European Research Studies Journal
Volume XXVIII, Issue 2, 2025

pp. 1018-1048

Unified Text Segmentation (TSEG) Deep Learning Function for Legal Text Recovery: Enhancing Generative-NLP and AI-Driven Regulatory Operations in Industry 5.0

Submitted 23/04/25, 1st revision 05/05/25, 2nd revision 20/05/25, accepted 30/06/25

Pascal Muam Mah[1], Tomasz Pelech-Pilichowski[2]

Abstract:

Purpose: *The emergence of Industry 5.0 demands intelligent systems that not only automate tasks but also collaborate meaningfully with humans, especially in highly regulated fields like law and governance. This study addresses the challenges of fragmented, redacted, and corrupted legal documents, which hinder compliance, auditing, and interpretability..*
Design/Methodology/Approach: *This research introduces a Unified Tseg Deep Learning Function—a multi-layered framework combining legal text segmentation (Tseg), generative NLP techniques, and legal ontology alignment. The model incorporates transformer architectures, attention mechanisms, and recursive state modeling to enable dynamic content recovery, contextual classification, and visibility-based annotation based on age and jurisdiction.*
Findings: *The proposed system effectively recovers missing or corrupted legal content, categorizes segments by legal domain, and applies jurisdiction-specific and age-based visibility rules. It demonstrates strong potential in automating and enhancing interpretability within legal workflows and regulatory compliance systems.*
Practical Implications: *This framework can support digital governance platforms, legal information retrieval systems, and compliance monitoring tools by enabling more accurate and automated reconstruction and classification of legal documents, thus improving transparency and accountability in legal processes*
Originality/Value: *The study offers a novel integration of Tseg, generative NLP, and legal ontologies tailored for Industry 5.0. Its unique contribution lies in bridging technical deep learning methods with domain-specific legal compliance requirements, supporting the evolution of intelligent, human-centric regulatory technologies.*

Keywords: *Legal Text Recovery, Unified Tseg, natural language processing, Artificial intelligence, Deep Learning, Industry 5.0, Name Entity Recognition.*

JEL codes: *G34, G11, G14, L25, C33.*

Paper Type: *Research article.*

[1]*AGH University of Krakow, 30-059 Krakow, Poland, e-mail: mah@agh.edu.pl;*
[2]*AGH University of Krakow, 30-059 Krakow, Poland, e-mail: tomek@agh.edu.pl;*

1. Introduction

The legal sector is experiencing rapid digital transformation (Mania, 2022), propelled by the emergence of Industry 5.0, in which artificial intelligence enhances human reasoning to promote personalization, resilience, and sustainable development. Nonetheless, legal texts present distinct challenges owing to their intricate nature, specialized terminology, and the critical importance of context and interpretation. Additionally, legal documents frequently suffer from incompleteness as a result of redaction, privacy limitations, or data loss that occurs during the processes of transmission and digitization.

To tackle this challenge, we propose a Unified Tseg Deep Learning Function designed for the recovery of legal texts. Tseg stands for text segmentation $T_{\text{Recovered}}$, which we develop into a hierarchical and domain-aware methodology, allowing for the classification of legal content into thematic, juris- dictional, and demographic categories.

The proposed deep learning function combines transformers, sequence-to-sequence models for recovery, and rule-based post-processing that is guided by legal knowledge bases. Brooks *et al.* (2020) investigate the impact of artificial intelligence on the legal industry, underscoring significant pressures including organizational resistance, regulatory ambiguity, and ethical dilemmas. They stress the difficulties associated with incorporating AI into legal practice while upholding professional standards and preserving client trust.

The incorporation of Natural Language Processing (NLP) into the recovery of legal texts has fundamentally transformed the methodologies employed in the processing and analysis of legal documents. As Artificial Intelligence (AI) continues to advance and the transition towards Industry 5.0 unfolds—characterized by a focus on human-machine collaboration—NLP has emerged as a crucial technology that significantly improves the efficiency and precision of legal text recovery operations.

The research conducted by Shiel *et al.* (2009) examines the role of hyperspectral imaging in the recovery and segmentation of text from historical manuscripts. This study highlights the significance of digital codicology techniques in uncovering obscured writings, which enhances both paleographic analysis and the preservation of manuscripts through the implementation of advanced spectral imaging and computational methods

An analysis of the Recovery and Resilience Facility, characterizing it as a significant legal innovation within the context of European governance (Fabbrini, 2025). The research delves into its effects on integration, economic

revitalization, and the dynamics of institutions within the European Union. The investigation focuses on governance structures, legal frameworks, and the implications for policy.

An investigation in to the ”bail-in” provisions established by the European Union, evaluating their effects on the stability of banks and the security of depositors was conducted (Smits, 2014). The research delves into the associated legal frameworks, financial risks, and regulatory strategies to ascertain if European banks provide sufficient safeguards for customer deposits. Mah (2024a) examines national AI strategies, emphasizing data governance, privacy, accountability, and regulatory compliance.

The study highlights how legal frameworks shape AI deployment across public sectors, ensuring ethical use, transparency, and national sovereignty. It underscores the need for adaptive AI in all sectors to balance innovation with rights protection. Muller (1990) investigates the resurgence of Justinian’s ”Digest” during the Middle Ages, emphasizing its significant influence on the development of legal scholarship in that era. The research delves into the processes of manuscript transmission, the nuances of legal interpretation, and the revitalization of Roman law within the frameworks of medieval canon and civil law traditions.

The application of natural language processing in the recovery of legal texts in the age of artificial intelligence and Industry 5.0 seeks to promote collaboration, consolidate existing resources for legal text analytics, and establish an open-source platform. This initiative aims to empower both individuals and organizations in their research endeavors, thereby enhancing and refining contemporary tools to collectively address the challenges posed by missing or deleted legal texts.

This study enhances the development of smart, regulation-conscious systems aimed at document auditing, compliance verification, and legal drafting. We anticipate uses in digital courts, e-governance, international law, and automated regulatory reporting, establishing it as a pivotal aspect of the AI-led transformation in legal operations.

2. Literature Review

2.1 Legal Text Analysis in NLP

Recent developments in legal natural language processing (NLP) have concentrated on retrieving case law, extracting contract clauses, and answering legal questions. Research conducted by (Chalkidis *et al.,* 2019) presented the EURLEX dataset for the purpose of legal classification, whereas (Alberts *et al.,* 2020) introduced LegalBERT aimed at optimizing embeddings within legal

corpora. Nonetheless, there are limited models that tackle the issue of text recovery, particularly in instances where legal documents are either partially absent or have been redacted.

Text Segmentation (Tseg) Approaches: Text segmentation (Tseg) entails partitioning text into meaningful units for subsequent analysis. Historically, conventional models depended on lexical or statistical segmentation techniques; however, contemporary approaches employ BERT-based attention mechanisms for more adaptive segmentation. Hierarchical Tseg, aligned with specific task objectives, has demonstrated potential in areas such as summarization and argument mining, yet it is still not widely adopted in legal recovery frameworks.

Generative Models for Text Completion: Generative transformers, such as GPT and T5, have transformed the fields of text completion and inpainting. Their use in the legal domain is on the rise, particularly in the conditional generation of clauses and summaries. However, these mod- els frequently fall short in providing the precise legal control and age-specific filtering necessary for regulatory contexts.

AI in Industry 5.0 Legal Systems: Industry 5.0 highlights the importance of collaboration between humans and machines, personalization, and the ethical use of artificial intelligence. In the realm of legal technology, this manifests as AI that is aware of compliance, interpretable machine learning, and adaptable legal support. Our comprehensive framework is in harmony with this vision, as it retrieves lost content while preserving legal reasoning, jurisdictional limits, and appropriate visibility for different age groups.

An investigation into the application of Natural Language Processing (NLP) in the classification of legal documents was examined (Joseph and Vijayalakshmi, 2025). Their research delves into various NLP techniques, algorithms, and tools that facilitate the automation of legal document categorization, thereby improving communication and decision-making processes within legal frameworks through enhanced efficiency and accuracy in document management.

Siino *et al.* (2025) conduct a comprehensive review of the applications of large language models (LLMs) in the context of legal natural language processing (NLP). Their analysis encompasses existing methodologies, the challenges faced, and the advancements made in this field. The research delves into various aspects such as the processing of legal texts, the analysis of contracts, the retrieval of case law, and the ethical implications involved, while also emphasizing emerging trends, inherent limitations, and potential avenues for future research in the context of AI-driven legal automation.

The work of Onami *et al.* (2025) unveils ”LegalViz,” a system that converts legal language into diagrams to enhance understanding. This study focuses on the generation of diagrams from text, various visualization techniques, and their

implications for legal analysis, with the goal of improving accessibility, interpretability, and efficiency in the processing of legal texts. An investigation into a stylus tablet discovered at Vindolanda, which provides valuable insights into the institution of slavery along the Northern Frontier of the Roman Empire (Meyer *et al.,* 2025).

The authors meticulously analyze the tablet's inscription and its historical context docu- mentaries, thereby enhancing our understanding of Roman slavery practices and the socio-economic dynamics prevalent in that region during the specified era. Dong *et al.* (2025) present 'TermDiffuSum', a novel 'term-guided diffusion model' designed for the 'extractive summarization' of 'legal documents'. This research elucidates the model's methodology for pinpointing essential terms and enhancing 'summary accuracy' through the utilization of 'semantic term relationships', thereby advancing 'legal text comprehension' and 'information extraction' within the realm of NLP applications.

2.2 Legal Text Recovery

An introduction of 'MEL', a 'Legal Spanish Language Model' aimed at natural language processing applications in the realm of legal texts (Sánchez *et al.,* 2025). Their study outlines the 'training framework, linguistic features, and operational performance' in tasks including 'classification of legal documents, summarization, and named entity recognition', thereby facilitating improved AI-driven legal analysis in the Spanish context.

An analysis of 'ChatGPT's capabilities in legal classification' is presented, scrutinizing its 'accuracy, limitations, and relevance' in the context of legal text processing (Weichbroth, 2025). The research evaluates 'benchmark findings, error trends, and the interpretability of the model', drawing attention to the strengths and obstacles inherent in the application of AI for legal reasoning and automation.

Plonka *et al.* (2025) perform a comparative assessment of document splitting methodologies tailored for large language models within legal frameworks. This research investigates the efficacy of different splitting techniques in document processing, focusing on their influence on accuracy, efficiency, and comprehension in the analysis of legal texts and applications driven by artificial intelligence. An investigation into 'comparative law' is conducted through the lens of the metaphor "breaking the vessels," focusing on the ways in which different legal systems interpret and implement laws (Costantini, 2025).

This analysis addresses 'legal symbolism, cultural influences, and the fragmentation' of legal norms, underscoring their relevance to the evolution of international legal theory and its practical applications.

Recovering legal texts is vital for preserving the integrity, accessibility, and continuity of legal information, especially when faced with redaction, damage, or data loss. Recovering legal texts supports the process of digital archiving, ensures compliance with legal requirements, and enhances transparency in judicial and regulatory systems.

Through the application of AI and NLP, legal text recovery plays a crucial role in reconstructing lost content, safeguarding legal meaning, and facilitating fair access to justice. This is especially relevant in the areas of digital governance, automated compliance, and international legal operations in Industry 5.0 settings.

2.3 AI and Industry 5.0 in Legal Text Recovery

Industry 5.0, and the role of artificial intelligence extends past simple automation, highlighting the collaborative dynamics between human operators and machines. Legal professionals are turning to AI systems that utilize natural language processing to improve the efficiency of document analysis, derive valuable information, and ensure conformity with legal requirements, thus enhancing both the speed and accuracy of legal proceedings.

A research conducted by (Pusztahelyi and Stef˝an, 2024) addresses the prospects and challenges of legal informatics alongside the legal metrology framework within the framework of Industry 6.0. This study assesses the ramifications of emerging technologies on legal standards, data regulation, and measurement systems, highlighting the opportunities and challenges that must be navigated in the context of new industrial developments.

The value of ”AI and Industry 5.0 in Legal Text Recovery” is significant due to their combined ability to humanize and automate sophisticated legal processes. AI allows for the intelligent reconstruction of incomplete or damaged legal documents by leveraging natural language understanding and generation. In the context of Industry 5.0, which highlights the importance of human-AI collaboration, this results in improved accuracy, personalization, and regulatory compliance. Together, they create resilient, ethical, and context-aware legal recovery systems, empowering digital courts, legal experts, and citizens in an increasingly data-oriented legal framework.

2.4 Industry 5.0’s Human-Centric Approach Supports

Holzinger *et al.* (2024) examine the shift from Industry 5.0 to Forestry 5.0, highlighting the significance of human-centered artificial intelligence in addressing existing disparities. The research investigates the potential of AI technologies to promote sustainable forestry practices, emphasizing the collaboration between humans and machines to enhance efficiency and decision-making in forest management.

Personalization: Artificial intelligence models can be customized to align with particular le- gal standards, including jurisdiction-specific language, relevant case law, and specialized legal terminology.

Enhanced Collaboration: Legal professionals collaborate with artificial intelligence to vali- date, enhance, and deliver nuanced judgments, utilizing natural language processing models for labor-intensive activities such as document classification, extraction, and summarization.

The framework for Explainable AI in Industry 5.0, as presented by (Trivedi *et al.,* 2024), delineates its vision, structural components, and potential future directions. This research delves into the significance of transparent and interpretable AI systems in enhancing collaboration between humans and AI, improving decision-making efficacy, and guaranteeing the ethical and reliable deployment of AI technologies in cutting-edge industrial applications.

In a study by Rozanec *et al.* (2023) advocate for a human-centric AI architecture specifically designed for applications in Industry 5.0, prioritizing the enhancement of collaboration between humans and machines. The authors explore various design principles, technological innovations, and integration methods for AI in advanced manufacturing, underscoring the significance of personalized and adaptive systems that contribute to improved productivity, creativity, and decision-making capabilities.

2.5 Role of NLP in Legal Text Recovery

Natural Language Processing (NLP) is crucial in streamlining the recovery of legal texts, facilitating the extraction of pertinent information from both structured and unstructured legal documents. The primary elements of this process encompass:

OCR-Based Legal Text Extraction: Optical Character Recognition (OCR) algorithms facilitate the transformation of scanned documents and images into text that can be processed by machines.

Named Entity Recognition (NER): This strategy identifies legal entities, which consist of corporate names, individual names, references to cases, and specific legal terminology.

Text Classification: Legal documents are classified into distinct sectors, such as criminal, civil, and corporate law, which facilitates more efficient retrieval and processing of these documents.

Contextual Reconstruction: The task of piecing together incomplete legal texts necessitates a contextual approach, aimed at safeguarding the legal coherence and integrity of the information presented.

Age-Based Visibility Rules: Legal documents are require to index based on age structure. This will align with target group, provide a concise format of preservation, and direct target information etc. Certain types of legal content, notably in delicate areas such as the rights of minors, require the establishment of visibility rules based on age. These regulations are crucial for ensuring that content is presented in a manner that is appropriate for the viewer's demographic background.

In a study, Frankenreiter and Nyarko (2022) analyze the influence of Natural Language Processing (NLP) on legal technology, with a specific emphasis on its role in civil justice. The authors investigate how NLP innovations are reshaping legal workflows, thereby enhancing efficiency, improving accessibility, and increasing the accuracy of legal decisions and document handling.

A propose methodology that employs natural language processing (NLP) techniques for the extraction of metadata in the context of legal text consolidation was examined (Spinosa *et al.,* 2009). The research emphasizes the application of NLP to facilitate the automated retrieval of pertinent metadata from legal documents, with the objective of enhancing the organization, accessibility, and management of legal information within digital frameworks.

2.6 Challenges in Legal Text Recovery

Although there have been notable improvements in the recovery of legal texts due to advancements in natural language processing (NLP), several challenges still endure.

Ambiguity in Legal Language: The intricate nature and ambiguities found within legal language require NLP models to tackle these challenges in order to minimize the risk of misinterpretation.

Domain-Specific Terminology: Given the abundance of specialized language in legal documents, it is essential for natural language processing systems to be trained on datasets tailored to the legal field in order to improve their accuracy.

Data Privacy and Security: The management of sensitive legal information necessitates the implementation of rigorous protocols to maintain confidentiality and adhere to data protection laws, such as the General Data Protection Regulation (GDPR).

A study by (Bommarito II *et al.,* 2021) explores LexNLP, a tool for natural language processing tailored for legal and regulatory texts. It highlights the utilization of NLP strategies to extract relevant information, enhance the efficiency of legal research, and confront the challenges posed by the interpretation of complex legal language. A study examines the improvement of legal case text representation through the

application of information extraction (IE) methodologies Bru¨ninghaus and Ashley (2001).

It underscores the role of IE in augmenting the analysis and interpretation of legal documents by automatically retrieving critical information, such as facts, rulings, and legal principles, which can assist in legal reasoning and the management of cases. Ashley (2018) investigates the prospects and difficulties involved in the automated interpretation of legal texts via artificial intelligence and natural language processing. The article outlines a range of methods for extracting pertinent legal information, including case facts and legal principles, and examines the challenges posed by the ambiguity and intricacy of legal language.

2.7 Technological Advancements in NLP for Legal Text Recovery

The advent of Industry 5.0 marks a significant evolution in the development of sophisticated AI models that integrate deep learning methodologies, transfer learning, and knowledge graphs. This integration enhances the capability of natural language processing systems to retrieve legal texts with greater efficiency.

Transformer Models like BERT, GPT etc: These models possess the capability to interpret the context of legal documents, thereby enhancing comprehension, facilitating summarization, and improving information extraction.

Deep Learning Architectures: Employed for the training of models on extensive collections of legal documents, this approach enhances efficacy in various tasks, including text generation, sentiment analysis, and text summarization.

Multimodal Learning: The incorporation of textual, auditory, and visual materials from legal sources aims to enhance the precision of recovery across various media formats.

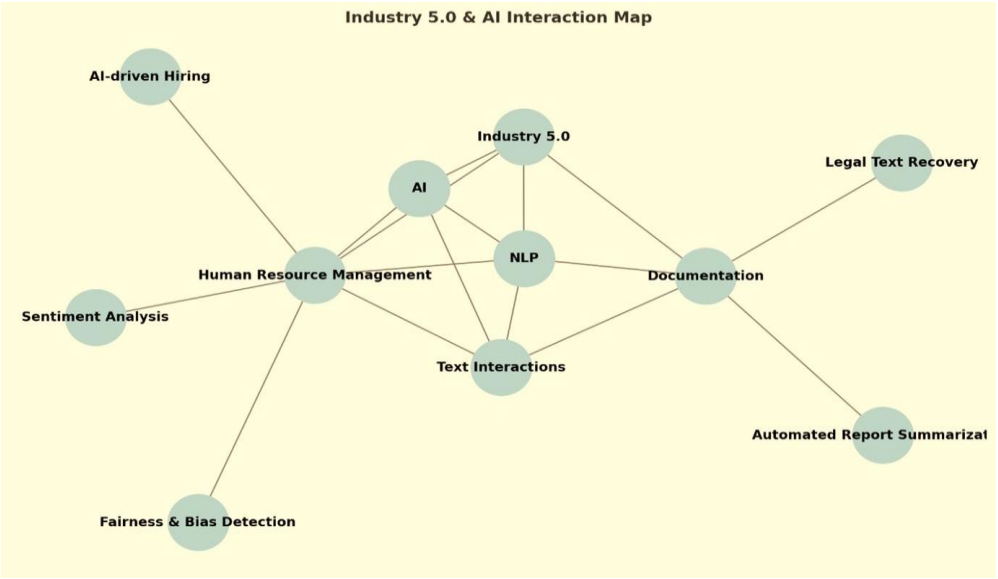
In a work by Moens (2001) delves into cutting-edge approaches for retrieving legal texts by lever- aging artificial intelligence and natural language processing technologies. The paper outlines various techniques designed to optimize search and information retrieval processes in legal databases, while also addressing challenges such as the complexity of legal terminology, the structuring of documents, and the prioritization of relevance, all aimed at bolstering the effectiveness of legal research.

Ariai and Demartini (2024) present an extensive overview of Natural Language Processing (NLP) applications within the legal field. Their analysis encompasses essential NLP tasks, relevant datasets, and models employed for the processing of legal texts, while also addressing significant challenges including ambiguity, domain specificity, and ethical considerations. The authors underscore recent advancements and propose future research trajectories in the realm of legal NLP.

2.8 Industry 5.0 & AI Interactive Integration

The illustration represents the connections among 'Industry 5.0', 'AI', 'NLP', 'Text Interactions', 'Documentation', and 'Human Resource Management (HRM)'. 'Industry 5.0' is associated with AI and NLP, which propel the development of automation and intelligent systems.

Figure 1. *Industry 5.0 & AI Interactive Integration*



Source: *Own study.*

The roles of 'NLP' and 'Text Interactions' are pivotal in facilitating sentiment analysis, processing documents, and recovering legal texts. 'Documentation' is integral to the organization of structured knowledge, while 'HRM' benefits from AI applications in recruitment, bias detection, and performance assessment. This framework emphasizes the collaborative relationship between humans and AI in the context of Industry 5.0 and its influence on business processes.

Khosravy *et al.* (2023) examine the intersection of human-collaborative artificial intelligence and Industry 5.0, highlighting the significance of social values and the ethical deployment of AI. Their survey presents an overview of contemporary advancements in AI-human collaboration, concentrating on cognitive systems and sustainable industrial innovations.

The study yields important insights regarding the future of industries that are increasingly driven by human-centric AI methodologies. In their study, (Martini *et al.,* 2024) explore the integration of human-centered and sustainable AI in the framework of Industry 5.0, identifying critical challenges and future directions. The authors underscore the pivotal role of AI

in cultivating ethical, sustainable, and collaborative industrial ecosystems. Their findings provide essential insights into achieving a balance between technological innovation and social as well as environmental responsibilities in emerging industries.

3. Method in Legal Text Recovery

This section outlines a series of chronological mathematical procedures that can be employed in the retrieval of legal texts. The methodology presented below integrates various natural language processing techniques designed to improve the efficacy of legal text recovery.

The mathematical representation of the Unified Tseg Function is summarized as follows:

$$T_{\text{Recovered}} = P\_T(R\_T(C\_T(N\_T(S\_T(E\_T(T))))))$$

where:

- ✓ (E_T) → OCR Extraction
- ✓ (S_T) → Segmentation
- ✓ (N_T)→ Named Entity Recognition (NER)
- ✓ (C_T) → Contextual Reconstruction
- ✓ (R_T) → Retrieval
- ✓ (P_T) → Post-processing (Correction, Summarization)

The subsequent subsections will provide a comprehensive discussion on how each phase contributes to the recovery of legal texts. These steps are fundamentally rooted in the principles of natural language processing.

Step 1: OCR-Based Legal Text Extraction

Transform scanned or deteriorated legal documents into a format that can be processed by machines. An image (Legal Text Recovery Document) I serves as the input for the OCR function, which facilitates the extraction of text (Legal Text Preview of first 1000 Words in Contract).

$$E_T(I) = \arg\max_X P\,(X|I)$$

Here, X signifies the text that has been acknowledged, and $P(X|I)$ indicates the probability that this text is correct. *Optical character recognition (OCR)* employs probability models to enhance its recognition capabilities.

$$P(X|I) = \prod_{i=1}^{n} P(x_i|I)$$

Here, every x_i denotes a character that has been acknowledged.

Ost-processing Correction: Implement the Levenshtein Distance method to amend words that have been misrecognized. Error model:

$$\text{Edit Distance} = \sum_{i=1}^{n} \delta(A_i, B_i)$$

where A is OCR text and B is ground truth.

Metrics: Character Error Rate (CER):

$$CER = \frac{\text{Insertions} + \text{Deletions} + \text{Substitutions}}{\text{Total Characters in Ground Truth}}$$

Character Error Rate (CER) is a metric used to evaluate the accuracy of character recognition systems by measuring the frequency of incorrect character predictions relative to the total number of characters in a given text. Word Error Rate (WER)

$$WER = \frac{\text{Total Errors}}{\text{Total Words in Ground Truth}}$$

Word Error Rate (WER) is an important evaluation criterion in the field of speech recognition, representing the ratio of erroneous words identified in a transcription when compared to an accurate reference.

Step 2: Text Segmentation (Tseg)

Segment legal materials based on age, field, jurisdictional context, and the format of presentation. Tokenization}

$$S_T(T) = \{S_1, S_2, \ldots, S_n\}$$

In this instance, S_i\refers to a particular segment that has been delineated.

Segmentation Through Contextual Attributes: Utilizing a Hidden Markov Model (HMM) to detect boundaries.

$$P(B|W) = \frac{P(W|B)P(B)}{P(W)}$$

In this scenario, B is identified as a boundary, whereas W is characterized as a word. BERT-Based Semantic Segmentation: Determine the cosine similarity for the embeddings of various sentences.

$$\cos(\theta) = \frac{A \cdot B}{|A||B|}$$

In this $scenario, \$A\$$ and B are sentences that have been transformed into their vectorized representations. The accuracy of segmentation is determined by analyzing key metrics, namely F1-score, precision, and recall.

Step 3: Named Entity Recognition (NER) for Legal Terms

Determine legal entities such as case law, relevant dates, names, and legislative statutes. Conditional Probability: Understanding conditional probability is essential for effectively identifying entities in various contexts.

$$P(E_i|S) = \frac{P(S|E_i)P(E_i)}{P(S)}$$

Here, E_i refers to an entity, and S signifies the text that has undergone segmentation.

Utilizing a transformer architecture for named entity extraction, the model processes the text T and determines a probability for each token, reflecting its potential categorization as an entity.

$$P(E|T) = \sum_{i=1}^{n} P(e_i|t_i)$$

In this context, $\$t\_i\$$ represents a token. Metrics for Evaluating Named Entity Recognition: Analyzing Precision, Recall, and F1-score.

Step 4: Contextual Text Reconstruction and Masked Language Modeling (MLM)

Reconstruct omitted terms or statements while ensuring the legal context remains intact. Masked Language Modeling (MLM): Masked Language Modeling (MLM) is a method utilized in the field of natural language processing that involves concealing specific words within a sentence.

$$P(x_m|X_{context}) = \frac{e^{h_{x_m}}}{\sum_j e^{h_{x_j}}}$$

In this context, h_x represents the concealed state associated with the token x.

Attention Mechanism (Transformer):The Attention Mechanism, a fundamental component of the Transformer architecture, facilitates the model's ability to focus on specific parts of the input data, thereby enhancing its performance in various tasks such as natural language processing and machine translation:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_k \exp(e_{ik})}$$

Contextual embedding update: Modification of contextual embedding framework:

$$h_i = \sum_j \alpha_{ij} v_j$$

Recursive Sentence Reconstruction: Recursive Sentence Reconstruction refers to the process of systematically reassembling sentences by utilizing recursive techniques:

$$C_T(S) = \sum_{i=1}^{n} P(S_i|S_{context})$$

Perplexity (PPL), BLEU, and ROUGE are metrics commonly employed in the evaluation of natural language processing models, particularly in tasks related to language generation and translation.

Step 5: Retrieval-Based Legal Text Recovery

Extract absent provisions from legal databases. BM25 for Legal Text Search: Implementing BM25 for the Retrieval of Legal Information.

$$R_T(Q) = \arg\max_D \sum_{i=1}^{n} BM25(Q, D_i)$$

BM25 score: The score derived from the BM25 algorithm:

$$BM25 = \sum_{i=1}^{n} \frac{IDF(t_i) f(t_i, D)(k_1 + 1)}{f(t_i, D) + k_1 \left(1 - b + b \cdot \frac{|D|}{avgD}\right)}$$

Semantic Search Using SBERT: Employing SBERT for Semantic Search significantly improves the retrieval process by focusing on the semantic relationships between sentences, thereby facilitating more accurate and contextually relevant search results.

$$S(A,B) = \frac{A \cdot B}{|A||B|}$$

MRR Mean Reciprocal Rank (MRR) and Normalized Discounted Cumulative Gain (NDCG) are two important metrics used for evaluating the effectiveness of information retrieval systems.

Step 6: Text Correction, Normalization, Summarization

Correct spelling errors, standardize legal terminology, and provide a concise summary of the text. Correction Using Levenshtein Distance: The implementation of correction techniques can be significantly enhanced by employing Levenshtein Distance.

$$C_T(A,B) = \sum \delta(A_i, B_i)$$

(Seq2Seq) Summarization via Sequence-to-Sequence (Seq2Seq):

$$P(Y|X) = \prod_{t=1}^{T} P(y_t|y_{1:t-1}, X)$$

ROUGE Score for Summary Evaluation: Utilizing the ROUGE Score for the Assessment of Summaries.

$$ROUGE - N = \frac{\sum_{match \in generated} count(match)}{\sum_{match \in reference} count(match)}$$

ROUGE, BLEU, Edit Distance: The aforementioned six steps present a succinct summary of the Unified Tseg Function aimed at recovering legal texts. This integrated process is known as Tseg, which denotes text segmentation, and consists of a series of mathematical formulas employed in the field of natural language processing (NLP).

An introduction to an automated methodology for the extraction of label data from images of herbarium specimens, employing Optical Character Recognition (OCR) and Named Entity Recognition (NER) was examined (Takano *et al.,*

2024). This innovative approach significantly enhances the process of database creation, thereby increasing the efficiency of botanical research.

The research showcases sophisticated techniques for the digitization and systematic organization of herbarium collections. Marti and Bunke (2001) propose a novel system aimed at achieving writer-independent handwriting recognition, concentrating on the segmentation of text lines and the recognition of individual words.

Their approach significantly boosts the accuracy of text extraction, thereby facilitating improved document analysis. This research is a significant contribution to the field of automated handwriting recognition, accommodating a variety of writing styles.

Takagi (2024) investigates the effects of diverse masking strategies within Masked Language Modeling (MLM) as applied to text-based person search. The research examines the ways in which different masking methods affect the performance of models, thereby enhancing the precision of retrieving person-related information from textual sources.

The work of ghosh2022indian focuses on a text normalization strategy designed for the summarization of Indian legal texts. Their methodology improves the readability of complex legal documents and supports effective information extraction. This study significantly contributes to the field of automated legal text processing and summarization.

4. Results

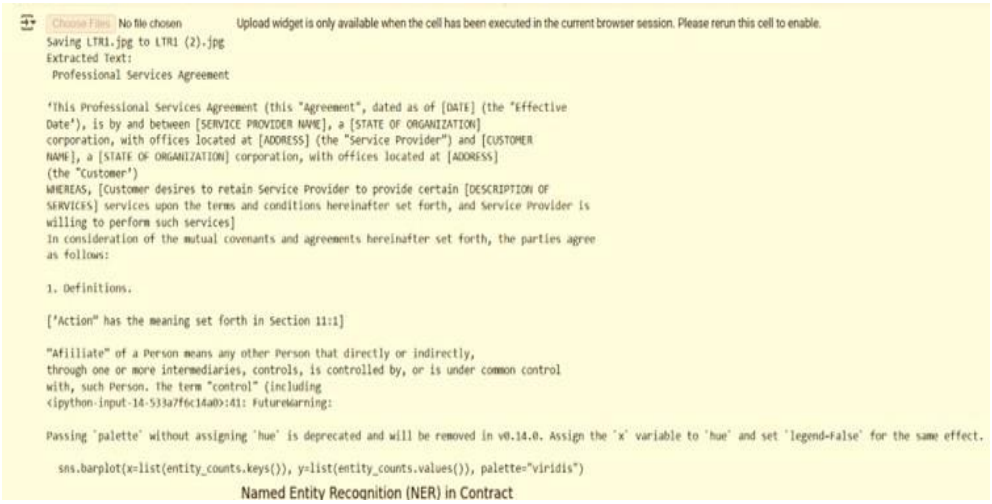
In this research, we explore and implement a re-burst technique in the domain of Natural Language Processing (NLP). Our focus is on recovering text data from images and analyzing it effectively. We combine several NLP methodologies to create an AI system designed to streamline the analysis of text data essential for decision-making processes in administrative sectors, particularly concerning legal documents.

The following paragraphs delineate the key steps taken during this study, as illustrated in the mathematical framework outlined in the methodology section. Additionally, this section showcases the results of our research through various visual representations.

4.1 Steps in Legal Text Analysis

Image Upload: An image file is systematically and manually uploaded to Google Colab through the use of files.upload (Figure 6), which enables the integrated code to access and manipulate the image.

Figure 2. *Legal Text Preview of first 1000 Words in Contract*



Source: *Own study.*

The extraction of information, thereby improving the efficiency of tasks such as summarizing texts, retrieving legal documents, and classifying various types of documents

OCR: During this stage, we utilized the code to extract textual information from the image of the legal document that was uploaded. The process of Optical Character Recognition (OCR) converts the visual data present in the image into a format that machines can read, thereby allowing for further analytical procedures, including Named Entity Recognition (NER).

Text Preview: The code outputs the initial 1000 characters of the text extracted from the legal document. This procedure enables us to assess and confirm that the optical character recognition (OCR) process has effectively retrieved pertinent information from the image.

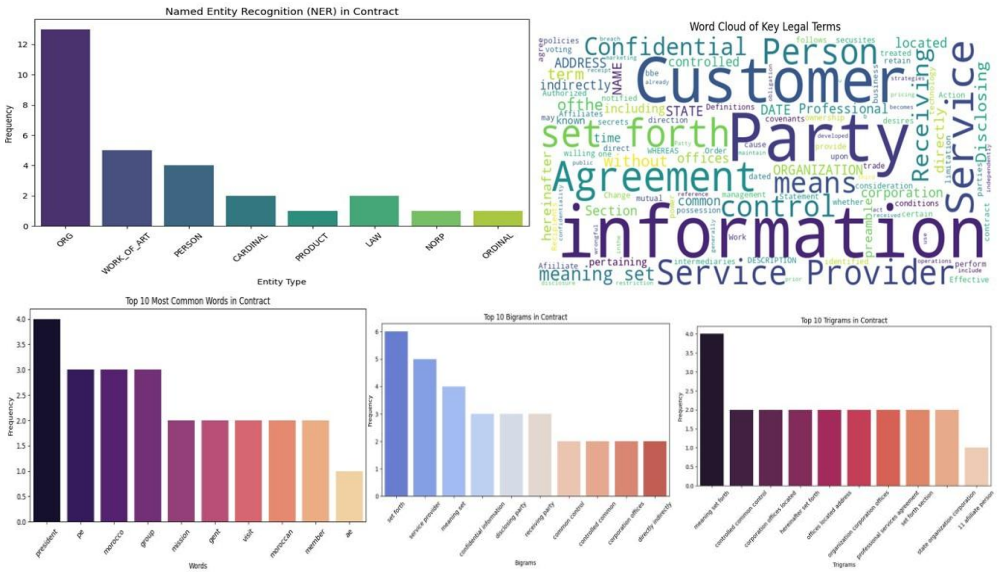
NLP Processing: The text that has been extracted undergoes processing through spaCy's natural language processing model (en core web sm). This involves tokenization and an analysis of diverse linguistic characteristics, thereby facilitating tasks such as Named Entity Recognition (NER).

This section provides the capability to segment documents into coherent legal units (e.g., clauses, articles, jurisdictions) by first reconstructing any missing or damaged sections. Deep learning models, like transformers, employ this recovered text to precisely segment, label, and classify legal content according to its structure, meaning, visibility of age, and jurisdiction.

Findings Analysis: This approach ensures that any missing clauses, citations, or contextual information are restored with precision, while also upholding the legal integrity and structural organization of the original text.

Named Entity Recognition (NER): The named entity recognition (NER) model integrated within SpaCy is capable of detecting named entities such as individuals, organizations, and dates within a given text. Once identified, these entities are extracted for subsequent categorization and analysis.

Figure 3. *(**a**) Named Entity Recognition (NER) in a Legal Contract, (**b**) Word Cloud of Key Legal Terms, (**c**) Top Ten Most Common Words in a Contract, (**d**) Top 10 Bigrams in Contract Text, (**e**) Top 10 Trigrams in Contract Text.*



Notes: *(a) Named Entity (b) Word Cloud (c) Top 10 Words (d) Bigrams (e) Trigrams*
Source: *Own study.*

Figure 3(a) Entity Visualization: Entity Visualization elucidates the prominence and distribution of these entities from the legal text document. The frequency of various entity types—namely persons, organizations, and dates—is represented in a bar chart generated with seaborn.

This technique significantly contributes to the extraction of information, thereby optimizing activities such as summarizing texts, recovering legal documents, and classifying documents effectively. Named Entity Recognition (NER) represents a crucial methodology in the field of Natural Language Processing (NLP), aimed at detecting and classifying specific entities, including names, dates, and geographical locations, found in written text.

Word Cloud: A word cloud is created utilizing the WordCloud tool. This visualization highlights the most commonly occurring terms within the extracted text. The word cloud displayed here visually conveys the frequency of words within the legal text, with larger words indicating a higher rate of occurrence.

Figure 3(b) Word Cloud. Word cloud thereby facilitating the identification of significant legal terminology or concepts according to their prominence within the document.

In the analysis of text recovery, this approach assists in identifying key terms, patterns, and themes, which is instrumental in understanding context, revealing missing information, or recognizing important entities. By simplifying text interpretation, word clouds prove to be beneficial tools in legal, medical, and other data-intensive sectors.

Common Words: We used another line of the code to identifies and extracts the most commonly occurring non-stopwords, which are significant to legal text context, from the analyzed document.

Figure 3(c) Common Words. These identified terms are then represented in a bar chart, emphasizing the key concepts that frequently appear within the document, thereby facilitating a more effective text analysis.

N-grams Extraction: We use another line code utilizes CountVectorizer to extract bigrams, which are pairs of words, and trigrams, which consist of three-word combinations, from the text. These multi-word constructs offer enhanced contextual insights and are instrumental in recognizing prevalent phrases or terminology within the document.

Bigrams: The top ten identified entities, along with the most common bigrams a are presented.

Figure 3(d) Bigrams. This step summarizes vital information and facilitates the readers and user's examination of significant entities and phrases obtained from the legal text document.

Trigrams: The top ten identified entities, along with the most common

trigrams, are presented.

Figure 3(e) Trigrams. This phase provides a concise overview of essential data, enabling the user to review significant entities and phrases derived from the document.

The experiments outlined in this section are detailed in the subsequent paragraphs, which we deemed are better informed by a selection of figures and tables.

4.2 Word Embeddings Correlation Matrix

The presented table illustrates a correlation matrix for various legal terms, elucidating the interrelationships among them. It serves as a tool for examining the frequency with which terms such as ”the,” ”of,” ”to,” and ”party” appear together in legal documents.

Table 1. *Selected Correlation Matrix Terms and scores*

	the	of	to	by	and	or	information	is	party
the	1.000	0.561	0.325	0.139	-0.229	-0.268	-0.163	-0.098	0.389
of	0.561	1.000	0.354	-0.169	-0.554	-0.384	-0.166	0.256	0.046
to	0.325	0.354	1.000	-0.063	-0.321	-0.143	-0.059	0.445	-0.095
by	0.139	-0.169	-0.06	1.000	-0.366	0.096	-0.083	0.314	0.218
and	-0.22	-0.554	-0.32	-0.366	1.000	0.237	0.600	-0.290	-0.317
forth	-0.72	-0.525	0.048	-0.239	0.029	. . .	-0.207	-0.117	0.381
notified	-0.20	0.240	-0.33	0.127	-0.319	-0.196	0.391	-0.302	0.020
persons	-0.11	0.089	-0.13	0.058	-0.196	-0.106	-0.144	-0.078	0.128
identified	0.381	-0.250	-0.21	0.393	0.019	-0.195	-0.173	0.009	0.014

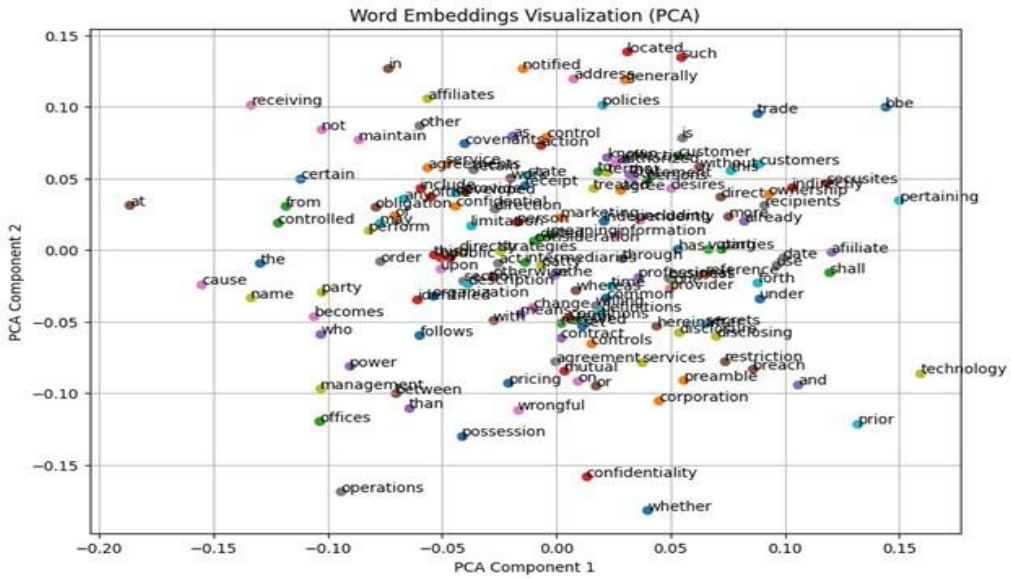
Source: *Own study.*

In Table 1 the analysis is crucial for the recovery of legal texts in the fields of natural language processing (NLP) and artificial intelligence (AI), as it facilitates the understanding of syntactic structures, the identification of significant terms, and the enhancement of models for tasks including information retrieval, entity recognition, and text summarization. Elevated correlations indicate prevalent legal phrasing, thereby improving AI’s grasp of the patterns inherent in legal language.

4.3 Principal Component Analysis (PCA)

The visualization of word embeddings through Principal Component Analysis (PCA) facilitates the reduction of high-dimensional word vectors into two or three dimensions, thereby enhancing the clarity of word relationships.

Figure 4. *Principal Component Analysis*



Source: *Own study.*

In Figure 4 Principal Component Analysis captures the variance present in the data. Principal Component Analysis (PCA) effectively projects information into a lower- dimensional space, which helps in the representation of semantic similarities and groupings among words. This technique is instrumental in comprehending word associations and clustering within natural language processing tasks.

5. Brief Case Study: Legal Text Format for Structural Analysis

NLP models proficient in managing the complexities of legal terminology will facilitate a more streamlined and collaborative legal experience for both legal professionals and end-users in the age of artificial intelligence.

Typically, the structure of current legal documents consists of a title that designates the document, a preamble that describes its purpose and the authority under which it operates, definitions that elucidate key terms, main provisions that delineate obligations, rights, and regulations, along with exceptions and limitations. Furthermore, it encompasses enforcement clauses and final provisions that cover amendments and the specifics of enactment.

There is need to advance more distinct techniques to enable this applications better assist users in times of distress. Some of this distinct processes require age structure legal text indexing, legal text display format and specificity of legal text based on age structure to enable quicker recovery processes.

$$W_t = f(W_y, W_r, W_n, W_o, W_h)$$

Based on the given legal text quote below we analyzed its standard structure format using the formula above:

"Funding from sources other than the state subsidy must secure a scholarship for candidates in an amount equal to or higher than the amount set out in Art. 209 of the Higher Education and Science Law Act of 20 July 2018 (Journal of Laws year 2023, item 742 as further amended), hereinafter referred to as the Act, for a minimum period of three years."

APPLYING THE FORMULA:

Table 2. *Legal text Breakdown: Key Factors Contributions & Descriptions*

Factor	Legal Text Analytical Breakdown
W_y	Purpose: The scholarship amount must meet or exceed the requirement set by the Act
W_r W_n	Source of funding: Must come from sources other than the state subsidy
W_o	Duration: Scholarship must be secured for a minimum period of three years
W_h	Beneficiaries: Candidates (likely doctoral students) are the beneficiaries
	Legal reference: Article 209 of the Higher Education and Science Law Act (as amended)

Source: *Own study.*

Table 2 presented indicates that the aforementioned legal text lacks a specific target demographic index. This absence suggests a potential vulnerability in the recovery process associated with the legal text. It is essential to categorize the legal text according to various age structures. Further details regarding the age structure index of legal text will be elaborated upon in the following paragraphs.

LEGAL TEXT RECOVERY RULE (INDEXER):

The below text suggest an AI-Age based indexer rule for future legal text. We belief thus rule will further enhances the recovery speed in the situation where there is a missing data within legal text.

"Funding from sources other than the state subsidy must secure a scholarship for candidates in an amount equal to or higher than the amount set out in Art. 209 of the Higher Education and Science Law Act of 20 July 2018 (Journal of Laws year 2023, item 742 as further amended), hereinafter referred to as the Act, for a minimum period of three years.

$_A(24+:35+:40-)$."

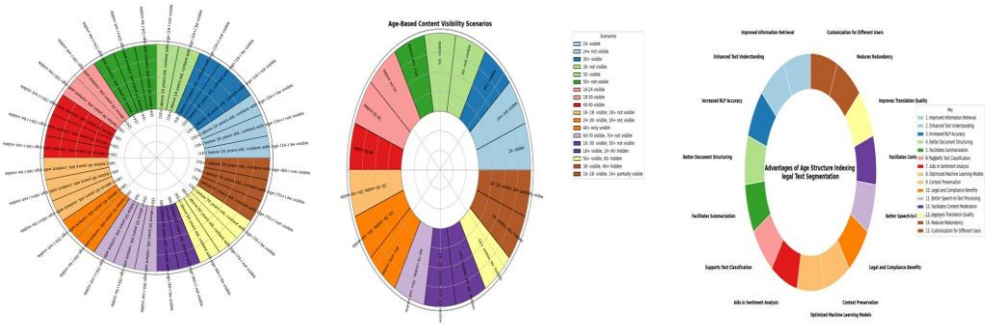
In the legal text above, (+) means above age, (-) mean below age while (:) mean in-between age or range in age.
$A^{(24+:35+:40-)}$

AGE-BASED CATEGORIZATION:

- **24+:** This applies to individuals who are 24 years of age and older.
- **35+:** This applies to individuals who are 35 years of age and older.
- **40-:** This applies to individuals who are 40 years old or younger.

Future legal texts should conventionally be structured to include a title, a preamble, definitions, essential provisions (articulating obligations, rights, and regulations), exceptions or limitations, en- forcement clauses, age specifications (indicating age-related restrictions or allowances), and concluding provisions (which incorporate amendments and details of enactment).

Figure 5. *Legal Text Recovery Based on Age Structure Indexing (A), Age-Based Content Visibility Scenarios (B), and Advantages of Age Structure Indexing for Legal Text Segmentation (C).*



Source: *Own study.*

5.1 Legal Text Recovery Based on Age Structure Indexing

Legal Text Recovery Based on Age Structure indexing is an AI-driven approach that encourages the indexation of legal text into serious age groups. This process allows a quicker restoration of missing legal text while ensuring age-appropriate content visibility to target group of users. It emphasize that legal documents be segmented into structured sections which is an applies NLP techniques to better structure concepts like Named Entity Recognition (NER), transformer-based text completion, and retrieval-based methods to recover missing words or sentences.

Figure 5A Legal Text Recovery Based on Age Structure Indexing. we briefly present some suggestions to how legal text can be structure across different age groups.

Significance in AI and NLP Advancements: This technique enhances the ability of machines to comprehend sophisticated legal language, leading to improved document analysis, automation of compliance, and greater accessibility. It ensures the ethical use of AI by incorporating visibility controls for age-restricted content, which allows AI to adapt legal information for various age categories. This innovation significantly enhances the efficiency of natural language processing in legal technology, supporting automated legal summarization, content moderation, and the reduction of bias, thereby making AI-driven legal systems more responsive, secure, and ethically sound.

Age-Based Content Visibility Scenarios: Age-based content visibility scenarios involve the enforcement of the indexation of legal text with stricker regulation access specific age group that legal text is targeting.

Figure 5B Age-Based Content Visibility Scenarios. These measures are designed to ensure that individuals are exposed to information that is suitable for their age, giving more preservation of content by those directly involve. Should technology faills to recover, this particular age group will assist in the recovery process quicker and faster.

This approach not only promotes compliance with legal standards but also enhances user safety and supports ethical practices in AI-driven content moderation and circulation.

Advantages of Age Structure Indexing Legal Text Segmentation: The following outlines a carefully developed step-by-step advantages of the need for framework designed to facilitate efficient segmentation and recovery.

Figure 5C Advantages of Age Structure Indexing legal Text Segmentation. To advance the chrono- logical methodology for Text Segmentation (Tseg) in legal text recovery, it is imperative to integrate text target groups, which include ”age, sectors, foreign and domestic,” as well as the text display format into the overall process.

6. Discussion

The proposed Unified Tseg Deep Learning Function improves the recovery of legal texts in four essential domains: Our cohesive framework seeks to align with the goal of rapidly restoring lost content while upholding legal logic, jurisdictional parameters, and age appropriate visibility across different age groups.

- **Hierarchical Text Segmentation for Legal Taxonomy:** In contrast to conventional text segmentation methods, our model organizes

documents into distinct legal categories (e.g., civil, crimi- nal, labor) and enriches each segment with metadata: country, domain, and age-visibility classification (e.g., (16+), (18+)). This allows for context-aware filtering and the adaptation of content to suit different legal jurisdictions and age groups.

- **Transformer-Based Generative Recovery:** Utilizing a carefully fine-tuned encoder-decoder transformer (like T5-legal), the model reconstructs omitted sections based on the surrounding context, legal definitions, and case references. Attention mechanisms empower the model to recover syntax while preserving the integrity of legal semantics.

- **Combining Ontology and Rule-Based Post-Processing:** The retrieved content is aligned with a legal ontology to maintain logical consistency. This involves name entity consistency (laws, articles, institutions), temporal reasoning (validity periods), and the logic surrounding obligations and permissions.

- **AI-Driven Regulatory Operations in Industry 5.0:** Our framework provides support for automated compliance reporting, the regeneration of legal documents, and the restriction of content by age or jurisdiction. For example, a contract clause considered unsuitable for audiences younger than 18 can be automatically masked or rewritten to ensure it is appropriate for readability.

Table 3. *Example Output of Legal Text Recovery using Unified Tseg Deep Learning Function*

Input	*According to Article [MASK], the tenant has the right to terminate the contract in case of. . .*
Recovered Text	According to Article 14b of the Civil Code, the tenant has the right to terminate the contract in case of severe sanitary hazards or structural failure.
Age Annotation	(18+)

Source: *Own study.*

Table 3 demonstrates a case of legal text recovery through the Unified Tseg Deep Learning Function, illustrating the precise reconstruction of a masked legal clause with contextual details and annotated for suitable age-based visibility (18+).

Industry 5.0 represents a paradigm shift that merges human creativity with artificial intelligence, significantly improving the recovery of legal texts through sophisticated Natural Language Processing (NLP) methodologies. AI-enhanced NLP models facilitate the extraction, summarization, and restora- tion of incomplete or damaged legal documents by comprehending context, entities, and semantics. This synergy

optimizes legal workflows, promoting precision, efficiency, and accessibility in document management, while enabling human professionals to enhance and interpret the results generated by AI.

In a study conducted by Farook et al. (2024), titled ”Enlightening Justice: Empowering Society Through AI-Driven Legal Assistance,” delves into the implications of artificial intelligence in legal aid. The authors examine the ways in which AI can broaden access to legal services, refine legal procedures, and bolster the effectiveness of justice delivery. Their findings highlight the significant role of AI in empowering individuals by making legal assistance more readily available and operationally efficient.

Also, ”A Mutual Legal Assistance Case Study” by (Swire *et al.,* 2016) presents an analysis of a mutual legal assistance scenario involving the United States and France. The authors assess the legal frameworks and difficulties associated with international legal cooperation, drawing attention to the complexities inherent in international law, issues of data privacy, and the processes involved in fulfilling legal requests.

7. Future Prospects in Legal Text Recovery

As artificial intelligence and natural language processing continue to evolve, it is likely that future advancements will include:

Real-Time Legal Assistance: Systems utilizing artificial intelligence to deliver instantaneous recommendations and conduct document analysis for legal practitioners.

AI-Driven Legal Drafting: Automated legal document generation facilitates the integration of customizable clauses and terminology that are appropriate for the given context.

Cross-Jurisdictional NLP: Developing natural language processing systems to achieve seam- less integration across multiple legal jurisdictions and languages, facilitating the removal of im- pediments in international legal practice. Yao et al. (2025) discuss their research titled ”Intelligent Legal Assistant: An Interactive Clarification System for Legal Question Answering”. This paper outlines the development of an AI-based system that seeks to refine the legal question answering process by implementing interactive clarification tools. The system is designed to boost the accuracy and relevance of the responses, thus facilitating a more interactive and user-centric approach to legal support.

Also, Basha *et al.* (2024) examine the integration of generative artificial intelligence within the context of legal drafting in their work, ”Generative Artificial Intelligence in Legal Drafting.” The authors articulate how AI technologies can aid in the

automation of legal document generation, thereby improving efficiency, precision, and uniformity in legal drafting processes. This innovation serves to optimize legal operations and promote sustainability within the legal field.

Nithya *et al.* (2024) examine the application of AI-driven legal automation in their work, ”AI-Driven Legal Automation to Enhance Legal Processes with Natural Language Processing.” The authors underscore the importance of Natural Language Processing (NLP) in facilitating the automation of various tasks, including contract analysis, legal research, and document review, which collectively contribute to greater efficiency and accuracy in legal operations.

Mah (2024b) offers an objective text categorization strategy for digital content associated with IoT, utilizing a Word-to-Graph model. This research delves into the role of graph-based representations in refining text classification by reducing bias. The outcomes indicate a significant improvement in the accuracy of digital content categorization, thus making the model applicable to a wide range of IoT applications.

Tyss *et al.* (2024) present a study entitled ”Beyond Borders: Investigating Cross-Jurisdiction Transfer in Legal Case Summarization,” in which they analyze the intricacies involved in the transfer of legal case summaries between different legal systems. Their research highlights the obstacles and prospects inherent in this process, aiming to advance the effectiveness of AI models in producing legal summaries that are relevant across multiple jurisdictions.

8. Conclusion

The Unified Tseg Deep Learning Function offers a unique and practical solution for the recovery of legal texts, addressing the challenges of incompleteness, segmentation, and the complexities of jurisdiction present in legal documents. By combining hierarchical text segmentation, transformer- based generation, and alignment with legal ontologies, this framework enables precise reconstruction and regulation of content that is sensitive to age.

This advancement significantly supports the vision of Industry 5.0, where AI enhances human judgment in regulatory activities, compliance, and digital legal services. Future work will focus on expanding multilingual capabilities, enabling real-time deployment in legal platforms, and integrating blockchain for traceable legal audits.

Industry 5.0 signifies a transformative moment in digital technology that combines human creativity with artificial intelligence, thereby enhancing the quick recovery of legal texts through innovative Natural Language Processing (NLP) strategies. AI-driven NLP frameworks are instrumental in extracting, summarizing, and recovering

lost or damaged legal content by effectively understanding context, recognizing entities, and interpreting semantics.

The integration of Natural Language Processing, artificial intelligence, and Industry 5.0 has opened up a novel domain for the recovery of legal texts. With the ongoing maturation of these technologies, we are likely to see considerable enhancements in the processing of legal documents, which will contribute to making legal services more accessible, efficient, and dependable. NLP models capable of addressing the complexities inherent in legal language will empower both legal experts and users, fostering a more streamlined and collaborative legal experience in the era of AI.

Availability of data and material used: We used this legal text code and document (Figure 6) to analyze and experiment our model. All other information underlying analysis used to developed the results are available as part of the article and no additional source data are required or reserved somewhere.

Competing Interest: No conflict of Interest.

Funding: No funding.

References:

Alberts, H., Ipek, A., Lucas, R., Wozny, P. 2020. Coliee 2020: Legal information retrieval and entailment with legal embeddings and boosting. JSAI International Symposium on Artificial Intelligence, 211-225.

Ariai, F., Demartini, G. 2024. Natural language processing for the legal domain: A survey of tasks, datasets, models, and challenges. arXiv preprint arXiv:2410.21306.

Ashley, K.D. 2018. Automatically extracting meaning from legal texts: Opportunities and challenges.
Ga. St. UL Rev., 35, 1117.

Basha, T.Y., Kalyani, B., Sandeep, Y. 2024. Generative artificial intelligence in legal drafting. 2024 International Conference on Computational Intelligence for Green and Sustainable Tech- nologies (ICCIGST), 1-6.

Bommarito II, M.J., Katz, D.M., Detterman, E.M. 2021. Lexnlp: Natural language processing and information extraction for legal and regulatory texts. In: Research handbook on big data law (pp. 216-227). Edward Elgar Publishing.

Brooks, C., Gherhes, C., Vorley, T. 2020. Artificial intelligence in the legal sector: Pressures and challenges of transformation. Cambridge Journal of Regions, Economy and Society, 13 (1), 135-152.

Bru¨ninghaus, S., Ashley, K.D. 2001. Improving the representation of legal case texts with in- formation extraction methods. Proceedings of the 8th international conference on Artificial Intelligence and Law, 42-51.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I. (2019). Extreme multi-label legal text classification: A case study in eu legislation. arXiv preprint arXiv:1905.10892.

Costantini, C. 2025. Comparative law and the breaking of the vessels. International Journal for the Semiotics of Law-Revue internationale de S´emiotique juridique, 38 (1), 163-175.

Dong, X., Li, W., Le, Y., Jiang, Z., Zhong, J., Wang, Z. 2025. Termdiffusum: A term-guided diffusion model for extractive summarization of legal documents. Proceedings of the 31st International Conference on Computational Linguistics, 3222-3235.

Fabbrini, F. 2025. The recovery and resilience facility as a new legal technology of european gover- nance. Journal of European Integration, 47 (1), 85-103.

Farook, A.M., Kingston, W., Kannaiah, S.K. 2024. Enlightening justice: Empowering society through AI driven legal assistance. 2024 Second International Conference on Advances in Information Technology (ICAIT), 1, 1-7.

Frankenreiter, J., Nyarko, J. 2022. Natural language processing in legal tech. Legal Tech and the Future of Civil Justice (David Engstrom ed.) Forthcoming.

Holzinger, A., Schweier, J., Gollob, C., Nothdurft, A., Hasenauer, H., Kirisits, T., H¨aggstr¨om, C., Visser, R., Cavalli, R., Spinelli, R. 2024. From industry 5.0 to forestry 5.0: Bridging the gap with human-centered artificial intelligence. Current Forestry Reports, 10 (6), 442-455.

Joseph, T., Vijayalakshmi, A. 2025. Natural language processing for legal document classification. In: Enhancing communication and decision-making with AI (pp. 295-316). IGI Global.

Khosravy, M., Gupta, N., Pasquali, A., Dey, N., Crespo, R.G., Witkowski, O. 2023. Human-collaborative artificial intelligence along with social values in industry 5.0: A survey of the state-of-the-art. IEEE Transactions on Cognitive and Developmental Systems, 16(1), 165-176.

Mah, P.M. 2024a. National AI strategies. European Research Studies Journal, XXVII, (4B), 96-115. DOI: 10.35808/ersj/3565.

Mah, P.M. 2024b. Unbiased text categorization in iot-based digital content using a word-to-graph model. Procedia Computer Science, 251, 31-40.

Mania, K. 2022. The digital transformation of legal industry: Management challenges and technological opportunities. ?????

Marti, U.V., Bunke, H. 2001. Text line segmentation and word recognition in a system for general writer independent handwriting recognition. Proceedings of Sixth International Conference on Document Analysis and Recognition, 159-163.

Martini, B., Bellisario, D., Coletti, P. 2024. Human-centered and sustainable artificial intelligence in industry 5.0: Challenges and perspectives. Sustainability, 16(13), 5448.

Meyer, A., Mullen, A., Tomlin, R. 2025. Slavery on the northern frontier: A stylus tablet from vindolanda. Britannia, 1-23.

Moens, M.F. 2001. Innovative techniques for legal text retrieval. Artificial Intelligence and Law, 9, 29-57.

Muller, W.P. 1990. The recovery of justinian's digest in the middle ages. Bull. Medieval Canon L., 20.

Nithya, M., Harini, S., Kavyadharshini, S., Srinidhi, K. 2024. Ai-driven legal automation to enhance legal processes with natural language processing. 2024 International Conference on IoT Based Control Networks and Intelligent Systems (ICICNIS), 1246-1253.

Onami, E., Miyanishi, T., Maeda, K., Kurita, S. 2025. Legalviz: Legal text visualization by text to diagram generation. arXiv preprint arXiv:2502.06147.

Pl-onka, M., Kocot, K., Holda, K., Daniec, K., Nawrat, A. 2025. A comparative evaluation of the effectiveness of document splitters for large language models in legal contexts. Expert Systems with Applications, 126711.

Pusztahelyi, R., Stef˝an, I. 2024. Improving industry 4.0 to human-centric industry 5.0 in light of the protection of human rights. 2024 25th International Carpathian Control Conference (ICCC), 1-6.

Ro˘zanec, J.M., Novalija, I., Zajec, P., Kenda, K., Tavakoli Ghinani, H., Suh, S., Veliou, E., Papamartzi- vanos, D., Giannetsos, T., Menesidou, S.A. 2023. Human-centric artificial intelligence architecture for industry 5.0 applications. International journal of production research, 61 (20), 6847-6872.

S´anchez, D.B., Garc´ıa, N.A., Jim´enez, B., Nieto, M.G., Morales, P.M., Salas, N.S., Hern´an, C.G., Coll, P.H., Ponsoda, E.M., Ib´an˜ez, P.C. 2025. Mel: Legal spanish language model. arXiv preprint arXiv:2501.16011.

Shiel, P., Rehbein, M., Keating, J. 2009. The ghost in the manuscript: Hyperspectral text recovery and segmentation. Codicology and Palaeography in the Digital Age, 1, 159-74.

Siino, M., Falco, M., Croce, D., Rosso, P. 2025. Exploring llms applications in law: A literature review on current legal nlp approaches. IEEE Access.

Smits, R. 2014. Is my money safe at european banks? reflections on the 'bail-in'provisions in recent eu legal texts. Capital Markets Law Journal, 9(2), 137-156.

Spinosa, P., Giardiello, G., Cherubini, M., Marchi, S., Venturi, G., Montemagni, S. 2009. NLP- based metadata extraction for legal text consolidation. Proceedings of the 12th international conference on artificial intelligence and law, 40-49.

Swire, P., Hemmings, J.D., Vergnollie, S. 2016. A mutual legal assistance case study: The united states and france. Wis. Int'l LJ, 34, 323.

Takagi, Y. 2024. Effect of masking strategies in masked language modeling used for text-based person search. Master's thesis Master's Programme in Imaging and Light in Extended Reality (IMLEX) School of Computing University of Eastern Finland.

Takano, A., Cole, T.C., Konagai, H. 2024. A novel automated label data extraction and data base generation system from herbarium specimen images using ocr and ner. Scientific Reports, 14(1), 112.

Trivedi, C., Bhattacharya, P., Prasad, V.K., Patel, V., Singh, A., Tanwar, S., Sharma, R., Aluvala, S., Pau, G., Sharma, G. 2024. Explainable AI for industry 5.0: Vision, architecture, and potential directions. IEEE Open Journal of Industry Applications.

Tyss, S., Venkatkrishna, V., Ghosh, S., Grabmair, M. 2024. Beyond borders: Investigating cross- jurisdiction transfer in legal case summarization. Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 4136-4150.

Weichbroth, P. 2025. AI and the law: Evaluating chatgpt's performance in legal classification. arXiv preprint arXiv:2502.12193.

Yao, R., Wu, Y., Zhang, T., Zhang, X., Huang, Y., Wu, Y., Yang, J., Sun, C., Wang, F., Liu, X. 2025. Intelligent legal assistant: An interactive clarification system for legal question answering. arXiv preprint arXiv:2502.07904.

Legal Text Recovery Document.

Figure 6. *Legal Text Recovery Document*

Professional Services Agreement

This Professional Services Agreement (this "**Agreement**"), dated as of [DATE] (the "**Effective Date**"), is by and between [SERVICE PROVIDER NAME], a [STATE OF ORGANIZATION] corporation, with offices located at [ADDRESS] (the "**Service Provider**") and [CUSTOMER NAME], a [STATE OF ORGANIZATION] corporation, with offices located at [ADDRESS] (the "**Customer**").

WHEREAS, [Customer desires to retain Service Provider to provide certain [DESCRIPTION OF SERVICES] services upon the terms and conditions hereinafter set forth, and Service Provider is willing to perform such services].

In consideration of the mutual covenants and agreements hereinafter set forth, the parties agree as follows:

1. Definitions.

["**Action**" has the meaning set forth in Section 11.1.]

"**Affiliate**" of a Person means any other Person that directly or indirectly, through one or more intermediaries, controls, is controlled by, or is under common control with, such Person. The term "control" (including the terms "controlled by" and "under common control with") means the possession, directly or indirectly, of the power to direct or cause the direction of the management and policies of a Person, whether through the ownership of voting securities, by contract or otherwise.

["**Authorized Service Recipients**" means the [Affiliates of Customer as may be notified by Customer to Service Provider from time to time/Persons identified as such in [the/a] Statement of Work].]

"**Agreement**" has the meaning set forth in the preamble.

"**Change Order**" has the meaning set forth in Section 5.2.

"**Confidential Information**" means any information that is treated as confidential by a party, including, without limitation, trade secrets, technology, information pertaining to business operations and strategies, and information pertaining to customers, pricing, and marketing. Confidential Information shall not include information that: (a) is already known to the Receiving Party without restriction on use or disclosure prior to receipt of such information from the Disclosing Party; (b) is or becomes generally known by the public other than by breach of this Agreement by, or other wrongful act of, the Receiving Party; (c) is developed by the Receiving Party independently of, and without reference to, any Confidential Information of the Disclosing Party; or (d) is received by the Receiving Party from a third party who is not under any obligation to the Disclosing Party to maintain the confidentiality of such information.

"**Customer**" has the meaning set forth in the preamble.