International Agreements Data Base Mining

Paulina Kulczyk
Warsaw University of Technology
email@domain

Mateusz Wiktorzak
Warsaw University of Technology
email@domain

Muhamed Fahim Asim
Warsaw University of Technology
01205609@pw.edu.pl

Mateusz Zagorski
Warsaw University of Technology
01161597@pw.edu.pl

Supervisor: Anna Wr´oblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

This project addresses the challenge of extracting structured and reliable information from a large collection of international legal agreements signed by U.S. states and cities. The source documents primarily exist as low-quality, inconsistently formatted PDF scans, making them difficult to access and analyze using standard computational methods. These agreements represent a valuable resource for research on subnational diplomacy, yet their usability is limited by OCR noise, metadata inconsistencies, duplicate records, and unreliable entity extraction.

The objective of this work is to transform these documents into a clean, analyzable dataset through a comprehensive information extraction pipeline. We first review state-of-the-art methods relevant to this task, including advanced OCR models for degraded scans, legal-domain named entity recognition, agreement-type and clause classification, text segmentation, zero-shot contract analysis, and semantic deduplication using Legal-SBERT and FAISS. This review identifies the most suitable techniques for handling the complexity and noise characteristic of subnational legal agreements.

We then analyze the dataset itself, examining document quality, variability across agreements, metadata completeness, and linguistic heterogeneity. Basic statistical analysis and qualitative observations highlight challenges related to scan degradation, inconsistent formatting, and multilingual content, motivating informed methodological choices.

In the proof-of-concept stage, the most promising techniques are evaluated on a subset of the corpus to assess practical performance under realistic computational constraints. OCR models, NER systems, clause extractors, and similarity-based deduplication methods are tested to determine which approaches deliver acceptable accuracy. Based on these findings, we design a modular, scalable extraction pipeline that integrates enhanced OCR, legal-domain entity recognition, agreement classification, metadata normalization, clause extraction, and robust duplicate detection. The final system supports reproducible large-scale processing and provides a foundation for further extensions to legal text analysis and international agreement corpora.

1 SOTA Analysis

1.1 Optical Character Recognition (OCR)

Recent work in document understanding shows that OCR alone is insufficient for reliably processing formal government PDFs that mix multi-page scans, complex layouts, stamps, signatures, and legal boilerplate (Smith, 2007; Cui et al., 2021). This project targets converting such heterogeneous, often low-quality in- puts into structured representations with correct page, paragraph, and reading-order reconstruction so that downstream systems can search, validate, and analyze official records at scale, including mea- suring agreement length in pages or words. The task lies at the intersection of OCR, layout-aware information extraction, and document-structure recovery, building on multimodal document AI, form understanding, and invoice-style IE for business and government workflows (Li et al., 2022b; Cui et al., 2021).

Modern OCR backbones are largely open-source and deep-learning based, with engines such as

Tesseract, PaddleOCR, TrOCR, and docTR forming standard building blocks (Smith, 2007; Cui et al., 2025; Li et al., 2022a; Mindee, 2021). Tesseract remains a widely adopted baseline for printed and multilingual text in large digitization efforts, includ- ing legal and governmental archives (Smith, 2007). Newer engines extend these capabilities with convolutional and transformer architectures that jointly model visual and textual context, improving robustness to noise, skew, and varying fonts common in legacy scans and faxed documents and enabling reliable word-level statistics for agreement-length estimation (Cui et al., 2025; Li et al., 2022a; Mindee, 2021). Work on degraded and historical doc- uments shows that binarization, denoising, background removal, and super-resolution further boost OCR quality in the presence of stamps, seals, marginal notes, and low-resolution scans (Ntirogiannis et al., 2019).

Beyond OCR, state-of-the-art layout-aware IE models treat documents as multimodal objects where text, layout, and visual signals are processed jointly (Cui et al., 2021). LayoutLM-style architectures pretrain on large corpora of business and administrative documents with 2D positional embeddings and visual features, then fine-tune for key-value extraction, form understanding, and document QA (Cui et al., 2021; Yashwant et al., 2025). Benchmarks such as FUNSD, CORD, SROIE, and LIE show that layout-aware models consistently outperform text-only baselines on entity extraction and relation prediction, which directly supports extracting areas of cooperation from cleaned agreement text as basic semantic fields (Li et al., 2022b; Cui et al., 2021). LIE is particularly relevant because it includes product and official documents from more than 150 organizations, with diverse templates and page lengths resembling real-world government PDFs (Li et al., 2022b).

At the structural level, work on PDF and document structure extraction focuses on recovering logical hierarchy, section boundaries, and reading order across multi-column and multi-page documents (Luz et al., 2011; Liu et al., 2025). Classical methods such as XY-cut and its variants are strong baselines for segmenting text blocks and columns but are sensitive to noise and layout variation and often require heavy heuristic tuning (Liu et al., 2025). Newer graph-based and neural approaches operate over layout graphs to recover headings, and tables of contents, which in this context support reliable paragraph reconstruction and the identification of recurring clause segments that can be standardized and counted across agreements (Luz et al., 2011; Cui et al., 2021). For this project, these techniques motivate a hybrid design that combines robust block segmentation with learning-based ordering and section classification tuned to government document genres.

Commercial systems such as Google Document AI, ABBYY, and Azure Document Intelligence achieve strong extraction accuracy on forms, invoices, and contracts. However, their closed-source nature, dependence on proprietary clouds, and limited transparency about training data and behavior complicate tuning them (Cui et al., 2021; AI, 2021). Open benchmarks such as CC-OCR and evaluations of layout-aware and multimodal models indicate that open-source and academic approaches can match or surpass these systems in adaptability and extensibility, especially with domain-specific fine-tuning (Yang et al., 2024; Cui et al., 2021). Collec- tively, this literature defines the current SOTA and supports a domain-specialized, open pipeline for formal government PDFs that combines proven components while addressing gaps in domain adapta- tion and full-document reconstruction, including agreement-length measurement, basic extraction of cooperation areas, and clause-frequency analysis over standardized segments.

1.2 Agreement type, sides, organizations extraction

There are three interconnected subtasks that can be grouped together based on the similarity of their outputs:

2. Identify the parties involved,
3. Identify the types of agreements,
5. Identify international organizations mentioned.

These tasks can be addressed using two primary NLP techniques: Named Entity Recognition (NER) for extracting entities such as parties and organizations (tasks 2. and 5.), and Document Classification for categorizing the types of agreements (task 3.). This grouping reflects the underlying workflows needed to transform unstructured legal texts into structured data, performing extraction and finding the key information.

1.2.1 Legal-Domain Named Entity Recognition (NER)

The most advanced model in legal NER is LegNER (Karamitsos et al., 2025), a domain-adapted transformer model pre- trained on extensive legal corpora with span-level supervision. LegNER achieves F1 scores over 99%, surpassing other models like Legal-BERT (Chalkidis et al., 2022) by several percentage points. Its architecture, based on BERT-base but fine-tuned on legal-specific language, enables precise recognition of key legal entities such as parties, laws, and organizations crucial for international agreements. The model maintains high precision and recall, making it reliable for clean entity extraction even on complex, noisy legal data.

Other competitive models include SpanBERT-Legal, which captures entity spans well, and ContractNLI (Henderson et al., 2021), which extends entity recognition to contract clause classification but typically scores lower than LegNER in benchmarks.

1.2.2 Legal agreement type classification

Transformers like RoBERTa, LegalBERT, and T5 are the leading models for legal document clas- sification tasks, including categorizing types of agreements. They are fine-tuned on domain-specific datasets to identify particular categories of contracts. These models leverage multi-label classification and contextual understanding, achieving near-human accuracy in distinguishing diverse legal document types.

1.2.3 Essential datasets

1. LEDGAR (Tuggener and Dániken, 2020): A large-scale, multi-label corpus for contract provision classification created from SEC filings (EDGAR). It contains nearly 100,000 provisions across more than 12,000 labels from over 60,000 contracts. LEDGAR is widely used for fine-tuning models like LegalBERT for clause and agreement type classification, supporting scalable legal NLP research,

2. CUAD (Contract Understanding Atticus Dataset) (Henderson et al., 2024): An expert-annotated dataset with over 2,000 clauses across 41 contract categories from 510 commercial contracts. CUAD enhances classification at the clause-level, useful for detailed contract analysis and extraction,

3. Contract NLI Dataset: A resource for document-level contract classification and clause iden- tification.

These datasets provide rich, annotated data essential for training and benchmarking the state-of-the-art models in legal NER and contract classification.

1.3 Deduplication (Legal-SBERT + FAISS)

Large collections of international agreements often contain duplicate or near-duplicate documents originating from multiple sources, OCR rescans, multilingual versions, or slightly edited reissues. Removing such duplicates is essential for producing reliable statistics, including the percentage of agreements signed under the patronage of Sister Cities International and the identification of partners with whom agreements tend to be more detailed. To address this, the pipeline applies state-of-the-art sentence- embedding–based deduplication, combining Legal-SBERT for semantic encoding and FAISS for fast large-scale similarity search.

1.3.1 Legal-SBERT for semantic encoding

Legal-SBERT (Askari et al., 2024) is a domain-adapted variant of Sentence-BERT fine-tuned on large legal corpora, achieving significant improvements over standard BERT and SBERT in semantic similarity tasks in- volving contracts, statutes, and judicial documents. Reimers & Gurevych's SBERT architecture (Reimers and Gurevych, 2019) enables fixed-size embeddings optimized for cosine-similarity retrieval, while legal-domain adaptations such as Legal-SBERT demonstrate state-of-the-art performance on legal sentence similarity bench- marks, with improvements of up to +7–10 points in Spearman correlation. Using Legal-SBERT ensures that semantically identical agreements—despite OCR noise, reformatting, or paraphrasing—receive nearly identical embeddings, enabling accurate duplicate detection.

1.3.2 FAISS for scalable similarity indexing

FAISS (Facebook AI Similarity Search) (Johnson et al., 2019) is the leading library for high-dimensional vector indexing and approximate nearest-neighbor search. It supports millions of embeddings with sub-second lookup times using optimized GPU/CPU indices such as HNSW and IVF-PQ. FAISS is widely used in legal NLP for clustering and deduplication in large corpora (e.g., EDGAR, global treaty databases). In this project, FAISS enables:

- efficient retrieval of nearest neighbors for each agreement

- identification of duplicate or near-duplicate records (e.g., cosine similarity > 0.92)

- removal of redundant documents prior to downstream analysis

This step is crucial for ensuring that downstream statistics—such as identifying detailed partnerships or calculating Sister Cities International patronage—are not distorted by duplicated agreements or OCR-generated copies.

1.4 Semantic Similarity Clustering (UMAP + HDBSCAN)

Beyond deduplication, semantic similarity clustering enables the discovery of latent patterns in the agreement corpus. This step directly supports tasks such as identifying partners associated with more detailed agreements, detecting coordination with other entities, and clustering agreements that reference other legal documents. These phenomena correspond to distinct semantic regions of the embedding space.

1.4.1 Dimensionality reduction with UMAP

Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) is one of the most advanced non- linear dimensionality-reduction algorithms for high-dimensional embeddings. Unlike PCA or t-SNE, UMAP preserves both local neighborhoods and global manifold structure, producing representations well suited for density-based clustering. Applied to Legal-SBERT embeddings, UMAP:

- reduces 768-dimensional vectors to 10–50 dimensions
- preserves semantic topology of legal texts
- enhances density separation between agreement types, detailed vs. minimal agreements, refer- ences to external legal documents, and coordination clauses

UMAP has become a standard preprocessing step in legal and scientific document clustering because it maintains structural coherence while enabling more robust cluster detection.

1.5 Density-based clustering with HDBSCAN

HDBSCAN (Campello et al., 2013) is the state-of-the-art clustering algorithm for noisy text datasets, outperforming k- means and DBSCAN in legal document grouping. It automatically identifies dense regions in the reduced embedding space and labels sparse regions as “noise,” which is ideal for heterogeneous legal corpora.

For this project, HDBSCAN enables:

- grouping agreements with similar structural complexity (to identify highly detailed partners)
- detecting clusters of agreements that reference other legal documents
- isolating agreements that mention coordination with external entities (e.g., federal agencies, NGOs, international organizations)
- discovering thematic clusters without requiring predefined labels

In combination, UMAP + HDBSCAN form a robust state-of-the-art pipeline for unsupervised legal document analysis, enabling empirical insight into agreement structure and recurring patterns relevant for social science research.

1.6 Determine Terms of Validity

Extracting and normalizing temporal expressions in contracts is essential to identify start and end dates, durations, and validity periods. Temporal information may be explicit or implicit, expressed via phrases such as ”for the term of the project” or ”until December 31, 2026”. Accurate extraction supports contract interpretation, compliance, and automated reasoning.

Rule-based systems such as HeidelTime remain a strong baseline, achieving high precision on legal corpora such as CUAD and LEDGAR by encoding patterns for dates, durations, and relative expressions (Strötgen and Gertz, 2010).

Neural approaches, particularly LegalBERT fine-tuned for token classification, capture implicit temporal phrasing and domain-specific variations (Singh et al., 2025). ContractNLI verification further refines extrac- tion, utilizing hybrid pipelines that combine rule-based normalization, neural token classification, and NLI verification, with F1 scores consistently above 0.85 (Koreeda and Manning, 2021).

1.7 Determine Conditions for Extending the Agreement

Renewal and extension clauses define when a contract may be automatically renewed, mutually extended, or optionally extended. They often include nested conditions, notice periods, or triggers, making detection challenging.

Transformer-based models such as LegalBERT and Longformer, fine-tuned on CUAD and LEDGAR, detect renewal clauses and distinguish automatic from optional extensions (Singh et al., 2025).

Hybrid pipelines combining neural span extraction, rule-based temporal/condition normalization, and NLI or QA verification improve precision and resolve cross-references (Koreeda and Manning, 2021), with F1 scores typically between 0.80 and 0.85. Simple rule-based or keyword approaches can supplement neural methods in less complex contracts.

1.8 Indicate Whether the Agreement Includes an Evaluation of Its Imple- mentation

Evaluation clauses monitor compliance or performance, specifying assessment, reporting, or auditing duties. They are often sparse or implicit, with verbs like ”review”, ”assess”, or ”audit”, sometimes spanning multiple sections.

Fine-tuned LegalBERT or LegalProBERTmodels detect evaluation clauses in CUAD and LEDGAR, while token-classification extracts roles and reporting frequency (Singh et al., 2025). NLI or QA verification captures implicit evaluation obligations, improving precision.

Hybrid pipelines, combining token-classification, rule-based heuristics, and NLI/QA verification, achieved F1 scores around 0.80–0.83, balancing recall and precision (Koreeda and Manning, 2021).

1.9 Metadata Reliability

Metadata for international agreements—such as dates, partner names, issuing authorities, agreement types, and references to Sister Cities International—are often inconsistent or incomplete due to variable formatting, or manual transcription. Ensuring metadata reliability is fundamental for tasks such as computing the percentage of agreements under Sister Cities International and systematically identifying coordination clauses, external references, and detailed agreements with specific partners.

1.9.1 Metadata normalization

Normalization consolidates metadata fields into standardized canonical forms. This includes:

- organization name normalization (e.g., “Sister Cities International,” “Sister City Int’l,” “SCI”) using fuzzy matching and embedding-based similarity partner
- entity normalization using Wikidata cross-referencing and legal-NER outputs
- date normalization using ISO-8601 standards
- agreement type harmonization based on controlled vocabularies from international treaty databases

State-of-the-art metadata normalization relies on neural fuzzy-matching models such as Ditto (Li et al., 2020) and embedding-based entity alignment (Sun et al., 2020), which significantly outperform rule-based approaches by capturing semantic-level similarity across heterogeneous entity representations. However, transformer- based models such as Ditto require substantial computational resources—particularly GPU acceleration and fine-tuning on domain-specific datasets—making them less practical for smaller-scale projects or environments with limited hardware. Consequently, for this study, we adopt a fuzzy string matching approach using RapidFuzz (Ye et al., 2021). This method provides sufficient accuracy for canonicalizing entities such as variations of ‘Sister Cities International’ while remaining computationally lightweight and feasible within the available CPU resources and project timeline.

1.9.2 Consistency checks

After normalization, consistency checks detect contradictions or missing information. Typical checks include:

- verifying that agreement dates are chronological
- confirming that referenced legal documents exist in the database
- ensuring that coordination entities (ministries, city councils, NGOs) match recognized NER categories
- checking cross-document consistency for Sister Cities International attribution
- validating that long agreements (as identified via clustering or word counts) correspond to ex- pected partner categories

Such checks rely on techniques from information quality management (Batini et al., 2016) and cross-document vali- dation frequently applied in treaty and contract corpora. In addition to these traditional approaches, modern LLMs can assist in consistency verification by detecting anomalous or contradictory metadata patterns—such as misaligned partner names, unexpected absence of coordination clauses, or references that do not match the text—providing a hybrid methodology that ensures metadata is reliable, coher- ent, and robust across the entire corpus.

Together, metadata normalization and consistency checks ensure that extracted insights—such as de- tailed agreement patterns or coordination with other entities—are based on validated and coherent metadata, increasing reliability for social science research.

2 Dataset

2.1 Data collection and preprocessing

The raw data were initially acquired in two formats: PDF and HTML, with a distribution ratio of approximately 2:1. During the inspection phase, several critical preprocessing steps were taken:

- Format Filtering: All HTML files were removed from the dataset. Inspections revealed that these files were low-quality extractions from PDFs, containing fragmented sentences and incon- sistent formatting that would impede reliable mining.
- OCR Processing: Optical Character Recognition (OCR) was applied to the remaining docu- ments to generate two output types:
  1. .txt files containing the raw agreement text.
  2. .json files containing structured statistical summaries of the text.
- Storage Structure: The processed files are organized into 50 distinct directories, each named after a US state, stored within a central OCR-output directory.

2.2 Dataset composition

The current dataset comprises a total of 298 agreements. While the structure accounts for all 50 US states, the actual distribution of documents is as follows:

Metric	Value
Minimum agreements per state	0
Maximum agreements per state	116
Average agreements per state	5.96
Median agreements per state	1.0
Standard Deviation ()	18.05

Table 1: Dataset Summary Statistics

Rank	State	Count
1	California	116
2	Texas	54
3	Utah	22
4	Arizona	17
5	Alaska	12

Table 2: Top 5 States by Document Count

- States with Agreements: 29 (58% coverage).
- States without Agreements: 21 (42% coverage).
- Total Sample Size: 298 documents.

2.3 Statistical Analysis

The distribution of agreements per state is highly skewed, characterized by a few significant outliers and many states with minimal data.

2.4 Geographical outliers

A small number of states contribute the vast majority of the dataset. The top 5 states by agreement count are shown in Table 2. Notably, California alone accounts for nearly 39% of the total dataset. Conversely, many states, such as Missouri, Florida, Georgia, and New Mexico, currently contain zero recorded agreements in this specific project iteration.

2.5 Conclusion

The dataset is robust in its total document count (298) but exhibits significant geographical im- balance. Preprocessing has successfully filtered out noisy HTML data, ensuring that the remaining OCR-processed text is of high quality for subsequent legal mining and metadata normalization tasks.

3 Exploratory data analysis

This section presents the key findings of EDA conducted on the International Agreements Database.

The dataset consists of OCR-processed agreement documents organized by U.S. state. The analysis focuses on the spatial distribution, document characteristics, and thematic content of these agreements.

3.1 Dataset overview and spatial distribution

The dataset comprises 298 agreements distributed across 50 directories, each corresponding to a U.S. state. The distribution is highly skewed, with only 29 states containing any agreement documents.

- Total Volume: There are 298 agreements in total.

- Geographic Concentration: The vast majority of agreements are concentrated in California (116 agreements) and Texas (54 agreements). Utah also holds a significant number (22), while many other states (e.g., Virginia, Illinois, New Hampshire) contain only a single document.

- Coverage: 21 out of 50 states have no representation in the dataset.

3.2 Document characteristics

An analysis of document length (page count and word count) reveals that most agreements are relatively concise.

- Page Count: The agreements range from 1 to 21 pages in length. The distribution is heavily weighted towards shorter documents, with the majority having 5 pages or fewer.

- Word Count: The total word count per document ranges from 94 to 7,368 words.

- Variation: States with the highest volume of agreements (California and Texas) exhibit the widest variance in document length. Conversely, outliers such as Alaska and Vermont also contain documents reaching the maximum length of 21 pages.

3.3 Thematic analysis and N-gram extraction

To identify key themes, frequentist analysis was performed on word tokens and bigrams after removing standard English stopwords and state names.

3.3.1 Top Words

The most frequent terms across the corpus highlight the administrative and cooperative nature of the texts. The top words include ”state” (921 occurrences), ”parties” (423), ”information” (422), and ”energy” (380). Terms like ”participants”, ”supervisors”, and ”understanding” are also prominent.

3.3.2 Bigram analysis

Bigram analysis reveals specific areas of international cooperation. After filtering out generic phrases (e.g., ”United States”), the following key themes emerged:

- Legal Frameworks: The most common bigram is ”memorandum understanding” (904 occur- rences), indicating the primary legal instrument used.

- Environmental Issues: A significant portion of the corpus focuses on climate and ecology, evidenced by bigrams such as ”climate change” (326), ”low carbon” (233), ”forest fire” (180), ”greenhouse gas” (156), and ”clean energy” (144).

- Regional Partners: Frequent references to ”British Columbia” (200) and ”Great Lakes” (138) suggest strong cross-border cooperation with Canada.

3.3.3 State-Specific Topics

State-level analysis highlights distinct regional priorities:

- New Jersey: Unique occurrences of ”salud chihuahua” and ”servicios salud” indicate specific health-related cooperation with the Mexican state of Chihuahua.

- Northern Border States: Maine, Michigan, and Wisconsin frequently feature terms like ”for- est fire”, ”crossing agreement”, and ”great lakes”, reflecting shared environmental and infras- tructure interests with Canada.

- Western States: California and Oregon are heavily associated with terms regarding ”climate change” and ”clean energy”.

4 Proof of concept

4.1 Tasks 1, 7, 9

The Proof of Concept (POC) regarding the OCR pipeline establishes the technical feasibility of trans- forming a large corpus of scanned legal agreements into a structured dataset optimized for compu- tational social science. The proposed system is anchored in a deep learning-based Optical Character Recognition (OCR) framework designed to digitize documents with high geometric fidelity, subse- quently enabling complex algorithmic analysis to determine critical metrics including document length, the prevalence of boilerplate language, and the substantive scope of international cooperation.

4.1.1 Determining documents length

The core extraction layer is built upon the docTR (Document Text Recognition) library, utilizing a pre- trained predictor to process PDF documents ingested from the project's directory structure. To ensure the integrity of the analysis, the workflow incorporates a preprocessing step that programmatically discards the first page of every file, effectively stripping away the administrative cover sheets and metadata often appended during the download process. Unlike simple text-dumping tools, the model generates a hierarchical JSON output for each agreement, mapping the document's geometry into a nested structure of pages, blocks, lines, and words. This geometric preservation is crucial for the subsequent metadata extraction task, where a custom algorithm traverses the JSON hierarchy to calculate document volume. By iterating through every text block and summing the token counts at the line level, the system produces a highly accurate word count that ignores whitespace anomalies, resulting in the dataset which facilitates precise quantitative comparisons of agreements (Task 7).

4.1.2 Identifying areas of cooperation

Building upon this digitized foundation, the module implements a sophisticated NLP architecture to analyze the legal text. For the analysis of recurring clauses (Task 9), the system distinguishes between standardized "boilerplate" templates and bespoke diplomatic terms by leveraging the visual block structure detected by the OCR engine. To filter out noise such as page numbers, headers, and artifacts, the system discards any text blocks containing fewer than ten words. The remaining substantive clauses are encoded into dense vector embeddings using the all-MiniLM-L6-v2 Sentence-Transformer, a model optimized for semantic similarity tasks. These embeddings are then subjected to HDBSCAN clustering with a Euclidean metric and a minimum cluster size of two, allowing the system to detect even minimal repetitions of legal phrasing. The classification logic is derived from these cluster assignments: clauses appearing in over 80% of the document corpus are categorized as standard diplomatic protocol, whereas clauses that fail to form clusters (labeled as -1) are identified as unique, custom-drafted terms specific to a particular negotiation.

4.1.3 Analyzing frequency of recurring clauses

In parallel with the structural analysis, the POC addresses the identification of cooperation areas (Task 1) through a Zero-Shot Classification approach powered by the facebook/bart-large-mnli model. This methodology allows the system to categorize agreements into specific sectors—such as "Green Energy," "Culture & Arts," or "Trade & Economic Development"—without the prohibitive requirement of a manually labeled training dataset. Recognizing the input constraints of Transformer-based models, the pipeline reconstructs the full text from the JSON output and applies a truncation strategy, limiting the input to the first 3,000 characters. This window is strategically selected to encompass the Preamble and the initial "Scope of Cooperation" articles, where the binding intent of the agreement is typically articulated. The model performs a multi-label classification on this truncated text, assigning tags only when the confidence score exceeds a 0.5 threshold. The resulting data is aggregated into a final CSV report, confirming that the infrastructure is fully operational and ready for the large-scale ingestion of the complete legal dataset.

4.2 Tasks 2, 3, 5

This section details the POC implementation for extracting agreements parties, agreement types, and identifying international organizations mantioned from the legal agreements corpus. The POC has been executed on the Alabama subdirectory of the parsed International Agreements Database.

4.2.1 Task 2: Extraction of Agreement Parties

Two distinct approaches were evaluated for the extraction of contracting parties from the legal texts: a Named Entity Recognition (NER) approach and a Large Language Model (LLM) prompting approach.

Method 1: Legal NER The initial attempt utilized a BERT-based model fine-tuned for legal Named Entity Recognition. The extraction logic involved tokenizing the input text, running inference to obtain labels, and aggregating tokens into entities based on the tagging scheme.

The NER approach proved insufficient for this specific use case. The primary issues observed included:

- Complex Naming Structures: Agreement parties often possess long, multi-word names containing embedded entities (e.g., ”Ministry of Economics, Small and Medium Business, and Tech- nology”). The NER model frequently fragmented these into separate entities rather than recog- nizing the single legal party.

- Model Availability: Finding robust, publicly available NER models fine-tuned specifically for this type of legal document construction was difficult, with several candidate sources being broken or deprecated.

Method 2: Generative extraction (Mistral 7B Instruct) Due to the limitations of the NER approach, the strategy shifted to using a quantized version of the Mistral 7B Instruct model. The model was prompted to act as a legal AI and return a JSON object containing the required information. The instruct-oriented fine-tuning of the Instruct varient of the Mistral model was supposed to ensure specifity of the answer. It could be interesting to compare if a model fine-tuned for legal purposes would do better in extraction than a model trained to consider exact instructions.

The LLM approach yielded significantly better results. The model successfully identified complex party names and provided context regarding their roles. However, minor errors persisted in complex scenarios, such as occasionally misclassifying constituent members of a party as separate agreement parties.

4.2.2 Task 3: Agreement Type Classification

Two methods were implemented to categorize the agreements into types such as Memorandum of Understanding, Trade Agreement, or Partnership Agreement.

Method 1: Zero-Shot Classification (BART) A BART-large model trained on the MNLI dataset was employed for zero-shot classification. The model classified texts against a predefined list of 13 agreement types.

It was difficult to assess the outcome validity without an expert knowledge, but some of the agreements like Memoranda of Understanding had this value written in the very first line, and for those, the classification worked. Some of the classified agreements had their confidence ratios for each type very close to each other, indicating that the predefined subset could not be accurate.

Method 2: Generative Extraction (Mistral 7B Instruct) To allow for more flexibility, the Mistral 7B model was prompted to analyze the text and generate the agreement type and a brief description without being restricted to a fixed list. This approach allowed for the identification of nuanced agreement types that might not fit strictly into preset categories.

Some of the agreements had their extracted type matching for both methods. Generative extraction yields better responses in the case of custom agreements – the exact type of the agreement may not be stated explicitly.

4.2.3 Task 5: Extraction of International Organizations

Building on the findings from Task 2, the extraction of international organizations was performed using the Mistral 7B Instruct model rather than NER.

Methodology The model was instructed to extract organizations operating across national borders (e.g., intergovernmental bodies) while explicitly ignoring purely domestic agencies unless they were part of an international body. The output was structured as a JSON object containing the organization’s name and a short context.

Results This method successfully filtered out local entities to focus on international actors, providing a cleaner list of relevant organizations involved in the agreements.

4.3 Tasks 6: Determine Terms of Validity

The following methods are implemented to determine the terms of validity

4.3.1 HeidelTime

This method applies the HeidelTime temporal tagger to identify `TIMEX3 DATE` and `DURATION` expressions in agreement text. The resulting TimeML output is parsed and temporally normalized into ISO-formatted dates. Validity-specific anchor heuristics are applied at the sentence level to associate temporal expressions with agreement validity semantics (e.g., *effective*, *enter into force* for start dates; *until*, *expires*, *expiration* for end dates). When explicit anchors are absent, fallback logic assigns the earliest detected date as the effective date and the latest as the end date. Although HeidelTime was evaluated both with and without TreeTagger-based POS tagging, the results reported here correspond to the POS-free configuration, which proved more robust on OCR-derived text and showed no consistent improvement from POS tagging.

4.3.2 POC Main Validity Output (Hybrid Validity Extractor)

The POC main method implements a hybrid heuristic extractor for agreement validity terms. Candidate sentences are retrieved using validity-related keyword triggers, and temporal expressions are extracted using rule-based date and duration patterns. Heuristic selection logic determines the most plausible effective date, end date, and duration based on local context and anchor phrases. The method outputs structured validity fields together with supporting evidence snippets and assigns a document-level validity status (*found*, *uncertain*, or *absent*). This approach balances recall and precision but exhibits lower coverage than HeidelTime for end-date extraction.

4.3.3 POC Baseline Keyword Validity (Presence-Only Flag)

This baseline performs presence-only detection of validity clauses. The document text is scanned for predefined validity-related keywords using regular expressions. If any keyword is detected, the document is marked as containing a validity clause. This method does not extract or normalize temporal information and serves as a minimal reference baseline for evaluating the added value of structured temporal extraction.

4.3.4 POC Baseline Rule-Based Validity (Simple Patterns)

This approach relies exclusively on hand-crafted regular expressions to extract validity-related temporal information. Patterns are used to identify effective dates, end dates, and durations directly from the text, and simple positional heuristics are applied to distinguish start and end dates. Unlike the hybrid method, this baseline does not apply semantic filtering or evidence ranking, making it computationally lightweight but sensitive to phrasing variation and OCR noise.

The results obtained by above discussed techniques are as following:

Technique	**N**	**Found**	**Start**	**End**
HeidelTime	30	100.0	100.0	83.3
POC main (hybrid)	30	76.7	46.7	16.7
POC keyword (presence)	30	80.0	0.0	0.0
POC rules (patterns)	30	80.0	53.3	10.0

Table 3: Term validity extraction coverage (%) on 30 California agreements.

Technique	**N**	**Dur.**	**End (explicit)**
HeidelTime	30	83.3	13.3
POC main (hybrid)	30	70.0	16.7
POC keyword (presence)	30	0.0	0.0
POC rules (patterns)	30	70.0	10.0

Table 4: Duration extraction and explicitly stated end dates (%) on 30 California agreements.

The results show that HeidelTime combined with validity-specific anchor heuristics provides the most reliable extraction of structured validity information, achieving the highest coverage for effective dates, end dates, and durations. Simpler baseline approaches are sufficient for detecting the presence of validity clauses but fail to consistently recover explicit temporal boundaries, highlighting the importance of structured temporal normalization.

4.3.5 Determine Conditions for Extending the Agreement

The proof-of-concept for identifying extension and renewal conditions follows SOTA methods for contract clause detection, commonly evaluated on CUAD and LEDGAR. Candidate renewal clauses are first retrieved using high-recall lexical and syntactic rules targeting renewal triggers, notice periods, and extension conditions in noisy OCR text. Semantic validation is then performed using a zero-shot NLI classifier, serving as a proxy for transformerbased models such as LegalBERT or Longformer used in prior work. Rule-based post-processing distinguishes automatic renewals from extensions requiring mutual agreement, reflecting the clausetype differentiation reported in the literature. This hybrid de- sign demonstrates how SOTA renewal detection techniques can be approximated without task-specific fine-tuning.

4.3.6 Indicate Whether the Agreement Includes an Evaluation of Its Implementation

The proof-of-concept for detecting evaluation clauses is informed by SOTA legal clause classification approaches developed on datasets such as CUAD and LEDGAR. High-recall lexical rules are first applied to identify candidate segments containing evaluation-related verbs and reporting terminology, accommodating OCR-induced variability. To approximate the semantic discrimination achieved by fine-tuned LegalBERT or LegalPro-BERT models, a zero-shot NLI classifier is used to verify whether candidate segments entail an obligation to evaluate or assess implementation. This verification step mirrors the function of ContractNLI-style reasoning in hybrid pipelines. The resulting binary indicator and evidence extraction operationalize SOTA evaluationclause detection in a reproducible proof-of- concept.

4.4 Tasks 4, 10, 12, 13

4.4.1 Introduction

The objective of this Proof of Concept (POC) is to develop a system for the automated analysis of legal agreements. The project focuses on leveraging State-of-the-Art (SOTA) Natural Language Processing (NLP) techniques to handle deduplication, semantic clustering, and metadata normalization for a dataset of approximately 30 legal documents.

4.4.2 Methodology and SOTA Tasks

Semantic Embeddings To capture the nuances of legal language, we utilized Legal-SBERT (based on paraphrase-multilingual-MiniLM-L12-v2). This model transforms raw text into high-dimensional dense vectors, preserving semantic relationships between documents.

Deduplication (SOTA) We implemented a deduplication pipeline using:

- Cosine Similarity Matrix: To compute pairwise similarity between all documents.
- FAISS (Facebook AI Similarity Search): For efficient indexing and retrieval of near- duplicate documents.

A similarity threshold of 0.85 was applied to identify redundant agreements.

Semantic Similarity Clustering For document discovery and categorization, we combined:

- UMAP (Uniform Manifold Approximation and Projection): To reduce dimensionality while preserving global structure.
- HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise): To automatically detect clusters of agreements based on their content without pre- defining the number of clusters.

Metadata Normalization & Consistency Check We addressed the inconsistency in extracted data (e.g., ”NYC” vs ”New York City”) using:

- RapidFuzz: Fuzzy matching to merge variations of the same entity.
- LLM-Assisted Normalization: Standardizing noisy fields (dates, organization names) into a uniform format (e.g., YYYY-MM-DD).
- Consistency Checks: Rule-based verification to ensure the extracted metadata is reliable and present in the source text.

4.4.3 POC Results and Project Tasks

Task ID	Description and Result
(4)	Sister Cities International %: Automated regex and semantic search identified the percentage of documents falling under this category.
(10)	Detailed Agreements: Identification of partners with more detailed agreements based on average word count and clause density.
(12)	Coordination Detection: Implementation of flags to identify whether agreements mention coordination with other entities.
(13)	Legal References: Automated detection of references to other legal documents and statutes.

Table 5: Fulfillment of Specific Project Tasks

References

Microsoft Azure AI. 2021. Contract data extraction with document intelligence. https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/prebuilt/contract. Online documentation describing prebuilt contract models for Document Intelligence.

Arian Askari, Suzan Verberne, Amin Abolghasemi, Wessel Kraaij, and Gabriella Pasi. 2024. Retrieval for extremely long queries and documents with rprs: a highly efficient and effective transformer-based reranker. *ACM Transactions on Information Systems*, 42(5):1–32.

Carlo Batini, Monica Scannapieco, and 1 others. 2016. Data and information quality. *Cham, Switzerland: Springer International Publishing*, 63.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Ilias Chalkidis, Michalis Fergadiotis, and I Androutsopoulos. 2022. Legalbert: The muppets straight out of law school. *arXiv preprint arXiv:2202.02503*. Accessed: 2025-11-25.

Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. PaddleOCR 3.0 technical report. *Preprint*, arXiv:2507.05595.

Lei Cui and 1 others. 2021. Document AI: Benchmarks, models and applications. Microsoft Research Technical Report.

Paul Henderson and 1 others. 2024. An expert-annotated dataset for legal contract clause retrieval. *ArXiv preprint arXiv:2408.06582*. Accessed: 2025-11-25.

Peter Henderson and 1 others. 2021. Contractnli: A dataset for document-level natural language inference on contracts. In *Proceedings of ACL 2021*. Accessed: 2025-11-25.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Ilias Karamitsos and 1 others. 2025. Legner: a domain-adapted transformer for legal named entity recognition. *Frontiers in Artificial Intelligence*, 6:1638971. Accessed: 2025-11-25.

Yuta Koreeda and Christopher D Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts. *arXiv preprint arXiv:2110.01799*.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2022a. TrOCR: Transformer-based optical character recognition with pre-trained models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11124–11132.

Yiheng Li, Yujia Xu, Lei Cui, Yutong Zhang, and 1 others. 2022b. Layout-aware information extraction for document-grounded dialogue. In *Proceedings of the ACM International Conference on Multimedia*.

Yuliang Li, Jinfeng Li, Yoshihiko Suhara, AnHai Doan, and Wang-Chiew Tan. 2020. Deep entity matching with pre-trained language models. *arXiv preprint arXiv:2004.00584*.

Shuai Liu, Youmeng Li, and Jizeng Wei. 2025. XY-Cut++: Advanced layout ordering via hierarchical mask mechanism on a novel benchmark. *arXiv preprint arXiv:2504.10258*.

Saturnino Luz and 1 others. 2011. Structure extraction from PDF-based book documents. In *Proceedings of the ACM Symposium on Document Engineering*. ACM.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Mindee. 2021. docTR: Document text recognition. `https://github.com/mindee/doctr`. GitHub repository for docTR.

Konstantinos Ntirogiannis, Basilis Gatos, and Ioannis Pratikakis. 2019. Degraded historical document binarization: A review on issues and methods. *Applied Sciences*, 9(16):1–30. Representative survey on degraded historical document binarization and preprocessing.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Amrita Singh, Aditya Joshi, Jiaojiao Jiang, and Hye young Paik. 2025. A survey of classification tasks and approaches for legal contracts. *arXiv preprint arXiv:2507.21108*. Version v1, July 2025.

Ray Smith. 2007. An overview of the tesseract OCR engine. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR)*, pages 629–633. IEEE.

Jannik Strötgen and Michael Gertz. 2010. Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 321–324.

Zequn Sun, Qingheng Zhang, Wei Hu, Chengming Wang, Muhao Chen, Farahnaz Akrami, and Chengkai Li. 2020. A benchmarking study of embedding-based entity alignment for knowledge graphs. *arXiv preprint arXiv:2003.07743*.

Daniel Tuggener and Fabian Dániken. 2020. Ledgar: A large-scale multi-label corpus for text classification of legal contracts. In *Proceedings of LREC 2020*. Accessed: 2025-11-25.

Zhibo Yang, Jun Tang, Zhaohai Li, Pengfei Wang, Jianqiang Wan, Humen Zhong, Xuejing Liu, Mingkun Yang, Peng Wang, Shuai Bai, LianWen Jin, and Junyang Lin. 2024. CC-OCR: A comprehensive and challenging OCR benchmark for evaluating large multimodal models in literacy. *Preprint*, arXiv:2412.02210.

Sai Yashwant, Anurag Dubey, Praneeth Paikray, and Thulsiram Gantala. 2025. Invoice information extraction: Methods and performance evaluation. *arXiv preprint arXiv:2510.15727*.

Aoshuang Ye, Lina Wang, Lei Zhao, Jianpeng Ke, Wenqi Wang, and Qinliang Liu. 2021. Rapidfuzz: Accelerating fuzzing via generative adversarial networks. *Neurocomputing*, 460:195–204.