International Agreements Data Base Mining

Paulina Kulczyk
Warsaw University of Technology
01151498@domain

Mateusz Wiktorzak
Warsaw University of Technology
01171275@pw.edu.pl

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

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

Supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

Research on subnational diplomacy is often hindered by the inaccessibility of international legal agreements signed by U.S. states and cities, which predominantly exist as low-quality, unstructured PDF scans. This project addresses the challenge of converting these degraded documents into a reliable, structured dataset through a comprehensive information extraction pipeline. We target thirteen specific data points, ranging from the identification of parties and areas of cooperation to the analysis of recurring clauses and validity terms. The methodology integrates a review of state-of-the-art techniques—including advanced Optical Character Recognition (OCR), legal-domain Named Entity Recognition (NER), and semantic deduplication using Legal-SBERT—with a rigorous dataset analysis to address noise and linguistic heterogeneity. Following a proof-of-concept evaluation of these methods on a corpus subset, we present a modular, scalable pipeline designed to handle OCR noise, classify agreement types, and extract metadata with high accuracy. The resulting system enables reproducible large-scale processing, providing a foundational tool for quantitative research into the legal frameworks of subnational international relations.

1 Introduction

[Need to slightly rephrase it to sound more "results-oriented" once our experiments are finalized.]

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 scope of tasks, which were predefined in the beginning of the work and will be referred to later, are:

1. Identification of areas of cooperation mentioned in the agreements.
2. Identification of the parties involved.
3. Identification of the types of agreements.
4. Determination of the percentage of agreements under the patronage of Sister Cities International.
5. Identification of international organizations mentioned in the agreements.
6. Determination of the terms of validity for each agreement.
7. Identification of the length of each agreement.
8. Determination of the conditions for extending each agreement.
9. Analysis of the frequency of recurring clauses in the agreements.
10. Identification of the partners with whom the agreements tend to be more detailed.

11. Indication of whether the agreement includes an evaluation of its implementation.

12. Identification of whether the agreement mentions any coordination of activities with other entities.

13. Identification of whether the agreement refers to other legal documents.

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.

[final results stage]

2 SOTA Analysis

2.1 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 inputs 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 measuring 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 (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). Tesseract remains a widely adopted baseline for printed and multilingual text in large digitization efforts, including 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 (Mindee, 2021). Work on degraded and historical documents 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). 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 (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.

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). 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. 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 (Cui et al., 2021). Collectively, 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 adaptation and full-document reconstruction, including agreement-length measurement, basic extraction of cooperation areas, and clause-frequency analysis over standardized segments.

2.2 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.

2.2.1 Legal-Domain Named Entity Recognition (NER)

The most advanced model in legal NER is LegNER (Karamitsos et al., 2025), a domain-adapted transformer model pretrained on extensive legal corpora with span-level supervision. LegNER achieves F1 scores over 99%, surpassing other models like Legal-BERT 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.

2.2.2 Legal agreement type classification

Transformers like RoBERTa, LegalBERT, and T5 are the leading models for legal document classification 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.

2.2.3 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): 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 identification.

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

2.3 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.

2.3.1 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 involving contracts, statutes, and judicial documents. Reimers & Gurevych's SBERT architecture 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 benchmarks, 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.

2.3.2 FAISS for scalable similarity indexing

FAISS (Facebook AI Similarity Search) 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.

2.4 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.

2.4.1 Dimensionality reduction with UMAP

Uniform Manifold Approximation and Projection (UMAP) is one of the most advanced nonlinear 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.

2.5 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.

2.6 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. ContractNLI verification further refines extraction, utilizing hybrid pipelines that combine rule-based normalization, neural token classification, and NLI verification, with F1 scores consistently above 0.85.

2.7 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.

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

2.8 Indicate Whether the Agreement Includes an Evaluation of Its Implementation

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 LegalPro-BERT models detect evaluation clauses in CUAD and LEDGAR, while token-classification extracts roles and reporting frequency. 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.

2.9 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.

2.9.1 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 and embedding-based entity alignment, 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.

2.9.2 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 validation 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, coherent, and robust across the entire corpus.

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

3 Dataset

3.1 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 inconsistent 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.

The original data shared with us by the Uniwersytet Łódzki team is stored at this address: `https:`

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 (σ)	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

//uniwersytetlodzki-my.sharepoint.com/personal/marcin_frenkel_wsmip_uni_lodz_pl/_layouts/15/onedrive.aspx?id=%2Fpersonal%2Fmarcin%5Ffrenkel%5Fwsmip%5Funi%5Flodz%5Fpl%2FDocuments%2FU%C5%81%2FProjekt%20z%20Politechnik%C4%85%20Warszawsk%C4%85%20%2D%20baza%20um%C3%B3w%2FBaza%20um%C3%B3w&ga=1.

All preprocessed data can be found at the following address: https://drive.google.com/drive/folders/12lgLmvTZY6Q3Rz2L9JxGmPggIAXDLJ7M?fbclid=IwY2xjawPIwK5leHRuA2FlbQIxMABicmlkETE4WTE3dzRGVFlRUmY1WDVPc3J0YwZhcHBfaWQQMjIyMDM5MTc4ODIwMDg5MgABHikF555zPJ0CW8CP25tOVf89vdwAn_tyd7s-nsh-fzPK306j5YHTIKBhPaGO_aem_REAubR4osJ8cwvnihpEqpA.

3.2 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:

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

3.3 Statistical Analysis

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

3.4 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, and top 10 in the Figure 1. 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.

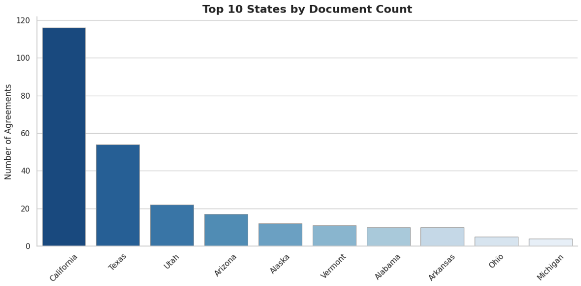


Figure 1: Barplot with count of agreements for top 10 states

3.5 Conclusion

The dataset is robust in its total document count (298) but exhibits significant geographical imbalance. 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.

4 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.

4.1 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 (Figure 2).

- **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 (Figure 2).

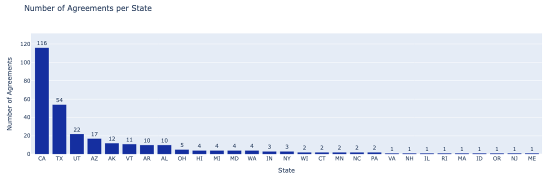


Figure 2: Barplot with count of agreements across all states

- **Coverage**: 21 out of 50 states have no representation in the dataset (Figure 3).

- **Outliers**: While California and Texas lead in absolute numbers, it is worth noting that Alaska and Vermont—which are among the least populous states—have a relatively large number of contracts (12 and 11, respectively). We can therefore classify them as outliers in terms of population versus number of contracts. These states appear as notable hubs for international activity relative to their population size, likely due to their geographic proximity to Canada.

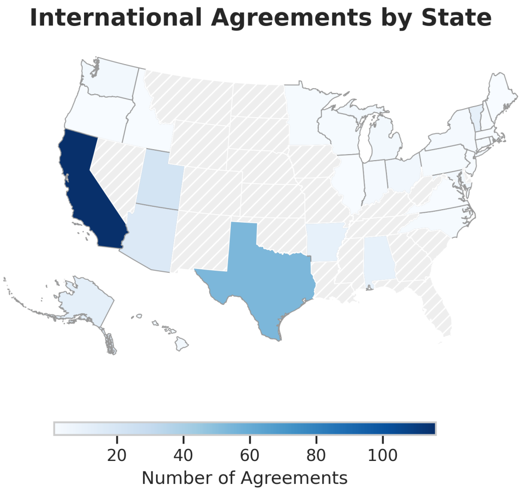


Figure 3: Map of distribution of documents through states

4.2 Document characteristics

An analysis of document length (page count and word count) reveals that most agreements are relatively concise, though significant complexity exists in specific regions (Figure 4).

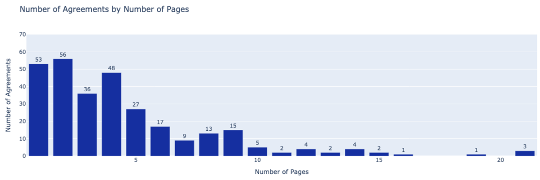


Figure 4: Distribution of document length (page count) by state

- **Page Count**: The agreements range from 1 to 21 pages. The distribution is heavily weighted towards shorter documents, with the majority having 5 pages or fewer (typically standard MOUs).

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

- Variation: While California and Texas have the highest volume, they also show the widest variance in document length. Interestingly, Michigan and Pennsylvania exhibit higher median page counts, suggesting their agreements may involve more complex regulatory or infrastructure frameworks compared to the single-page declarations found in other states. Maximum length documents (21 pages) were identified in California, Vermont, and Alaska.

4.3 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.

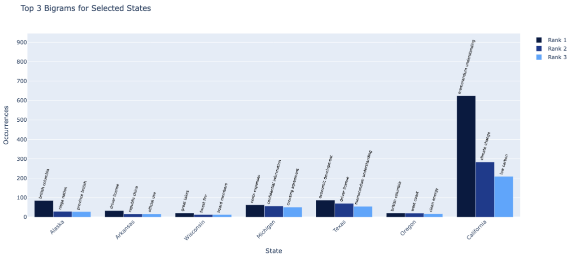


Figure 5: Key thematic bigrams for selected states.

4.3.1 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.

4.3.2 Bigram analysis

Bigram analysis reveals specific areas of international cooperation (Figure 5). After filtering out generic phrases (e.g., ”United States”), the following key themes emerged:

- **Legal Frameworks**: The most common bigram is ”memorandum understanding” (904 occurrences), confirming that MOUs are 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).
- **Emergency Management**: The prevalence of ”forest fire” (180) indicates shared cross-border resources for disaster response.
- **Regional Partners**: Frequent references to ”British Columbia” (200), ”Great Lakes” (138), and ”Republic [of] China” (124) highlight the primary geopolitical partners: Canada and East Asia.

Table 3: Top 10 words and bigrams.

Word	Count	Bigram	Count
state	921	memorandum understanding	904
parties	423	climate change	326
information	422	economic development	240
energy	380	low carbon	233
la	262	confidential information	214
understanding	241	British Columbia	200
supervisors	233	forest fire	180
participants	215	greenhouse gas	156
agreement	162	clean energy	144
member	138	Great Lakes	138

4.3.3 State-Specific Topics

State-level analysis highlights distinct regional priorities:

- **Western States (Environment)**: California and Oregon are heavily associated with terms regarding ”climate change” and ”clean energy”.
- **Texas and Arkansas (Administrative)**: Unlike the environmental focus of the West Coast, these states feature high frequencies of ”driver license” and ”economic development,” indicating agreements focused on commerce, reciprocity, and bureaucracy.
- **New Jersey (Health Cooperation)**: Unique occurrences of ”Salud Chihuahua” and ”servicios salud” indicate specific health-related cooperation with the Mexican state of Chihuahua, likely reflecting consular or local health initiatives.
- **Alaska (Indigenous Relations)**: Analysis of Alaska’s documents reveals specific references to the ”Nisga Nation,” highlighting unique cross-border cooperation with Indigenous First Nations in British Columbia.
- **Northern Border States**: Maine, Michigan, Indiana, and Wisconsin frequently feature terms like ”forest fire”, ”crossing agreement”, ”Great Lakes”, and ”Ontario Quebec,” reflecting shared infrastructure and environmental stewardship with Canada.

5 Proof of concept

5.1 Tasks 1, 7, 9

The Proof of Concept (POC) regarding the OCR pipeline establishes the technical feasibility of transforming a large corpus of scanned legal agreements into a structured dataset optimized for computational 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, subsequently enabling complex algorithmic analysis to determine critical metrics including document length, the prevalence of boilerplate language, and the substantive scope of international cooperation.

5.1.1 Determining documents length

The core extraction layer is built upon the docTR (Document Text Recognition) library, utilizing a pretrained 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).

5.1.2 Analyzing frequency of recurring clauses

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 are classified to groups appearing more than 10, 5, and 1 time, those more than 5 times could be perceived as standard diplomatic protocol (within one state), whereas clauses that fail to form clusters (labeled as -1) are identified as unique, custom-drafted terms specific to a particular negotiation.

5.1.3 Identifying areas of cooperation

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.

5.2 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.

5.2.1 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 Technology"). The NER model frequently fragmented these into separate entities rather than recognizing 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.

5.2.2 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.

5.2.3 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.

5.3 Tasks 6: Determine Terms of Validity

The following methods are implemented to determine the terms of validity

5.3.1 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.

To assess the computational overhead of temporal validity extraction, we measured wall-clock runtime and process-level memory usage for the HeidelTime-based pipeline on the evaluation corpus. The results compare configurations with and without TreeTagger-based POS annotation, highlighting the modest runtime cost of POS tagging and stable memory behavior.

HeidelTime config	Total (s)	s/doc	RSS Δ (MB)
POS=NO	135.05	4.50	-28.68
TreeTagger	144.06	4.80	+0.03

Table 4: Runtime and memory for HeidelTime-based validity extraction on 30 documents.

5.3.2 POC Main Validity Output (Hybrid Validity Extractor)

The POC main method implements a hybrid extraction pipeline combining rule-based retrieval, neural semantic filtering, and heuristic temporal parsing. Validity candidate sentences are first retrieved using keyword triggers and context expansion. These candidates are then filtered using zero-shot classification and NLI-style verification to retain only sentences that semantically entail a validity clause. From the verified evidence, dates and durations are extracted using rule-based temporal parsing and assigned to effective date, end date, or duration using anchor-based heuristics. The method outputs a document-level validity status together with supporting evidence spans.

5.3.3 POC Baseline Keyword Validity (Presence-Only Flag)

This baseline detects the presence of validity language without performing any temporal extraction. The document text is scanned for predefined validity-related keywords using regular expressions. If any keyword is found, the document is marked as containing a validity clause. This method does not extract or normalize dates or durations and serves as a minimal reference point for evaluating the benefit of structured temporal extraction.

5.3.4 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. Extracted temporal expressions are normalized where possible, and simple positional heuristics are applied to distinguish start and end dates. Unlike the hybrid method, this baseline does not employ semantic filtering or neural verification, 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 5: 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 6: Duration extraction and explicitly stated end dates (%) on 30 California agreements.

The above 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.

State	#Docs	Wall (s)	CPU (s)	RSS Peak (MB)	RSS Δ (MB)	Py Heap (MB)
California	30	137.39	948.31	840.84	1.76	0.59

Table 7: Runtime and memory usage for Tasks 8 and 11 on 30 international agreements from California. Results correspond to a single end-to-end execution including lexical retrieval, zero-shot MNLI classification, NLI-style verification, and rule-based post-processing.

5.4 Task 8: Determine Conditions for Extending the Agreement

The proof-of-concept for identifying conditions for extending or renewing an agreement follows a hybrid clause-detection pipeline inspired by state-of-the-art contract analysis approaches evaluated on datasets such as CUAD and LEDGAR. In the first stage, high-recall lexical rules are applied to OCR-extracted agreement text to retrieve candidate clauses containing renewal- and extension-related triggers, with additional context expansion to account for OCR noise and clause fragmentation.

In the second stage, candidate clauses are subjected to zero-shot clause-type classification using an MNLI-based transformer model, serving as a lightweight proxy for task-specific legal language models without requiring fine-tuning. To further reduce false positives, an NLI-style verification step is applied to assess whether the candidate text semantically entails the presence of renewal or extension conditions. Finally, rule-based post-processing distinguishes between automatic renewals and extensions requiring mutual agreement, and extracts associated notice periods where explicitly stated. Supporting evidence spans are retained for interpretability and manual validation.

Only 40 percent of the analyzed agreements contain explicit evaluation or review clauses, confirming that evaluation of implementation is optional and far less standardized than validity or renewal provisions.

Evaluation status	Count	Percent (%)
Evaluation clause found	12	40.0
Evaluation clause absent	18	60.0

Table 8: Presence of the clauses determining the condition of extension across 30 California agreements

5.4.1 Task 11: Indicate Whether the Agreement Includes an Evaluation of Its Implementation

The proof-of-concept for detecting clauses related to the evaluation of agreement implementation is implemented using a hybrid retrieval and semantic verification approach informed by prior work on legal clause classification. Initially, high-recall lexical patterns are used to identify candidate text segments referencing evaluation, review, monitoring, or assessment of an agreement's implementation, ensuring robustness to OCR-induced variability.

Candidate segments are then processed using a zero-shot MNLI-based classifier to assess their relevance to evaluation semantics. An additional NLI-style verification step determines whether the candidate text entails an explicit obligation or mechanism for evaluating implementation, analogous to ContractNLI-style reasoning but without task-specific fine-tuning. The final output is a binary indicator denoting the presence or absence of implementation evaluation clauses, accompanied by extracted evidence spans to support transparency and downstream analysis.

Renewal type	Count	Percent (%)
By mutual agreement	19	63.3
Automatic renewal	6	20.0
No renewal clause (absent)	5	16.7

Table 9: Task 11: Distribution of renewal clauses across 30 California agreements.

Most agreements specify renewal by mutual agreement, while automatic renewal is comparatively rare and a non-negligible fraction of agreements contain no explicit renewal provision.

5.5 Tasks 4, 10, 12, 13

5.5.1 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.

5.5.2 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.

5.5.3 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 10: Fulfillment of Specific Project Tasks

6 Final Results

6.1 Task 1: Identification of Areas of Cooperation

To identify the thematic scope of the agreements, we employed a multi-label zero-shot classification approach. Unlike traditional supervised learning, which requires annotated training data for every specific label, zero-shot classification allows the model to predict class relevance based on descriptive labels alone.

We utilized the `facebook/bart-large-mnli` model, a pre-trained textual entailment model capable of determining the probability that a given text sequence corresponds to a specific topic label. We defined a taxonomy of potential cooperation areas, including "Technology & Innovation," "Environment & Green Energy," "Trade & Economic Development," and "Education & Students."

For each agreement, the full text was passed to the model, which assigned a confidence score to each label. A threshold of 0.5 was applied; labels exceeding this confidence score were considered "active topics" for the document. This multi-label configuration acknowledges that a single international agreement often covers multiple overlapping sectors (e.g., a trade agreement that also includes educational exchange provisions).

The analysis revealed a strong emphasis on modernization and sustainability within state-level international agreements. As illustrated in Figure 6, "Technology & Innovation" and "Environment & Green Energy" emerged as the most frequent areas of cooperation, surpassing traditional categories such as "Culture & Arts."

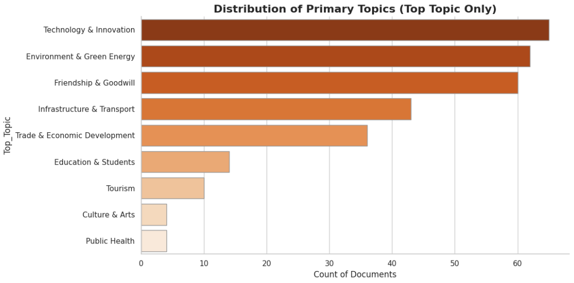


Figure 6: Frequency of all detected topics across the dataset. Technology and Environmental cooperation are the dominant themes.

Geographically, the distribution of topics varies significantly by state. Figure 7 highlights that larger economies like California and Texas exhibit the highest volume of agreements in "Trade & Economic Development" and "Technology." Conversely, states with fewer agreements often focus on specific niche partnerships. The heatmap analysis confirms that while some topics are universally relevant, the intensity of cooperation in high-tech and green energy sectors is concentrated in key states with established international policy frameworks.

The time and memory consumption for this task is described in Table 11.

Metric	Value
Number of documents	298
Total runtime (s)	410
Average time per document (s)	1.38
Peak Python memory (GB)	4.7

Table 11: Runtime and memory profile for facebook/bart-large-mnli on the full dataset.

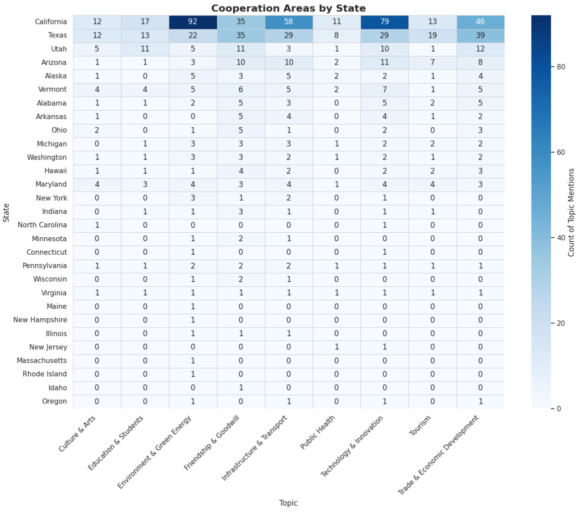


Figure 7: Heatmap of cooperation areas by state, showing the density of agreements in specific sectors.

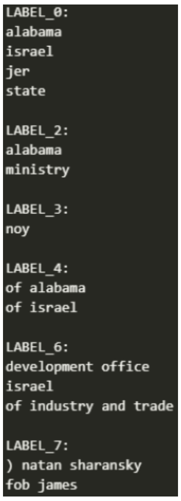
6.2 Task 2 - Agreement Parties

6.2.1 Task 2 methodology

For Task 2, we extracted the contracting parties of each agreement. We first implemented a baseline using a legal NER model (dandoune/legal-NER) as a token classification approach. We ran the model on a representative sample document per state, converted predicted token labels into spans, and saved the resulting entity fragments. We used this primarily as a sanity check, because we observed that token-level NER often breaks long party names into fragments and struggles when party strings include nested entities such as countries, agencies, and individual names.

We then implemented our main approach using an instruction-tuned LLM (Mistral-7B Instruct, loaded from a GGUF file via llama-cpp-python). We prompted the model to extract only the direct contracting parties and to return a strict JSON object containing a list of parties with name and role. We stored both the raw model output and a parsed JSON version when parsing succeeded. We measured extraction quality at the system level by tracking JSON parse success.

6.2.2 Task 2 results discussion



```
LABEL_0:
alabama
israel
jer
state

LABEL_2:
alabama
ministry

LABEL_3:
noy

LABEL_4:
of alabama
of israel

LABEL_6:
development office
israel
of industry and trade

LABEL_7:
) natan sharansky
fob james
```

Figure 8: Sample NER output.

The output on Figure 8 shows that the NER model captures isolated tokens and partial spans, but it does not reliably assemble the complete contracting-party strings that are typical in agreements (which often contain multiple nested entities, long titles, and jurisdiction descriptors). Because the extracted entities are fragmented and the label IDs do not map cleanly to party-specific categories, the approach does not support accurate party extraction

Mistral prompting method worked great for the task goal because it produced structured party lists that can be aggregated across documents. JSON parsing succeeded for almost all outputs (98% parse success across processed documents), which made the results usable in a real-world scenario.

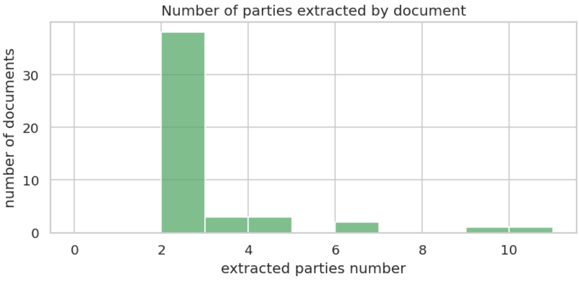


Figure 9: Number of parties extracted histogram.

Mistral usually extracted 2 parties only (Figure 9i), which is an expected outcome.

6.2.3 Task 2 time and memory efficiency

Table 12: NER profiling summary

Metric	Value
Max RSS (MB)	5736.453
Mean time (s)	1.282

Table 13: Mistral inference profiling summary

Metric	Value
Max RSS (MB)	5833.953
Mean time (s)	304.395

6.3 Task 3 - Agreement Type

6.3.1 Task 3 methodology

For Task 3, we identified the agreement type of each document using two complementary methods. We first applied zero-shot classification using facebook/bart-large-mnli through the HuggingFace pipeline. We defined a fixed label set of common agreement categories and we classified each document into exactly one label. We saved the predicted label and its confidence score, and we also kept the top-3 candidates for later comparison and reporting. We used this approach because it produces consistent categories suitable for aggregate plots and tables, while still providing a confidence measure that can be used for triage.

We then performed open-ended agreement type extraction with Mistral. We prompted the model to return a JSON object containing agreement type and a short description. We stored both the raw output and the parsed JSON when possible, and we tracked parsing success to quantify how reliably the model produced structured responses. We used this method because it is not constrained by a fixed label set and can capture types that were not included in the zero-shot taxonomy, at the cost of potentially more variability in naming and normalization.

6.3.2 Task 3 results discussion

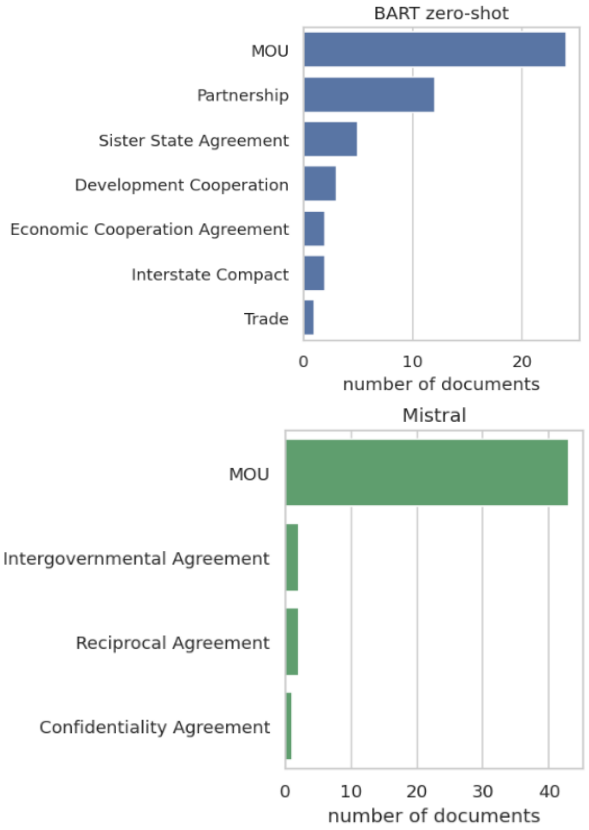


Figure 10: Types of agreements detected by both methods.

As visible on Figure 10 BART spreads predictions across several labels (MOU, Partnership, Sister State Agreement), suggesting the fixed label set strongly shapes the output. Mistral, on the other hand, strongly collapses most documents into *Memorandum of Understanding*. This can be good if the corpus truly contains mostly MOUs, but it can also indicate a bias toward a common legal template phrase. Zero-shot classification results are much more realistic.

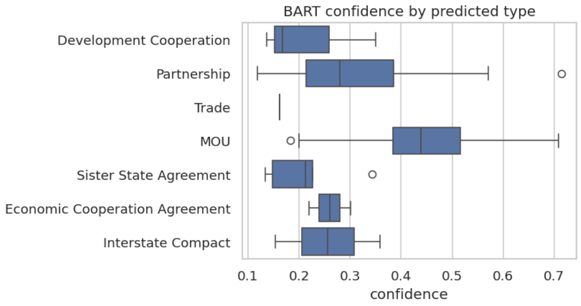


Figure 11: BART classification confidence.

Confidence from Figure 11 is moderate overall (many scores 0.2–0.6), which is expected for zero-shot on OCR text. The higher-confidence predictions tend to be for common classes (MOU/Partnership). Rare classes are less stable.

6.3.3 Task 3 time and memory efficiency

Table 14: BART zero-shot classification profiling summary

Metric	Value
Max RSS (MB)	7488.699
Mean time (s)	186.082

6.4 Task 5 - International Organizations Mentioned

6.4.1 Task 5 methodology

For Task 5, we extracted international organizations mentioned in the agreements using the Mistral model. We initially considered reusing an NER-based approach similar to Task 2, but we did not proceed with NER for this task because the Task 2 NER baseline produced fragmented and unreliable entity spans on this OCR-heavy legal text (especially for long, compound entity names). Based on that outcome, we chose to rely directly on LLM-based extraction for Task 5.

We prompted Mistral to output a JSON object containing a list of international organizations with name and a short type description, and we explicitly instructed the model to focus on cross-border or intergovernmental bodies while ignoring purely domestic entities unless they were clearly part of an international structure. We stored both the raw model output and the parsed JSON (when parsing succeeded) and tracked JSON parse success as a basic reliability metric.

6.4.2 Task 5 results discussion

Mistral produced, similarly to Task 2, parseable structured output with a high success rate (92% parse success rate). Most analyzed documents did not mention any international organizations according to model's predictions (Figure 12). The main observed limitation was semantic: *international organization* boundaries are sometimes ambiguous in analyzed texts, and the model can occasionally include national ministries or governments depending on context. Still, the results are strong enough for descriptive analysis.

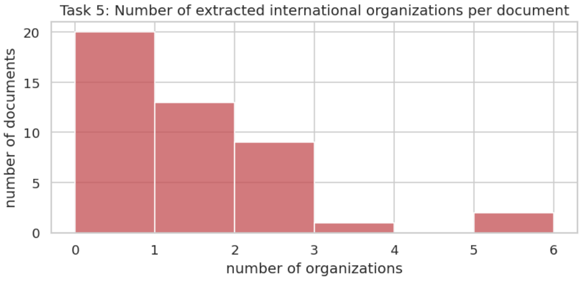


Figure 12: Number of extracted international organizations histogram.

6.4.3 Task 5 time and memory efficiency

Time and memory efficiency for Task 5 align with these reported for Task 2 as the same method was used to extract the information.

6.5 Runtime and Memory Profiling for task 8 and 11

Metric	Value
Number of agreements	298
Total wall-clock time (s)	1423.41
Total CPU time (s)	9511.17
Average time per agreement (s)	4.78
RSS memory at start (MB)	755.17
RSS memory at end (MB)	774.96
Peak RSS memory (MB)	779.70
RSS memory increase (MB)	19.79
Peak Python heap (MB)	2.83

Table 15: Runtime and memory profiling for the full dataset evaluation of Tasks 8 and 11 across all U.S. state agreements. The entire dataset was processed in a single pass without re-execution for profiling.

We evaluated the complete dataset of 298 U.S. state-level agreements for Tasks 8 and 11 in a single end-to-end run, ensuring that runtime and memory measurements were collected without requiring repeated executions.

These results demonstrate that the proposed pipeline scales reliably to several hundred long-form legal agreements and can be executed efficiently in a single-pass setting, making it suitable for large-scale agreement mining without repeated recomputation for performance analysis.

6.6 Task 6: Term of Validity

The following table reports end-to-end runtime and memory consumption for TreeTagger-based HeidelTime validity extraction on the full evaluation corpus.

Metric	Value
Number of documents	298
Documents with errors	3
Total runtime (s)	813.66
Average time per document (s)	2.71
Median time per document (s)	2.55
95th percentile time per document (s)	3.71
Peak Python memory (MB)	2.16

Table 16: Runtime and memory profile for TreeTagger-based HeidelTime validity extraction on the full dataset.

Table 12 summarizes the computational cost of running the TreeTagger-based HeidelTime validity extraction pipeline on the full dataset of 298 agreements. All documents are processed regardless of extraction outcome, and runtime and memory usage reflect end-to-end temporal normalization and clause-level analysis.

End date source	Agreements
Explicitly stated	28
Implicit / inferred	161
Not available	109

Table 17: Source of extracted end dates across agreements. Explicit end dates are directly stated in the text, while implicit end dates are inferred from temporal context.

Table 13 reports the availability and source of agreement end dates identified by the system. While all agreements are processed, only a subset explicitly states an end date, with additional cases inferred via heuristic fallback rules, and the remainder lacking sufficient temporal information.

Table 18: Extracted agreement validity periods with document evidence.

Document	Start Date	End Date	Duration	Extracted Evidence
Texas_48	2019-09-27	2021-08-31	–	"This Statement of Mutual Cooperation is intended to be in place until August 31, 2021."
Texas_50	2023-04-27	2025-03-31	–	"DURATION OF OPERATION ... applied from April 27, 2023, until March 31, 2025 ..."
NewJersey_9	2008-05-07	2010-09-30	–	"... will terminate on September 30, 2010 ..."
Alabama_3	–	–	three years	"This Memorandum of Intent shall enter into effect upon signature and shall remain in force for a period of three years."
Arkansas_6	2018-01-01	–	five years	"This Agreement shall be in full force and effective for a period of 5 years."

Table 14 presents representative high-confidence agreement validity periods extracted by the system, including start dates, end dates, or explicitly stated durations, along with the supporting textual evidence. These examples illustrate cases where temporal validity is expressed using clear legal force language, such as effective dates, termination clauses, or fixed-term durations.

6.7 Task 8: Renewal and Extension

Table 19: Distribution of renewal and extension clause types across $N = 298$ agreements.

Renewal type	Count	Percent (%)
By mutual agreement	102	34.23
No renewal clause	101	33.89
Automatic renewal	85	28.52
Uncertain / ambiguous	10	3.36

Table 12 shows that renewal provisions are distributed relatively evenly across agreement types. Slightly over one third of the agreements require explicit mutual agreement for renewal, while a comparable proportion contains no renewal clause at all. Automatic renewal is present in just under one third of the documents, indicating that fixed-term agreements with implicit continuation are common. Only a small fraction of cases remain uncertain, typically due to vague or underspecified language that prevents a confident classification.

Table 20: Renewal and extension clauses extracted by the system (Task 8).

Document	Renewal Type	Extracted Evidence
Alabama_4	By mutual agreement	"Additional items may be added to this document as mutually agreed upon by both parties."
Alaska_2	Automatic	"Renewed as of September 11, 2015, September 11, 2017, September 11, 2019 and September 11, 2021."
California_1July5	Automatic	"Unless canceled by any party, this MOU will automatically be renewed for one-year periods."

6.8 Task 7: Quantitative Analysis of Agreement Length

We performed a quantitative analysis of the physical and lexical length of the agreements to understand the complexity and detail level of these legal documents.

The length of each agreement was measured using two metrics: page count and total word count. Page counts were extracted directly from the PDF metadata and layout analysis, while word counts were derived following Optical Character Recognition (OCR) and text extraction. This dual approach helps account for formatting differences, where font size or layout might skew the perception of length based on page count alone.

The distribution of agreement lengths is highly right-skewed. As shown in Figure 13, the vast majority of agreements are concise, typically spanning between 1 to 5 pages. This suggests that most state-level agreements are high-level Memoranda of Understanding (MoUs) rather than detailed, multi-chapter treaties.

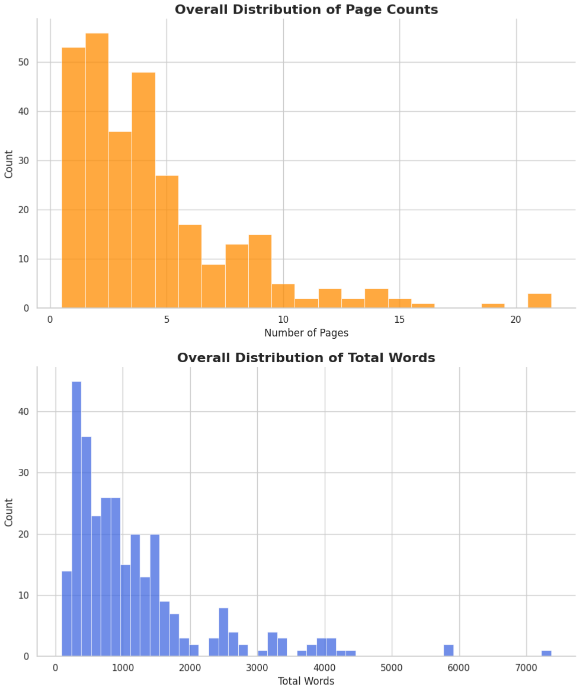


Figure 13: Overall distribution of page counts (top) and total words (bottom). The data indicates a prevalence of short, high-level documents.

However, significant outliers exist. Figure 14 presents a boxplot analysis by state, revealing that California and Texas not only have the most agreements but also the greatest variance in document length. California, in particular, exhibits several extensive agreements that significantly exceed the median, likely reflecting complex regulatory alignments or multi-jurisdictional frameworks. In contrast, states like Arkansas and Alabama show more uniformity in document length, suggesting a standardized template or scope for their international engagements.

This task didn't require substantial memory or runtime resources.

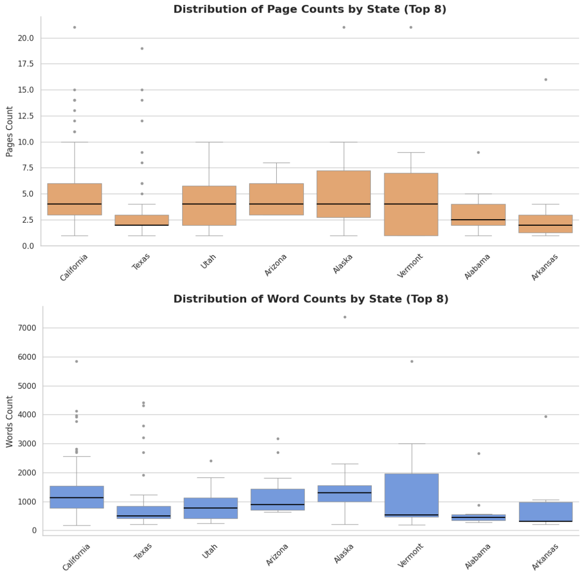


Figure 14: Distribution of page (top) and word (bottom) counts by state. Boxplots highlight the variability and presence of complex, lengthy agreements in major states.

6.9 Task 9: Analysis of Recurring Clauses

To assess the standardization of legal language across the corpus, we analyzed the frequency of recurring clauses and common lexical patterns.

We adopted a semantic clustering approach to identify recurring clauses, as exact string matching fails to capture variations in wording that convey the same legal meaning. The text of the agreements was segmented into clause-level blocks. We then generated semantic embeddings for each block using the `all-MiniLM-L6-v2` model.

These high-dimensional embeddings were clustered using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). This method allows for the identification of semantically similar groups of text (e.g., liability waivers or termination clauses) without requiring a pre-defined number of clusters.

The lexical analysis (Figure 15) highlights the terminology central to these documents. Unigrams such as “cooperation,” “development,” and “energy” dominate, while bigram analysis reveals specific focus areas like “clean energy,” “economic development,” and “climate change.”

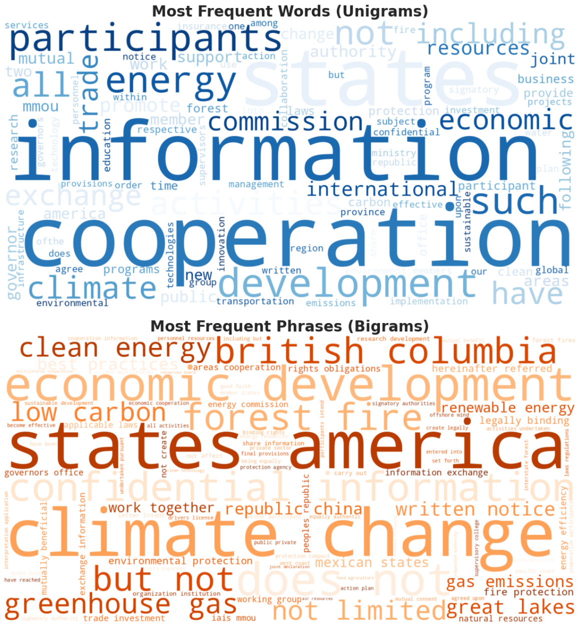


Figure 15: Word cloud of most frequent unigrams (top) and list of top bigrams (bottom), illustrating the core vocabulary of the agreements.

The structural analysis of clauses indicates a low degree of standardization for boilerplate provisions and a significant uniqueness in substantive sections. Figure 16 shows the distribution of clause types by frequency.

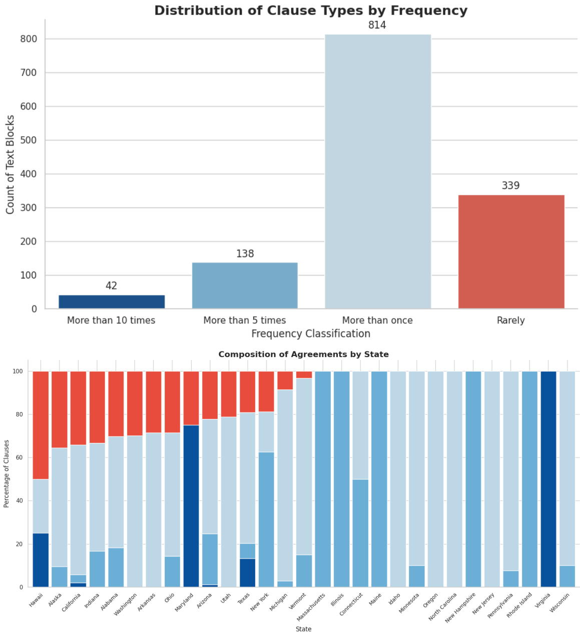


Figure 16: Distribution of clause types by frequency (top) and composition by state (bottom). A large portion of text blocks are unique (Rarely), less portion of them represents highly recurring language.

A large portion of the text blocks (labeled ”Rarely”) appear only once, representing the specific, custom content of each agreement (e.g., specific project details). Lower number of clauses appear ”More than 10 times” or ”More than 5 times”, corresponding to standard legal provisions such as entry into force, dispute resolution, and validity terms. This distribution confirms that the core subject matter is tailored to each partnership, only sometimes the content relies on recurring, standardized language.

The time and memory consumption for this task is described in Table 21.

Metric	**Value**
Number of documents	298
Total runtime (s)	122
Average time per document (s)	0.41
Peak Python memory (GB)	4.7

Table 21: Runtime and memory profile for all-MiniLM-L6-v2 on the full dataset.

6.10 Task 11: Evaluation of Implementation

Table 22: Task 11: Presence of implementation evaluation or review clauses in the agreements across $N = 298$ agreements.

evaluation_present	count	percent
Absent	192	64.430000
Present	106	35.570000

As reported in Table 13, explicit clauses requiring evaluation, monitoring, or review of implementation are present in approximately one third of the agreements. The majority of documents do not include such provisions, suggesting that implementation oversight is not systematically formalized in state-level agreements and is often left implicit or to informal coordination mechanisms.

Table 23: Representative examples of clauses indicating evaluation or review of implementation automatically extracted by the system (Task 11).

Document	Eval. Status	Extracted Evidence
Alaska_7	Present	“The Parties shall hold an annual meeting to review the terms of this Agreement and its implementation.”
California_1July5	Present	“This report will evaluate the implementation of the program and suggest proposals for improvement.”
Alaska_9	Present	“The parties shall review the status and progress made under this Memorandum of Understanding at the end of the first year.”
Arkansas_1	Present	“The Parties shall regularly review the execution of the Action Plan developed under this Agreement.”
Utah_14	Present	“Representatives of the Parties will evaluate the activities and results related to program development and implementation.”

7 Work distribution

Task	Workload Assessment [hours]	Asignee
Dataset preprocessing and OCR implementation	5	Mateusz Wiktorzak
EDA	3	Paulina Kulczyk
Project proposal document	1	Paulina Kulczyk
SOTA analysis	5	Muhammad Fahim Asim, Paulina Kulczyk, Mateusz Wiktorzak, Mateusz Zagórski
Tasks 1, 7, 9	20	Mateusz Wiktorzak
Tasks 4, 12, 13, 14	20	Paulina Kulczyk
Tasks 2, 3, 5	20	Mateusz Zagórski
Tasks 6, 8, 11	20	Muhammad Fahim Asim
Review 1	3	Muhammad Fahim Asim
Review 2	3	Mateusz Zagórski
Combining deliverables	3	Mateusz Zagórski

Table 24: Project responsibilities divided among all team members

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.

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

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.

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.

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

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.

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.

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.

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.