This study fills that gap by introducing RechtBERT, a family of Dutch legal BERT models further pre-trained on existing general Dutch BERT models.

Furthermore, this study compares general monolingual BERT models with domain-specific BERT models in the Dutch language, contributing to the question of whether generic BERT models underperform in comparison to domain-specific models.

Besides, the literature review conducted in Section 2.1 of this research shows a growing interest in the use of AI in legaltech. The findings of this research can serve as a foundational source for future investigations into the use of AI within the legal domain.

Contribution to practice This research explores several applications for legal BERT models, which can be used for contract generation and document analysis. It shows that a significant dataset is needed to fit the model, which is difficult to gather, even for large law firms. Furthermore, these models can be used for legal searches, but this implementation has limitations.

Legal BERT models such as RechtBERT have significant potential when the limitations in this research are lifted. The literature review reveals numerous applications for AI in legaltech. Some of these implementations have been further examined in this research, but more applications are possible. The RechtBERT models can be utilised to achieve these implementations, thereby increasing efficiency in the legal workflow.

5.3 Limitations & Future Work

Limitations This research has several limitations. As discussed earlier in the paper, finding suitable legal data published in Dutch was challenging. We identified and used three data sources, all made available from the public domain. Some potential applications of AI could not be developed due to the lack of valuable training data. Additionally, although our dataset for developing the RechtBERT models was sufficiently large, it lacked variety in the types of texts and sources. Gathering other datasets, possibly from private or published domains, could enhance the performance of the models.

The limitation on available data also affected the tasks that could be used to validate the RechtBERT models. Within the scope of this research, we could not explore whether BERT models can be used for document generation and contract analysis. Moreover, the options for legal search were limited due to hardware and security constraints. Future research could explore these applications.

Hardware limitations influenced some design choices. Initially, the tokenization process did not fit in the memory of the GPU, which led to a reduction in sequence length. This was adjusted to align with the average sentence length, ensuring that only two sentences were used per training instance. The sequence length is typically fixed, with sentences being filled or truncated with tokens. This alternative approach may have significantly impacted the model's performance.

Another limitation is that the design cycle was completed only once, without the iterative process that typically allows for improvements to the artefact, in this case, the RechtBERT models. We did not modify the training strategy after the validation results became clear. Since training a model like RechtBERT can take up to five full days, it was

not feasible to iterate this process frequently. Therefore, it would be beneficial to assess whether modifying the training strategies could yield better results for the RechtBERT models.

Future Work Future work could focus on refining the RechtBERT models by optimizing training parameters and experimenting with varying sequence lengths of tokens to enhance their performance further. Additionally, retraining the models with a more balanced dataset by incorporating a more significant proportion of actual legal texts, such as statutes and case law, could improve the models' contextual understanding.

Another promising direction for future research is developing a new RechtBERT model trained entirely from scratch. Chalkidis et al. [21] outlined three approaches to adapting BERT for the legal domain. While our research explored two of these approaches, namely using BERT out of the box and further pretraining existing BERT models on domain-specific corpora, the results showed that the first approach had better performance. Investigating the third option, which involves building a domain-specific BERT model from scratch, would significantly increase the contribution to science.

Lastly, efforts could focus on implementing RechtBERT in real-world legal practices. While this research did not gather sufficient data to explore direct applications within law firms or other legal environments, future studies could fill this gap by obtaining relevant datasets and testing RechtBERT in practical scenarios. Such implementations would validate the model's effectiveness in real-world settings and increase its contribution to practice.

Bibliography

[1] S. Agrawal, A. Sahu, and G. Kumar. A conceptual framework for the implementation of industry 4.0 in legal informatics. *Sustainable Computing: Informatics and Systems*, 33, 2022.

[2] M. AL-Qurishi, S. AlQaseemi, and R. Soussi. AraLegal-BERT: A pretrained language model for Arabic Legal text. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 338–344, 2022.

[3] B. Alarie, A. Niblett, and A. H. Yoon. Law in the future. *University of Toronto Law Journal*, 66:423–428, 10 2016.

[4] B. Alarie, A. Niblett, and A. H. Yoon. How artificial intelligence will affect the practice of law. *University of Toronto Law Journal*, 68:106–124, 1 2018.

[5] J. Armour, R. Parnham, and M. Sako. Unlocking the potential of ai for english law. *International Journal of the Legal Profession*, 28:65–83, 1 2021.

[6] J. Armour, R. Parnham, and M. Sako. Augmented lawyering. *University of Illinois Law Review*, 2022:71–138, 2022.

[7] J. Armour and M. Sako. Ai-enabled business models in legal services: From traditional law firms to next-generation law companies? *Journal of Professions and Organization*, 7:27–46, 2020.

[8] K. D. Ashley. Case-based reasoning and its implications for legal expert systems. *Artificial Intelligence and Law*, 1:113–208, 1992.

[9] M. Ayodele, R. Allmendinger, and K. N. Papamichail. Heuristic search in legaltech: Dynamic allocation of legal cases to legal staff. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 12566 LNCS, pages 326–338, 2020.

[10] J. Barnett and P. Treleaven. Algorithmic dispute resolution-the automation of professional dispute resolution using ai and blockchain technologies. *Computer Journal*, 61:399–408, 2018.

[11] P. Baser and J. R. Saini. Ai-based intelligent solution in legal profession. In *ICT Systems and Sustainability. Lecture Notes in Networks and Systems*, volume 516, pages 75–84, 2023.

[12] A. Beebeejaun and R. P. Gunputh. A study of the influence of artificial intelligence and its challenges: The impact on employees of the legal sector of mauritius. *Global Business Review*, 2023.

[13] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*, 2019.

[14] D. Braun, E. Scepankova, P. Holl, and F. Matthes. Consumer protection in the digital era: The potential of customer-centered legaltech. In *Lecture Notes in Informatics (LNI), Proceedings - Series of the Gesellschaft fur Informatik (GI)*, volume 294, pages 407–420, 2019.

[15] J. A. Burt. The revolutionary impact of artificial intelligence on the future of the legal profession. *Kutafin Law Review*, 8:390–402, 2021.

[16] P. D. Callister. Law, artificial intelligence, and natural language processing: A funny thing happened on the way to my search results. *Law Library Journal*, 112:161–212, 2020.

[17] Salvatore Caserta. Digitalization of the legal field and the future of large law firms. *Laws*, 9(2), 2020.

[18] Salvatore Caserta. New technologies and law firms-an uneasy relationship: A european perspective. *Law, Technology and Humans*, 4(2):183–196, 2022.

[19] Salvatore Caserta and Mikael Rask Madsen. The legal profession in the era of digital capitalism: Disruption or new dawn? *Laws*, 8(1):1, 2019.

[20] EUR-Lex - 01958R0001-20130701 - EN - EUR-Lex, November 2024. [Online; accessed 12. Nov. 2024].

[21] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos. Legal-bert: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pages 2898–2904, 2020.

[22] I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, D. M. Katz, and N. Aletras. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 4310–4330, 2022.

[23] Ilias Chalkidis, Manos Fergadiotis, and Ion Androutsopoulos. Multieurlex–a multi-lingual and multi-label legal document classification dataset for zero-shot cross-lingual transfer. *arXiv preprint arXiv:2109.00904*, 2021.

[24] G. Chandra, R. Gupta, and N. Agarwal. Role of artificial intelligence in transforming the justice delivery system in covid 19 pandemic. *International Journal on Emerging Technologies*, 11:344–350, 2020.

[25] S. Chen, J. Wang, and Q. Zhang. Informetric analysis of researches on application of artificial intelligence in legal practice. In *Proceedings of the International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE 2023*, pages 406–408, 2023.

[26] K. R. Chowdhary. *Fundamentals of Artificial Intelligence*. Springer, India, April 2020.

[27] Michael Chui, Bryce Hall, Helen Mayhew, Alex Singla, and Alex Sukharevsky. The state of AI in 2022 - and a half decade in review. *McKinsey & Company*, December 2022.

[28] G. M. Csányi, R. Vági, D. Nagy, I. Üveges, J. P. Vadász, A. Megyeri, and T. Orosz. Building a production-ready multi-label classifier for legal documents with digital-twin-distiller. *Applied Sciences (Switzerland)*, 12, 2022.

[29] E. da Luz Scherf, M. Silva, and J. Silva. In tech we trust? some general remarks on law in the technological era from a third world perspective. *Journal Juridical Opinion*, 17:107–123, 05 2019.

[30] R. Dale. Law and word order: Nlp in legal tech. *Natural Language Engineering*, 25:211–217, 2019.

[31] H. Darji, J. Mitrović, and M. Granitzer. Exploring semantic similarity between german legal texts and referred laws. In *Communications in Computer and Information Science*, volume 1718 CCIS, pages 37–50, 2023.

[32] E. L. de Siles. Ai, on the law of the elephant: Toward understanding artificial intelligence. *Buffalo Law Review*, 69:1389–1469, 2021.

[33] Wietse de Vries, Andreas van Cranenburgh, Arianna Bisazza, Tommaso Caselli, Gertjan van Noord, and Malvina Nissim. Bertje: A dutch bert model. *ArXiv*, pages 1912–09582, 2019.

[34] S. Delacroix. Computer systems fit for the legal profession? *Legal Ethics*, 21:119–135, 2018.

[35] F. Delgado, S. Barocas, and K. Levy. An uncommon task: Participatory design in legal ai. *Proceedings of the ACM on Human-Computer Interaction*, 6, 2022.

[36] Pieter Delobelle, Thomas Winters, and Bettina Berendt. Robbert: a dutch roberta-based language model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3255–3265, 2020.

[37] Jianyang Deng and Yijia Lin. The Benefits and Challenges of ChatGPT: An Overview. *FCIS*, 2(2):81–83, 2022.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[39] Directorate-General for Communication, European Union. Languages in the european union, n.d. Accessed: 2024-11-12.

[40] J. Donnelly and A. Roegiest. The utility of context when extracting entities from legal documents. In *International Conference on Information and Knowledge Management, Proceedings*, pages 2397–2404, 2020.

[41] Stella Douka, Hadi Abdine, Michalis Vazirgiannis, Rajaa El Hamdani, and David Restrepo Amariles. Juribert: A masked-language model adaptation for french legal text. *arXiv preprint arXiv:2110.01485*, 2021.

[42] M. D'Rosario. Fair use defences during copyright litigation: Is the success of a fair use defence strategy predictable? *International Journal of Strategic Decision Sciences*, 8:31–51, 4 2017.

[43] European e-Justice Portal. European e-Justice Portal - European Case Law Identifier (ECLI), may 2019.

[44] O.A. Alcántara Francia, M. Nunez del Prado, and H. Alatrista-Salas. Survey of text mining techniques applied to judicial decisions prediction. *Applied Sciences (Switzerland)*, 12, 2022.

[45] D. Gingras and J. Morrison. Artificial intelligence and family odr. *Family Court Review*, 59:227–231, 2021.

[46] G. Gordon, B. Rieder, and G. Sileno. On mapping values in ai governance. *Computer Law and Security Review*, 46, 2022.

[47] P. Gowder. Transformative legal technology and the rule of law. *University of Toronto Law Journal*, 68(supplement 1):82–105, 01 2018.

[48] W. Gravett. Is the dawn of the robot lawyer upon us? the fourth industrial revolution and the future of lawyers. *Potchefstroom Electronic Law Journal*, 23:1–37, 2020.

[49] A. Harzing and S. Alakangas. Google scholar, scopus and the web of science: a longitudinal and cross-disciplinary comparison. *Scientometrics*, 106(2):787–804, November 2015.

[50] R. Hilhorst and T. Van Engers. E-dossier at the dutch council of state: Design, implementation and lessons learned. In *CEUR Workshop Proceedings*, volume 582, pages 13–32, 2009.

[51] K. Hilt. What does the future hold for the law librarian in the advent of artificial intelligence? *Canadian Journal of Information and Library Science*, 41:211–227, 09 2017.

[52] C. Hogan, R. Bauer, and D. Brassil. Human-aided computer cognition for e-discovery. In *Proceedings of the International Conference on Artificial Intelligence and Law*, pages 194–201, 2009.

[53] L. Humphreys, G. Boella, L. Di Caro, L. Robaldo, L. van der Torre, S. Ghanavati, and R. Muthuri. Populating legal ontologies using semantic role labeling. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pages 2157–2166, 2020.

[54] D. Hunter. The death of the legal profession and the future of law. *University of New South Wales Law Journal*, 43:1199–1225, 2020.

[55] E. Iatrou. A normative model of explanation for binary classification legal ai and its implementation on causal explanations of answer set programming. In *CEUR Workshop Proceedings*, volume 3193, 2022.

[56] A. Ivaschenko, O. Golovnin, I. Syusin, A. Krivosheev, and M. Aleksandrova. Ontology based text understanding and text generation for legal technology applications. In *Lecture Notes in Networks and Systems*, volume 739 LNNS, pages 1080–1089, 2023.

[57] S. Jayasinghe, L. Rambukkanage, A. Silva, N. de Silva, and A. S. Perera. Legal case winning party prediction with domain specific auxiliary models. In *ROCLING 2022 - Proceedings of the 34th Conference on Computational Linguistics and Speech Processing*, pages 205–213, 2022.

[58] R. Keeling, R. Chhatwal, N. Huber-Fliflet, J. Zhang, F. Wei, H. Zhao, Y. Shi, and H. Qin. Empirical comparisons of cnn with other learning algorithms for text classification in legal document review. In *Proceedings - 2019 IEEE International Conference on Big Data, Big Data 2019*, pages 2038–2042, 2019.

[59] Kenniscentrum vooor beleid en regelgeving, Ministerie van Justitie en Veiligheid. Draaiboek voor de regelgeving, 7 2024. Nr. 20a (Consultatie van openbare lichamen Bonaire, Sint Eustatius en Saba).

[60] A. S. Khabibullina, S. B. Seleckaya, and A. N. Shpagonov. The problems of robotization of legal profession. *Rev Genero Direito*, 8(6):397–405, 2019.

[61] B. A. Kitchenham and S. Charters. Guidelines for performing Systematic Literature Reviews in Software Engineering. *ResearchGate*, 2(3), January 2007.

[62] Ron Kohavi and Foster Provost. Glossary of terms. *Machine Learning*, 2:271–274, 01 1998.

[63] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[64] C. Kronblad, J. E. Pregmark, and R. Berggren. Difficulties to digitalize: ambidexterity challenges in law firms. *Journal of Service Theory and Practice*, 33:217–236, 2023.

[65] S. Kuleshov, A. Zaytseva, and K. Nenausnikov. Legal tech: Documents' validation method based on the associative-ontological approach. In *International Conference on Speech and Computer*, pages 244–254. Springer, 2020.

[66] Z Lan. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[67] M. Lauritsen. Technology report: Building legal practice systems with today's commercial authoring tools. *Artificial Intelligence and Law*, 1:87–102, 1992.

[68] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[69] R. E. Leenes. Burden of proof in dialogue games and dutch civil procedure. In *Proceedings of the International Conference on Artificial Intelligence and Law*, pages 109–118, 2001.

[70] P. Leith. The application of ai to law. *AI & Society*, 2:31–46, 1988.

[71] E. Leitner, G. Rehm, and J. Moreno-Schneider. Fine-grained named entity recognition in legal documents. In *International Conference on Semantic Systems*, pages 272–287. Springer, 2019.

[72] Yinhan Liu. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[73] M. Livson, S. Eshtokin, V. Vasyukov, E. Yudina, A. Baybarin, and S. Pivneva. Impact of digitalization on legal regulation: formation of new legal practices. *Journal of Law and Sustainable Development*, 9(2):e0749–e0749, 2021.

[74] Batta Mahesh. Machine learning algorithms-a review. *International Journal of Science and Research (IJSR)*, 9(1):381–386, 2020.

[75] K. Mania. Legal technology: Assessment of the legal tech industry's potential. *Journal of the Knowledge Economy*, 14:595–619, 2023.

[76] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Cross-entropy loss functions: Theoretical analysis and applications. In *International conference on Machine learning*, pages 23803–23828. PMLR, 2023.

[77] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*, 2019.

[78] Mihai Masala, Radu Cristian Alexandru Iacob, Ana Sabina Uban, Marina Cidota, Horia Velicu, Traian Rebedea, and Marius Popescu. jurbert: A romanian bert model for legal judgement prediction. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 86–94, 2021.

[79] John O McGinnis and Russell G Pearce. The great disruption: How machine intelligence will transform the role of lawyers in the delivery of legal services. *Actual Probs. Econ. & L.*, page 1230, 2019.

[80] T. McKeown, J. Mustafina, R. Magizov, and C. Gataullina. Ai in law practices. In *Proceedings - International Conference on Developments in eSystems Engineering, DeSE*, volume 2020-Decem, pages 27–32, 2020.

[81] S. McLachlan, E. Kyrimi, K. Dube, N. Fenton, and L. C. Webley. Lawmaps: enabling legal ai development through visualisation of the implicit structure of legislation and lawyerly process. *Artificial Intelligence and Law*, 31:169–194, 2023.

[82] O. Metsker, E. Trofimov, and G. Kopanitsa. Application of machine learning metrics for dynamic e-justice processes. In *Conference of Open Innovation Association, FRUCT*, volume 2021-Janua, 2021.

[83] O. Metsker, E. Trofimov, S. Sikorsky, and S. Kovalchuk. Text and data mining techniques in judgment open data analysis for administrative practice control. In *Communications in Computer and Information Science*, volume 947, pages 169–180, 2019.

[84] Dan Milmo. ChatGPT reaches 100 million users two months after launch. *the Guardian*, February 2023.

[85] Ministerie van Veiligheid en Justitie. Auteurswet, October 2022. Auterswet Artikel 11. BWBR0001886.

[86] O. I. Miroshnichenko and D. S. Proscurina. The role of artificial intelligence in professional legal sphere: Development tool or existential threat? In *Lecture Notes in Networks and Systems*, volume 198, pages 928–937, 2021.

[87] J. R. Mok, W. Y. Mok, and R. V. Mok. Sentence classification for contract law cases: A natural language processing approach. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021*, pages 260–261, 2021.

[88] P. Mongeon and A. Paul-Hus. The journal coverage of web of science and scopus: a comparative analysis. *Scientometrics*, 106(1):213–228, October 2015.

[89] J. Moreno-Schneider, G. Rehm, E. Montiel-Ponsoda, V. Rodríguez-Doncel, A. Revenko, S. Karampatakis, M. Khvalchik, C. Sageder, J. Gracia, and F. Maganza. Orchestrating nlp services for the legal domain. In *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*, pages 2332–2340, 2020.

[90] A. Munthuli, V. Socatiyanurak, S. Sangchocanonta, L. Kovudhikulrungsri, N. Saksakulkunakorn, P. Chairuangsri, and C. Tantibundhit. Transformers for multi-intent classification and slot filling of supreme court decisions related to sexual violence law. *IEEE Access*, 11:76448–76467, 2023.

[91] A. Nazarenko and A. Wyner. Legal nlp introduction. *TAL Traitement Automatique des Langues*, 58:7–19, 2017.

[92] J. Niman and S. William. Leveraging technology to become a better lawyer. In *ACM International Conference Proceeding Series*, 2013.

[93] E. Nissan. Digital technologies and artificial intelligence's present and foreseeable impact on lawyering, judging, policing and law enforcement. *AI and Society*, 32:441–464, 2017.

[94] Publications Office of the EU. EuroVoc - EU Vocabularies - Publications Office of the EU, November 2024. [Online; accessed 26. Nov. 2024].

[95] Y. Ogurlu, R. V. Shagieva, and A. S. Bersanov. A debate on artificial intelligence in area of law and the legal professions. In *Lecture Notes in Networks and Systems*, volume 372, pages 365–374, 2022.

[96] Pariwat Ongsulee. Artificial intelligence, machine learning and deep learning. In *2017 15th International Conference on ICT and Knowledge Engineering (ICT&KE)*, pages 1–6, 11 2017.

[97] T. Orosz, R. Vági, G. M. Csányi, D. Nagy, I. Üveges, J. P. Vadász, and A. Megyeri. Evaluating human versus machine learning performance in a legaltech problem. *Applied Sciences (Switzerland)*, 12, 2022.

[98] A. Oskamp and M. Lauritsen. Ai in law practice? so far, not much. *Artificial Intelligence and Law*, 10:227–236, 2002.

[99] A. Oskamp, M. W. Tragter, and A. R. Lodder. Mutual benefits for ai & law and knowledge management. In *Proceedings of the 7th international conference on Artificial intelligence and law*, pages 126–127. Association for Computing Machinery, 6 1999.

[100] Oxford English Dictionary. artificial intelligence, 12 2008.

[101] R. Pande and S. Alam. Predicting the outcome of judicial cases using semantic analysis. In *2020 IEEE Symposium Series on Computational Intelligence*, pages 1757–1761, 2020.

[102] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.

[103] F. Di Porto. Algorithmic disclosure rules. *Artificial Intelligence and Law*, 31:13–51, 2023.

[104] Y.-Y. Rhim and K. Park. The applicability of artificial intelligence in international law. *Journal of East Asia and International Law*, 12:7–30, 2019.

[105] L. Robaldo, S. Batsakis, R. Calegari, F. Calimeri, M. Fujita, G. Governatori, M. C. Morelli, F. Pacenza, G. Pisano, K. Satoh, I. Tachmazidis, and J. Zangari. Compliance checking on first-order knowledge with conflicting and compensatory norms: a comparison among currently available technologies. *Artificial Intelligence and Law*, 2023.

[106] L. Roberge, S. Long, P. Hassett, and D. Burnham. Technology and the changing practice of law: An entrée to previously inaccessible information via trac. *Artificial Intelligence and Law*, 10:261–282, 2002.

[107] Esther Salmerón-Manzano. Legaltech and lawtech: Global perspectives, challenges, and opportunities. *Laws*, 10(2):24, 2021.

[108] F. M. De Sanctis. Artificial intelligence and innovation in brazilian justice. *International Annals of Criminology*, 59:1–10, 2021.

[109] M. Scherer. Artificial intelligence and legal decision-making: The wide open? *Journal of international arbitration*, 36(5), 2019.

[110] A. Shelar and M. Moharir. A comparative study to determine a suitable legal knowledge representation format. In *3rd International Conference on Electrical, Electronics, Communication, Computer Technologies and Optimization Techniques, ICEECCOT 2018*, pages 514–519, 2018.

[111] R. Sil, A. Roy, B. Bhushan, and A. K. Mazumdar. Artificial intelligence and machine learning based legal application: The state-of-the-art and future research trends. In *Proceedings - 2019 International Conference on Computing, Communication, and Intelligent Systems, ICCCIS 2019*, volume 2019-Janua, pages 57–62, 2019.

[112] H. Silva, N. António, and F. Bacao. A rapid semi-automated literature review on legal precedents retrieval. In *EPIA Conference on Artificial Intelligence*, volume 13566 LNAI, pages 53–65, 2022.

[113] N. Sivaranjani, J. Jayabharathy, and M. Safa. A broad view of automation in legal prediction technology. In *Proceedings of the 3rd International Conference on Electronics and Communication and Aerospace Technology, ICECA 2019*, pages 180–185, 2019.

[114] J. Soukupová. Ai-based legal technology: a critical assessment of the current use of artificial intelligence in legal practice. *Masaryk University Journal of Law and Technology*, 15:279–300, 2021.

[115] C. Sun, Y. Zhang, X. Liu, and F. Wu. Legal intelligence: Algorithmic, data, and social challenges. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2464–2467, 2020.

[116] H. Surden. Values embedded in legal artificial intelligence. *IEEE Technology and Society Magazine*, 41:66–74, 2022.

[117] M. Thornton. Legal professionalism in a context of uberisation. *International Journal of the Legal Profession*, 28:243–263, 2021.

[118] G. Vaciago. Opportunities and challenges in the legal tech services in the italian and european framework. *Frontiers in Artificial Intelligence and Applications*, 317:280–289, 2019.

[119] H. Vardhan, N. Surana, and B. K. Tripathy. Named-entity recognition for legal documents. In *Advances in Intelligent Systems and Computing*, volume 1141, pages 469–479, 2021.

[120] D. Vianna and E. Silva De Moura. Organizing portuguese legal documents through topic discovery. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3388–3392, 2022.

[121] M. A. Vladimirovich and E. K. Sergeevich. Alternative dispute resolution in digital government. *Revista Brasileira de Alternative Dispute Resolution*, 4:119–146, 2022.

[122] B. Waltl, G. Bonczek, E. Scepankova, J. Landthaler, and F. Matthes. Predicting the outcome of appeal decisions in germany's tax law. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10429 LNCS, pages 89–99, 2017.

[123] Z. Wang, B. Wang, X. Duan, D. Wu, S. Wang, G. Hu, and T. Liu. Iflylegal: A chinese legal system for consultation, law searching, and document analysis. In *2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, Proceedings of System Demonstrations*, pages 97–102, 2019.

[124] Roel J Wieringa. *Design science methodology for information systems and software engineering*. Springer, 2014.

[125] Wolters Kluwer NV. Algemene Leveringsvoorwaarden Legal Intelligence 2022, Article 29.1, 2022.

[126] N. Xu, K.-J. Wang, and C.-Y. Lin. Technology acceptance model for lawyer robots with ai: A quantitative survey. *International Journal of Social Robotics*, 14:1043–1055, 2022.

[127] E. Yang, D. D. Lewis, and O. Frieder. On minimizing cost in legal document review workflows. In *DocEng 2021 - Proceedings of the 2021 ACM Symposium on Document Engineering*, 2021.

[128] Z. Zhang, J. Nandhakumar, J. T. Hummel, and L. Waardenburg. Addressing the key challenges of developing machine learning ai systems for knowledge-intensive work. *MIS Quarterly Executive*, 19:221–238, 2020.

[129] H. Zhao, S. Ye, and J. Yang. An empirical study on transfer learning for privilege review. In *2021 IEEE International Conference on Big Data*, pages 2729–2733, 2021.

Appendix A

Appendices

A.1 Background

A.1.1 Retrieved literature in SLR

Library	Total	English	Article or conference
Scopus	275	251	179
Web of Science	86	72	67
ScienceDirect*	9	9	8
ACM	6	6	6
IEEE Xplore	3	3	3
JSTOR	1	1	1
LegalIntelligence	0	0	0
Total	380	342	264

Table A.1: Retrieved Literature per Database

A.1.2 Legaltech

Legal technology describes all applications of innovative technologies in the legal world. These technologies traditionally carry out tasks by lawyers and other legal professionals[107]. The goal of legaltech is to use technical solutions to improve the provision of legal services. These solutions aim to enhance efficiency, productivity, cost-effectiveness, and overall client outcomes [107].

The use of technology in the legal sector is not new. Historically, legal services were delivered very differently. Before the start of the computer era, lawyers were dependent on their ability to acquire and consolidate knowledge [4]. This changed in 1976, when Westlaw and Lexis introduced a computer-assisted tool to perform legal research, marking a significant shift in the use of legaltech. Lawyers could search through legal materials using a terminal and, later, a personal computer. Keyword searches enabled lawyers to access relevant information in extensive libraries rapidly [4].

Another significant shift within the legal profession was made when many materials were digitalized. As society continued its digital transformation, legal businesses also needed to evolve. Numerous law firms have made substantial investments in legaltech solutions to address increasing digitalisation in the legal field. Digitalization had a transformative impact on the organization of legal work [19]. More work is done digitally to get or maintain a competitive advantage towards other law firms [17].

In recent years, concerns about potential disruption in the legal services market, mainly due to the rise of artificial intelligence, have been raised [79]. In 2014, experts predicted that artificial intelligence would eventually surpass many lawyers in providing legal services, with lawyers remaining essential only in highly specialized areas of law, court appearances, or service contexts where human relationships are central to quality [79]. However, these claims, sometimes considered sensationalist, are substantially inaccurate. Recent research in 2022 by [18] suggests that, particularly among large firms, there is little intention to use new technologies to disrupt legal practice. Instead, firms are increasing their investments in new technologies to support and enhance their services rather than disrupting them.

A.1.3 Machine Learning

Supervised learning Supervised learning is an algorithm that trains the model based on a labelled dataset. The input is associated with the corresponding output or target. Experts manually label a subset of the learning data, and the algorithm tries to predict the most probable output for the unlabelled data. Supervised learning is often used for classification and regression.

Unsupervised learning Unsupervised learning is an algorithm that trains a model based on an unlabelled dataset. The model must discover patterns and structures in the data without explicit guidance, making it often less effective than supervised learning. However, it is frequently used for clustering data and association.

Reinforced learning Reinforced learning is an algorithm involving an agent that learns to make decisions by interacting with an environment. The model learns when users are interacting with the model. Due to this interaction, agents receive feedback through rewards or penalties. Reinforcement learning is often used for game-playing AI and robotics. Examples of reinforcement learning algorithms include AlphaZero and Q-learning.

Deep learning Deep learning is a subbranch of machine learning. A deep learning algorithm involves training a neural network with multiple hidden layers [96]. A deep learning algorithm "learns" gradually to understand and make decisions independently. It aims to mimic the way a human brain processes. It learns from information by using interconnected layers of nodes, called neurons or units, to identify patterns and features within complex datasets [26]. Deep neural networks have been demonstrated to outperform other machine learning algorithms [63]. They are often applied to fields like computer vision, automatic speech recognition and natural language processing [96].

A.1.4 Papers on Legal Domain

Legal Domain	Papers
Not specified	[2], [8], [21], [22], [25], [28], [32], [34], [35], [44], [52], [55], [56], [58], [60], [69], [70], [71], [75], [86], [89], [91], [95], [98], [99], [105], [111], [113], [117], [118], [120], [123], [127], [129]
Lawyers & Law firms	[1], [3], [4], [5], [6], [7], [9], [11], [12], [15], [16], [29], [30], [31], [35], [40], [42], [45], [46], [48], [51], [54], [57], [64], [65], [73], [80], [81], [87], [93], [97], [101], [104], [106], [109], [110], [114], [115], [119], [122], [126], [128]
Justice & Court	[1], [10], [24], [29], [47], [50], [73], [87], [89], [93], [108], [112], [114], [115], [121]
Clients	[10], [14], [47], [67], [90], [115]
Government & Law makers	[1], [73], [82], [83], [103], [116]
Others	[93]

TABLE A.2: Legal Domains in Research Papers

A.1.5 Papers on AI Applications

AI Application	Papers
Litigation & Prediction analysis	[3], [4], [5], [6], [7], [10], [11], [12], [15], [16], [22], [34], [35], [42], [44], [46], [48], [54], [57], [73], [82], [86], [87], [90], [91], [93], [97], [101], [104], [109], [111], [113], [114], [115], [122]
eDiscovery	[4], [2], [5], [6], [10], [11], [14], [15], [16], [22], [24], [25], [28], [29], [30], [31], [35], [48], [52], [53], [54], [58], [60], [71], [83], [89], [91], [93], [95], [97], [111], [117], [119], [120], [123], [127], [128], [129]
Legal search	[1], [4], [5], [6], [7], [11], [12], [15], [16], [21], [24], [25], [28], [29], [30], [48], [51], [53], [60], [91], [95], [97], [98], [111], [114], [115], [119], [123], [128]
Document generation	[3], [4], [11], [12], [16], [24], [30], [45], [50], [52], [56], [60], [65], [67], [89], [91], [98], [108], [111], [113], [114], [115]
Contract analysis	[1], [5], [6], [7], [11], [12], [15], [24], [30], [45], [48], [54], [60], [86], [95]
Due diligence reviews	[5], [6], [7], [11], [15], [24], [40], [113], [114]
Dispute resolution	[5], [6], [34], [90], [97], [121]
Not specified	[32], [55], [70], [103], [116]
Others	[6], [7], [8], [10], [12], [42], [54], [69], [80], [82], [83], [87], [90], [92], [93], [104], [110], [126]

TABLE A.3: AI Applications in Research Papers

A.1.6 Papers on ML technologies

Supervised algorithm	Descriptive papers	Non-Descriptive papers
Support Vector Machines (SVM)	[32], [44], [113]	[2], [22], [28], [53], [58], [83], [90], [97]
Logistic Regression (LR)	[32], [44], [113]	[28], [58], [83]
Decision Trees (DT)	[32], [44], [113]	[58], [83]
Random Forest (RF)	[32], [44], [113]	[28], [58], [83]
K-Nearest Neighbors (KNN)	[32], [44],	[28], [83]
Naive Bayes (NB)	[32], [44], [113]	[14], [16], [28], [83], [122]
Gradient Boosting (GB)	[44]	[83]
Conditional Random Fields (CRF)		[71], [89]
General	[5], [6], [35], [54]	[52], [55]
Other	[32], [44]	

TABLE A.4: Supervised Machine Learning Algorithms present in Research Papers

Unsupervised algorithm	Descriptive papers	Non-Descriptive papers
k-means	[32]	
Latent Dirichlet Allocation (LDA)	[44], [112]	[65], [101]
Latent Semantic Indexing (LSI)		[65]
Correlated Topic Model (CTM)		[120]
Global Vectors for Word Representation (GloVe)	[44]	
Other		[32]

TABLE A.5: Unsupervised Machine Learning Algorithms present in Research Papers

Deep Learning algorithm	Descriptive papers	Non-Descriptive papers
Artificial Neural Network (ANN)	[44]	[42]
Convolutional Neural Network (CNN)	[44], [112]	[58], [101], [119]
Recurrent Neural Network (RNN)	[44]	[57], [101], [123]
Capsule Networks (CapsNet)		[101]
Gated Recurrent Unit (GRU)	[44]	[57]
Multilayer Perceptron (MLP)	[44]	[119]
Long Short-Term Memory (LSTM)	[44], [112]	[57]
Bidirectional Long Short-Term Memory (BiLSTM)	[44]	[71], [89], [123]
Markov Logic Network (MLN)	[44]	
General	[48], [104], [113]	[40]

TABLE A.6: Deep Learning Algorithms present in Research Papers

A.1.7 Papers on NLP technologies

Technique	Descriptive papers	Non-Descriptive papers
BERT	[44], [112]	[2], [21], [22], [31], [57], [90], [120], [123], [129]
GPT		[22]
Transformer encoder		[57]
YARGY		[56]

TABLE A.7: Families of Large Language Models identified in Literature

Technique	Descriptive papers	Non-Descriptive papers
BERT	[44], [112]	[2], [21], [22], [31], [57], [90], [120], [123], [129]
AraLegal-BERT		[2]
AraBERTv2-Large		[2]
ARBERT		[2]
BART		[22]
BERTopic		[120]
BERTimbau		[120]
BERTikal		[120]
BigBird		[22]
BiMPM		[123]
CamemBERT	[44]	
CaseLaw-BERT		[22]
DeBERTa		[22]
DistilBERT		[57]
FlauBERT	[44]	
ftBERT	[44]	
LegalBERT	[44]	[21], [22]
LegalBERT-SMALL		[21]
Longformer		[22]
mBERT		[2]
MiniLM		[31]
ooBERT	[44]	
RoBERT	[44]	
RoBERTa		[22], [31]
XLM-RoBERTa		[90]
Sentence-BERT		[57]

TABLE A.8: Types of BERT identified in Literature

A.2 Interview Guideline

A.2.1 Structure

Topic The use of artificial intelligence within legal technology

Purpose

1. To gain insight into work processes (what do people do in a day?) and therein potential AI applications.
2. Gain insight into which repetitive tasks are performed and which of these require labour power

A.2.2 Introduction

Introducing

- Thanks for the time made available
- Introduction
- Explaining why this interview is being conducted (Purpose of study)
- The place of this interview in the study

Practical information

- Voluntary, anonymous and non-binding
- Asking if interview can be recorded

A.2.3 Questions

- Daily Work and Potential AI Applications:
    1. Can you tell us about your daily work and the nature of the tasks you perform?
    2. Do you see any areas within your work where AI could be valuable? If so, which specific tasks or processes?
- Repetitive Work Processes and Labour-intensive Tasks:
    1. Are there any work processes that you experience as repetitive or involve a lot of labour hours?
    2. Could you give specific examples of such tasks or processes?
- Bottlenecks and Inefficiencies in Work Area:
    1. Do you have any general insights about possible bottlenecks or inefficiencies within your work area?
    2. Are there any specific challenges that you think AI solutions could help with?
- Document Use and Automation Capabilities:

1. What types of documents do you use in your work and to what extent are they proprietary or publicly accessible?
2. Do you see opportunities for automation or use of AI in managing, analysing, or processing these documents?

- Personal Vision on AI within [Company]:
  1. What is your personal vision for the use of AI within our organisation?
  2. Do you see any specific areas where AI can contribute to our growth or efficiency?
- Personal experience with AI tools
  1. Have you personally used AI tools before, like ChatGPT or DeepL?

A.2.4 Closing

- Thanks participants for participation
- Stop recording

A.3 Interview results

All four respondents work as fee-earners in a law firm. Their daily tasks include advising clients on the law, drawing up contracts, documents, and notary acts. For example, Lawyer 3 advises and informs companies on employment law related to court filings, the Dutch Employee Insurance Administration Agency (UWV), or reorganizations. Lawyer 2 often drafts and translates documents into English or German.

Their daily work involves repetitive and labor-intensive work. According to Lawyer 1, most drafted documents or contracts are about 10% to 15% unique, with the rest being similar to other documents. Lawyer 3 supports this claim, adding that large parts of legal documents are the same, with subtle differences based on the document's purpose and which party has drafted it. Lawyer 3 claims that for every use case, there is a model or an example document to use. Lawyer 4 believes that work is not repetitive, but acknowledges that drafted documents have much overlap, and his department heavily relies on automated models, while the other lawyers' departments do not. The other lawyers all mentioned that their departments do not have automated models in the system. Lawyer 1 mentioned that templates and automated models are not developed for his department as this development takes time, and people do not invest this time to create proper models. Lawyer 3 said that everyone uses his own created templates and guidelines, as everyone has his own style. Lawyer 2 feels that colleagues are hesitant to share their templates from a competitive point of view.

Much value is connected to blogs and published documents of experienced colleagues, according to Lawyer 1. The firm has licenses for these articles and can access them through LegalIntelligence. According to Lawyer 2, LegalIntelligence is used for information, but its searching algorithms and interface only sometimes give precisely the results you want. Lawyer 3 mentioned that he is looking for court rulings on Google and then searches them in LegalIntelligence. This lawyer participated in a test with an AI-powered search for a publisher. He mentioned that this software provided a decision tree to see how the algorithm came to its answer. You usually have many results but cannot find what you are looking for. According to Lawyer 2, searching the firm's legal files is even harder since the IT solution does not allow a thorough search through documents, only the document filename and subject title of e-mails. Lawyer 4 supports this claim, calling the search method inefficient and time-consuming.

The firm also often works on translating legal documents for clients. Often, more and more clients are international, and German customers are not uncommon. Therefore, translating is repetitive, and the interviewed lawyers think AI tools can support this. Lawyer 2 mentioned that translating documents is often done without premium translating tools since the firm needs licenses for these tools. DeepL is a tool often used by many employees, according to Lawyer 2. Lawyer 4 mentioned that he uses ChatGPT to translate large documents. He finds the benefit of this system is that he can ask to use terms consistently. DeepL is also tried but uses many synonyms, which he finds not preferable for legal documents. Lawyer 2 mentioned that it is hard to charge for translations due to the upcoming translation tools. How do you charge for things? Do you charge per translation or hour?

When looking at the lawyers' experience and personal vision of AI, it can be noted that only some have the same vision of AI. Lawyer 1 mentioned that he did not try AI himself but does not understand why everyone is so hyped about the evolution of AI. Lawyer 3 thinks the innovation is not going fast enough and believes the firm should take action immediately. Lawyer 2 mentioned that the firm is reserved to use new IT. Previous

IT project implementations did not always go well, and there are considerable privacy concerns regarding the use of AI tools. Lawyer 1 confirms the privacy statement of Lawyer 2. Lawyer 3 mentions that when AI is implemented, it should start small with a few lawyers since he believes that it first needs to be proved on a small scale. It will only work when other employees are convinced of its needs and values, which will take time.

A.4 Loss Results training models

A.4.1 BERTje

Table A.9: Cross-Entropy Loss BERTje

STEP	MLM	SOP	TOTAL
0	11,1670	0,7027	11,8697
31250	0,7043	0,6934	1,3977
62500	0,9855	0,6932	1,6787
93750	1,2833	0,6932	1,9765
125000	1,3610	0,6932	2,0542
156250	1,5306	0,6931	2,2237
187500	1,3801	0,6931	2,0732
218750	1,2326	0,6931	1,9257
250000	1,2021	0,6931	1,8952
281250	1,1996	0,6931	1,8927
312500	1,1881	0,6931	1,8812
343750	1,1874	0,6931	1,8805
375000	1,1799	0,6931	1,8730
406250	1,1731	0,6931	1,8662
437500	1,1704	0,6931	1,8635
468750	1,1729	0,6931	1,8660
500000	1,1682	0,6931	1,8613
531250	1,1540	0,6931	1,8471
562500	1,1518	0,6931	1,8449
593750	1,1551	0,6931	1,8482
625000	1,1519	0,6931	1,8450
656250	1,1549	0,6931	1,8480
687500	1,1505	0,6931	1,8436
718750	1,1482	0,6931	1,8413
750000	1,1445	0,6931	1,8376
781250	1,1278	0,6931	1,8209
812500	1,1251	0,6931	1,8182
843750	1,1365	0,6931	1,8296
875000	1,1318	0,6931	1,8249
906250	1,1198	0,6931	1,8129
937500	1,1181	0,6931	1,8112
968750	1,1157	0,6931	1,8088
1000000	1,1132	0,6931	1,8063
1031250	1,1255	0,6931	1,8186
1062500	1,1206	0,6931	1,8137
1093750	1,1190	0,6931	1,8121
1125000	1,1157	0,6931	1,8088
1156250	1,1025	0,6931	1,7956
1187500	1,1010	0,6931	1,7941
1218750	1,1003	0,6931	1,7934
1250000	1,0974	0,6931	1,7905

Continued on next page

TABLE A.9: Cross-Entropy Loss BERTje (continued)

STEP	MLM	SOP	TOTAL
1281250	1,0936	0,6931	1,7867
1312500	1,0912	0,6931	1,7843
1343750	1,1255	0,6931	1,8186
1375000	1,1186	0,6931	1,8117
1406250	1,0798	0,6931	1,7729
1412691	1,0767	0,6931	1,7698
1443941	1,1626	0,6931	1,8557
1475191	1,0884	0,6931	1,7815
1506441	1,1832	0,6931	1,8763
1537691	1,1375	0,6931	1,8306
1568941	1,2738	0,6931	1,9669
1600191	1,2077	0,6931	1,9008
1631441	1,1539	0,6931	1,8470
1662691	1,1524	0,6931	1,8455
1693941	1,1611	0,6931	1,8542
1725191	1,1568	0,6931	1,8499
1756441	1,1587	0,6931	1,8518
1787691	1,1553	0,6931	1,8484
1818941	1,1503	0,6931	1,8434
1850191	1,1489	0,6931	1,8420
1881441	1,1517	0,6931	1,8448
1912691	1,1486	0,6931	1,8417
1943941	1,1355	0,6931	1,8286
1975191	1,1343	0,6931	1,8274
2006441	1,1374	0,6931	1,8305
2037691	1,1348	0,6931	1,8279
2068941	1,1389	0,6931	1,8320
2100191	1,1354	0,6931	1,8285
2131441	1,1339	0,6931	1,8270
2162691	1,1309	0,6931	1,8240
2193941	1,1153	0,6931	1,8084
2225191	1,1131	0,6931	1,8062
2256441	1,1138	0,6931	1,8069
2287691	1,1115	0,6931	1,8046
2318941	1,1247	0,6931	1,8178
2350191	1,1205	0,6931	1,8136
2381441	1,1090	0,6931	1,8021
2412691	1,1075	0,6931	1,8006
2443941	1,1055	0,6931	1,7986
2475191	1,1034	0,6931	1,7965
2506441	1,1160	0,6931	1,8091
2537691	1,1113	0,6931	1,8044
2568941	1,1099	0,6931	1,8030
2600191	1,1068	0,6931	1,7999
2631441	1,0939	0,6931	1,7870

Continued on next page

TABLE A.9: Cross-Entropy Loss BERTje (continued)

STEP	MLM	SOP	TOTAL
2662691	1,0921	0,6931	1,7852
2693941	1,0879	0,6931	1,7810

A.4.2 mBERT

TABLE A.10: Cross-Entropy Loss mBERT

STEP	MLM	SOP	TOTAL
0	10,595	0,7027	11,2977
31250	0,7941	0,6934	1,4875
62500	0,7889	0,6932	1,4821
93750	0,9122	0,6932	1,6054
125000	0,9142	0,6932	1,6074
156250	1,0045	0,6931	1,6976
187500	0,9630	0,6931	1,6561
218750	0,9178	0,6931	1,6109
250000	0,9101	0,6931	1,6032
281250	0,9215	0,6931	1,6146
312500	0,9148	0,6931	1,6079
343750	0,9142	0,6931	1,6073
375000	0,9093	0,6931	1,6024
406250	0,9090	0,6931	1,6021
437500	0,9048	0,6931	1,5979
468750	0,9125	0,6931	1,6056
500000	0,9075	0,6931	1,6006
531250	0,8959	0,6931	1,5890
562500	0,8928	0,6931	1,5859
593750	0,8948	0,6931	1,5879
625000	0,8912	0,6931	1,5843
656250	0,8983	0,6931	1,5914
687500	0,8935	0,6931	1,5866
718750	0,8926	0,6931	1,5857
750000	0,8889	0,6931	1,5820
781250	0,8758	0,6931	1,5689
812500	0,8727	0,6931	1,5658
843750	0,8735	0,6931	1,5666
875000	0,8695	0,6931	1,5626
906250	0,8828	0,6931	1,5759
937500	0,8782	0,6931	1,5713
968750	0,8701	0,6931	1,5632
1000000	0,8679	0,6931	1,5610
1031250	0,8659	0,6931	1,5590
1062500	0,8628	0,6931	1,5559
1093750	0,8757	0,6931	1,5688
1125000	0,8702	0,6931	1,5633

Continued on next page

TABLE A.10: Cross-Entropy Loss mBERT (continued)

STEP	MLM	SOP	TOTAL
1156250	0,8815	0,6931	1,5746
1187500	0,8769	0,6931	1,5700
1218750	0,8535	0,6931	1,5466
1250000	0,8504	0,6931	1,5435
1281250	0,8537	0,6931	1,5468
1312500	0,8509	0,6931	1,5440
1343750	0,8492	0,6931	1,5423
1350191	0,8463	0,6931	1,5394
1381441	0,7941	0,6931	1,4872
1412691	0,7890	0,6931	1,4821
1443941	0,9125	0,6931	1,6056
1475191	0,9144	0,6931	1,6075
1506441	1,0043	0,6931	1,6974
1537691	0,9629	0,6931	1,6560
1568941	0,9177	0,6931	1,6108
1600191	0,9101	0,6931	1,6032
1631441	0,9211	0,6931	1,6142
1662691	0,9140	0,6931	1,6071
1693941	0,9219	0,6931	1,6150
1725191	0,9083	0,6931	1,6014
1756441	0,9088	0,6931	1,6019
1787691	0,9049	0,6931	1,5980
1818941	0,9123	0,6931	1,6054
1850191	0,9078	0,6931	1,6009
1881441	0,8960	0,6931	1,5891
1912691	0,8930	0,6931	1,5861
1943941	0,8948	0,6931	1,5879
1975191	0,8912	0,6931	1,5843
2006441	0,8983	0,6931	1,5914
2037691	0,8933	0,6931	1,5864
2068941	0,8925	0,6931	1,5856
2100191	0,8889	0,6931	1,5820
2131441	0,8750	0,6931	1,5681
2162691	0,8719	0,6931	1,5650
2193941	0,8728	0,6931	1,5659
2225191	0,8684	0,6931	1,5615
2256441	0,8820	0,6931	1,5751
2287691	0,8744	0,6931	1,5675
2318941	0,8696	0,6931	1,5627
2350191	0,8672	0,6931	1,5603
2381441	0,8657	0,6931	1,5588
2412691	0,8624	0,6931	1,5555
2443941	0,8753	0,6931	1,5684
2475191	0,8699	0,6931	1,5630
2506441	0,8813	0,6931	1,5744

Continued on next page

TABLE A.10: Cross-Entropy Loss mBERT (continued)

STEP	MLM	SOP	TOTAL
2537691	0,8765	0,6931	1,5696
2568941	0,8531	0,6931	1,5462
2600191	0,8502	0,6931	1,5433
2631441	0,8535	0,6931	1,5466
2662691	0,8507	0,6931	1,5438
2693941	0,8488	0,6931	1,5419
2725191	0,8461	0,6931	1,5392
2756441	0,8401	0,6931	1,5332
2762882	0,8344	0,6931	1,5275
2794132	0,8303	0,6931	1,5234
2825382	0,8326	0,6931	1,5257

A.4.3 RobBERT

TABLE A.11: Cross-Entropy Loss RobBERT

STEP	MLM
0	12,6422
31250	1,43
62500	2,2659
93750	2,1911
125000	1,8776
156250	1,851
187500	1,6049
218750	1,4242
250000	1,3854
281250	1,3902
312500	1,3827
343750	1,3813
375000	1,3752
406250	1,3678
437500	1,3645
468750	1,3702
500000	1,3667
531250	1,3488
562500	1,3477
593750	1,3486
625000	1,3457
656250	1,3522
687500	1,3488
718750	1,3467
750000	1,3441
781250	1,3217
812500	1,3195

Continued on next page

TABLE A.11: Cross-Entropy Loss RobBERT (continued)

STEP	MLM
843750	1,3198
875000	1,3171
906250	1,3359
937500	1,3322
968750	1,3159
1000000	1,3147
1031250	1,3117
1062500	1,3097
1093750	1,3266
1125000	1,3213
1156250	1,3249
1187500	1,3222
1218750	1,2962
1250000	1,295
1281250	1,2948
1312500	1,2927
1343750	1,2891
1375000	1,2869
1406250	1,2715
1412691	1,2685
1443941	1,3937
1475191	1,3353
1506441	1,4628
1537691	1,4078
1568941	1,5321
1600191	1,444
1631441	1,3657
1662691	1,3616
1693941	1,3728
1725191	1,3677
1756441	1,3686
1787691	1,3638
1818941	1,3572
1850191	1,355
1881441	1,3603
1912691	1,3572
1943941	1,3397
1975191	1,3389
2006441	1,3403
2037691	1,3374
2068941	1,3436
2100191	1,34
2131441	1,3382
2162691	1,3354

Continued on next page

Table A.11: Cross-Entropy Loss RobBERT (continued)

STEP	**MLM**
2193941	1,3135
2225191	1,3111
2256441	1,3114
2287691	1,3084
2318941	1,3275
2350191	1,3233
2381441	1,3075
2412691	1,3061
2443941	1,3035
2475191	1,3016
2506441	1,3182
2537691	1,3129
2568941	1,3151
2600191	1,312
2631441	1,288
2662691	1,2869
2693941	1,2867
2725191	1,2846
2756441	1,2813
2787691	1,2794
2818941	1,2638
2825382	1,2608

A.5 Validation dataset

Eurovoc ID	# TRAIN	# TEST	# EVAL	Description (NL)
100142	2151	335	408	politiek
100143	8444	1086	944	internationale betrekkingen
100144	13519	1347	1499	EUROPESE UNIE
100145	2233	531	571	recht
100146	7065	444	478	economie
100147	30222	2769	2690	economie, verkeer en handelsverkeer
100148	8803	974	1211	financiën
100149	6056	774	778	sociale vraagstukken
100150	3640	513	469	opvoeding, onderwijs en communicatie
100151	877	57	65	wetenschappen
100152	5065	394	445	onderneming en concurrentie
100153	1926	164	203	werkgelegenheid en arbeid
100154	3317	600	512	transport
100155	3096	610	515	milieu
100156	19431	1960	2061	landbouw, bosbouw, en visserij
100157	26931	1641	2013	landbouwvoedingsmiddelen
100158	4175	622	578	productie, technologie en onderzoek
100159	1122	185	154	energie
100160	6150	653	568	industrie
100161	22975	2208	2082	geografie
100162	541	102	69	internationale organisaties

TABLE A.12: Number of EUROVOC IDs and Descriptions for NL monolingual part of MultiEURLEX dataset

ID	Label	Train Count	Validation Count	Test Count
100143	4	8444	944	1086
100144	6	13519	1499	1347
100146	9	7065	478	444
100147	3	30222	2690	2769
100148	2	8803	1211	974
100149	0	6056	778	774
100156	5	19431	2061	1960
100157	7	26931	2013	1641
100160	1	6150	568	653
100161	8	22975	2082	2208

TABLE A.13: Label counts in train, validation, and test datasets.