International Agreements Data Base mining
Project Proposal for NLP Course, Winter 2024

Tomasz Siudalski
Warsaw University of Technology
01161590@pw.edu.pl
Weronika Plichta
Warsaw University of Technology
01194060@pw.edu.pl
Michał Taczała
Warsaw University of Technology
01149437@pw.edu.pl
supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

This project, realized in cooperation with faculty members from the University of Lodz, aims to explore the application of Natural Language Processing (NLP) techniques to analyze a vast collection of international agreements from U.S. states and municipalities. The research will focus on automating the extraction of 13 key attributes proposed by legal researchers, such as areas of cooperation, parties involved, agreement types, and recurring clauses. Leveraging state-of-the-art NLP tools the project will investigate tasks including Named Entity Recognition, relation extraction, or clause frequency analysis and address challenges such as ambiguity and the formal structure of legal documents. By streamlining the analysis process, the project aims to reduce reliance on manual document reviews and establish a foundation for efficient analysis of similar legal datasets, enhancing the accessibility of actionable insights from complex legal agreements.

1 Introduction

The scientific goal of this project is to analyze a comprehensive database of international agreements concluded by U.S. states to uncover patterns, trends, and

key elements of these agreements. These agreements, which cover diverse topics such as economic development, cultural exchange, and environmental collaboration, provide valuable insights into the evolving dynamics of global partnerships. Traditionally, extracting relevant information from such agreements has been a difficult and time-consuming process requiring expert knowledge. By employing advanced data analysis and NLP techniques, this project seeks to answer several pivotal questions regarding the nature and structure of these agreements. The study will focus on identifying key attributes such as areas of cooperation, parties involved, agreement types, and clauses related to duration, extension conditions, and coordination with other entities. Our project also aims to accelerate and streamline the information extraction process for future agreements enabling faster and more consistent analysis of these documents, eliminating the need for exhaustive manual reviews.

1.1 Research questions

The primary goal of the analysis is to address the following challenges:

- Identification of areas of cooperation mentioned in the agreements.
- Identification of the parties involved (states, institutions, local partners).
- Identification of the types of agreements (e.g., Memorandum of Understanding, Sister Cities Agreement, etc.).
- Determination of the percentage of agreements under the patronage of Sister Cities International.
- Identification of international organizations mentioned in the agreements.
- Determination of the terms of validity for each agreement.
- Identification of the length of each agreement (number of pages or words).
- Determination of the conditions for extending each agreement (automatic/ by decision).
- Analysis of the frequency of recurring clauses in the agreements (always, often, rarely) – the level of detail in the agreements.
- Identification of the partners with whom the agreements tend to be more detailed.

- Indication of whether the agreement includes an evaluation of its implementation.
- Identification of whether the agreement mentions any coordination of activities with other entities (e.g., government, other cities/states, international organizations).
- Identification of whether the agreement refers to other legal documents.

1.2 Significance of the project

This project addresses the need for efficient document analysis in legal and administrative contexts by automating the extraction of critical information from international agreements. By reducing the time and effort required for manual reviews, the methodology enables stakeholders to quickly retrieve actionable insights, identify trends, and make decisions. Furthermore, the methodology of our project can be applied to similar datasets, such as treaties or trade agreements, showcasing the broader applicability and impact of this research.

2 Literature review

Analysis of legal and formal documents like ”paradiplomacy” is quite a different task from the analysis of a standard text. There are a lot of specific words or complex language. However, in the last few years there has been an increase in legal NLP research [6], which causes more data availability and code reproducibility.

To analyze formal text, NLP and computational methods are helpful, to extract information from legal texts. For that fine-tuned models like BERT can handle domain language. Also, Named-entity-recognition and relation extraction can help find relationships between words in documents.[10] What is worth noting is that legal documents often contain ambiguous words that might be hard to analyze. Another potential problem might be that the dataset is unbalanced and contains most of the agreements of a certain type, which might cause the model to learn the most common type of agreement instead of the pattern.

In another study [9] for quite a similar task to paradiplomacy, after annotating data by hired specialists, authors used BERT and Named Entity Recognition for modeling and sentiment analysis. It turned out, that in this type of document, most of the sentences are neutral. A small number of sentences were positive, and few were negative(mostly encountered in the ”unmet goals” part). The preprocessing to achieve such results contained: text extraction (because some PDFs were saved as images), then splitting text into separate paragraphs, spelling corrections(mostly due to unsatisfactory quality of the image-pdf data), and converting to English (because documents were French).

The next study [8] addresses the same problem as the paradiplomacy task. It shows that the framework's use of NER and rule-based extraction is quite useful for handling the structured and specific language that is often used in legal agreements. Since both tasks are very similar, using these techniques to identify agreement types, and international organizations could improve the results of the paradiplomacy project. However, the problem with ambiguity is mentioned in this study once more, as the authors stress this problem and its importance, as it's very common for legal documents and is very likely to appear in paradiplomacy documents as well.

2.1 Blackstone

Blackstone presents a solution for automatically identifying references to legal documents within agreements. Built on spaCy, this open-source legal text processing model is specifically trained to recognize various types of legal references through its Named Entity Recognition (NER) component. It can identify citations, case names, legal instruments (like acts and conventions), specific provi-

sions within those instruments, and court references. While initially trained in UK case law, the model has shown good generalization to other legal systems. However, it's important to note that this is still a prototype with around 70% accuracy (F1 score) for its NER component. Despite these limitations, Blackstone's specialized legal NLP capabilities make it a valuable tool for systematically extracting and analyzing legal document references from agreements, which we want to use in reference to point 13 of our guidelines. Link to GitHub: `https://github.com/ICLRandD/Blackstone/blob/master/README.md`

2.2 LEGAL-BERT

LegalBERT [2] is a family of BERT-based language models specifically trained for legal text processing. Unlike the original BERT, which is pre-trained on general-purpose texts like Wikipedia, LegalBERT adapts to the legal domain by using domain-specific corpora, such as EU and UK legislation, US court cases, and contracts. The training data, spanning 12GB, consists of diverse legal documents that enhance the model's understanding of legal-specific vocabulary and syntax. The authors compared three strategies for BERT in legal tasks: using base BERT, further pre-training BERT on legal corpora, and pre-training BERT from scratch. They showed that the two latter options yield superior performance in legal tasks compared to using the base BERT model by evaluating them on several legal tasks, including multi-label text classification, binary classification, and Named Entity Recognition (NER) for contract elements. LegalBERT is an open-source resource available on Hugging Face and may be a valuable tool in our research.

2.3 CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review

The Contract Understanding Atticus Dataset (CUAD) [5] is an NLP dataset created to support automated legal contract review as part of The Atticus Project. It contains over 13,000 expert annotations spanning 41 categories of contract clauses, extracted from 510 diverse contracts. These include clauses like governing law, anti-assignment, perpetual licenses, and non-compete agreements. CUAD aims to automate the time-consuming process of extracting key clauses from lengthy contracts which is usually performed manually by legal experts. CUAD serves as a benchmark for assessing NLP models in specialized domains. The authors fine-tuned Transformer-based models such as BERT and DeBERTa [4] on the dataset showing promising but not ideal performance. For example, DeBERTa-xlarge achieves a Precision of 44% at 80% Recall, highlighting substantial room for improvement. The dataset is particularly valuable for its high-quality annota-

tions, which include rigorous quality checks by trained legal professionals. Models fine-tuned on CUAD have the potential to significantly reduce the time and cost of extracting valuable information from contracts.

2.4 Phi-3

Phi-3-mini-4k-Instruct [7] is a 3.8 billion parameter language model developed by Microsoft, designed to balance performance and computational efficiency. Its compact size makes it ideal for projects with resource constraints, such as ours, where scalability and precision are critical. The model employs a data-optimal training approach, utilizing heavily curated web and synthetic data, enabling it to match much larger models like GPT-3.5 on key NLP benchmarks such as MMLU and HellaSwag. Its transformer-based architecture incorporates advanced features like LongRope encoding for extended context handling and block-sparse attention for efficient memory usage. Post-training refinements, including supervised fine-tuning (SFT) and Direct Preference Optimization (DPO), further enhance its capabilities in tasks like reasoning and information extraction. We chose Phi-3-mini-4k-Instruct due to its proven effectiveness in specialized tasks, making it particularly suitable for analyzing complex legal documents within our project's computational constraints.

2.5 LLama 3

LLama-3-8B-Instruct [3] is an 8-billion parameter language model developed with a focus on delivering state-of-the-art performance across a wide range of NLP tasks. Its larger size and more sophisticated architecture, compared to smaller models like Phi-3-mini-4k-Instruct, make it ideal for tasks requiring superior reasoning, contextual understanding, and nuanced information extraction. LLama-3-8B-Instruct leverages cutting-edge advancements in transformer-based architectures, including extended context handling and highly optimized attention mechanisms, enabling exceptional performance on benchmarks such as MMLU and HellaSwag. The model's training regimen combines large-scale curated datasets with robust fine-tuning techniques, such as supervised fine-tuning (SFT) and Direct Preference Optimization (DPO), ensuring high precision in specialized domains.

We selected LLama-3-8B-Instruct for some of the tasks due to its superior capabilities in handling complex legal documents and achieving a higher degree of accuracy than Phi-3-mini-4k-Instruct. Despite its increased computational demands and longer processing times, the model's enhanced ability to extract, reason, and analyze intricate information proved invaluable for our project's objectives, where precision and depth of understanding are critical.

2.6 Flair

The Flair/ner-english-ontonotes-large model [1] is a state-of-the-art Named Entity Recognition (NER) model built on the Flair NLP framework. Designed to excel in extracting entities from text with high accuracy, this model is fine-tuned on the OntoNotes 5.0 dataset, covering a wide range of entity types such as PERSON, DATE, ORGANIZATION, GPE (geopolitical entities), and more. Its architecture leverages contextual string embeddings, a unique feature of Flair, which combines character-level language modeling with pre-trained embeddings to deliver rich, context-aware representations of text.

Flair/ner-english-ontonotes-large is particularly well-suited for tasks requiring detailed entity recognition in complex documents. Its performance is comparable to larger transformer-based models, yet it benefits from a more lightweight implementation, which can be advantageous in scenarios with resource constraints. Additionally, its ability to handle fine-grained entity distinctions makes it a robust choice for applications like legal, biomedical, or financial text analysis.

We incorporated Flair/ner-english-ontonotes-large to speed up named entity recognition and support the LLM model. By offloading NER tasks to Flair, we reduced the LLM's computational load, enabling faster processing while maintaining high accuracy through Flair's specialized entity extraction capabilities.

3 Description of dataset

In a collaborative research initiative with faculty members from the University of Lodz and the University of Warsaw, our team obtained a collection of over 600 legal documents, all sourced from the HeinOnline legal database. These documents represent an array of international agreements, specifically focusing on two main categories: first, formal agreements between individual US states and their counterparts (states or provinces) in other countries, and second, city-to-city agreements such as Memoranda of Understanding and Sister Cities Agreements, which establish cultural and economic partnerships between municipalities worldwide. Initially, each document was preserved in its original form as a scanned PDF or as PDF files containing photographic images that were digitally inserted into the PDF format. To enhance accessibility and enable digital analysis, HeinOnline's database administrators employed optical character recognition (OCR) technology. This OCR process systematically converts the scanned images into text files.

Variation in the structure and level of detail

A key feature of this document collection is the wide variation in their structure and level of detail. Although some agreements are extensively detailed, containing comprehensive sections on objectives, responsibilities, implementation procedures, and legal frameworks, others are considerably more concise, presenting only basic terms and general principles of cooperation. This heterogeneity in document structure reflects the diverse nature of international agreements, which can range from highly formalized legally binding documents to more informal memoranda of understanding. The variation in detail and structure also appears to correlate with factors such as the scope of cooperation and the jurisdictional level of the participating entities.

Formal language

What's consistent across all these documents, however, is their use of formal, legal language. As official international agreements, they maintain a high level of formality in their writing style, employing specialized legal terminology, complex sentence structures, and standardized diplomatic phrases.

4 Data preprocessing

4.1 PDF to TXT

Data are divided into two main categories: some of them are PDF files with text that can be directly copied and interpreted, and the rest are PDFs created from scans or other image files. We used ORT from the Tesseract library to retrieve text values from images.

4.2 Text Cleaning

The text has been converted to lowercase. Stopwords(like then, moreover, so) have been deleted deleted, and special characters removed.

4.3 Fixing misspelled words

As mentioned in the study [9], Sentences obtained with OCR might contain errors in spelling (especially if the quality of the images is not satisfactory). For the POC stage, we haven't implemented fixing words yet, but it is an important step that will be done in the future if needed.

4.4 Annotating entities

Some entities like countries or state names, had to be annotated manually and in dependence on the task.

5 Exploratory data analysis

Figure 1: Most common words

As we can see in the picture 1, the most common words and phrases in agreements are ”party” and ”united state” which isn’t surprising as we deal with agreements between different American states.

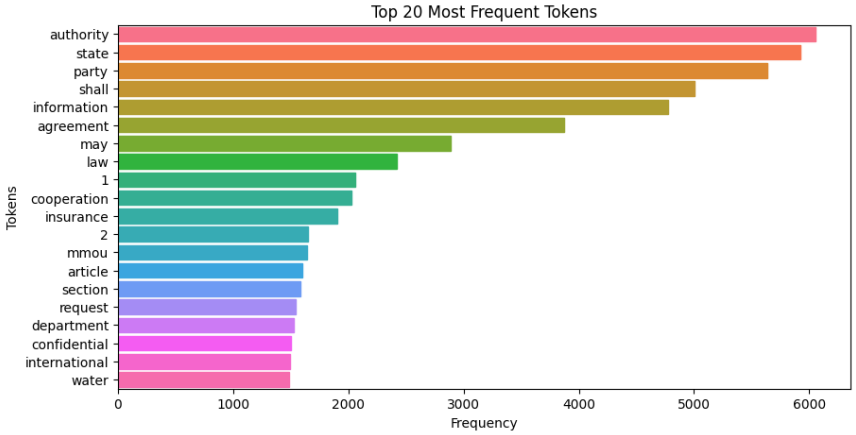

Figure 2: Most frequent words

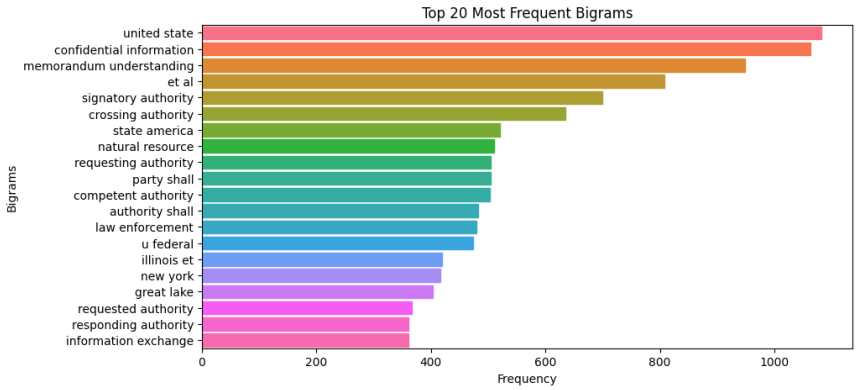

Figure 3: Most frequent bigrams

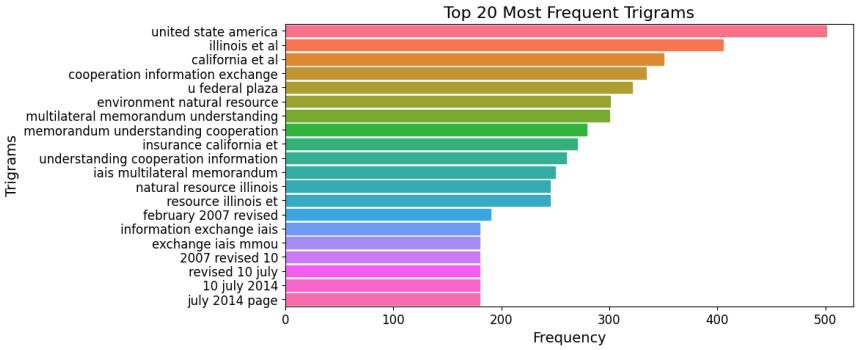

Figure 4: Most frequent trigrams

If we look at the most frequent words, bigrams, and trigrams, we can see some pattern, that a lot of words from the ”most frequent words” plot are present also in the bigram and trigram chart.

Moreover, most of the agreements consist of less than a thousand words 5. There are also some outliers with up to 20 thousand words.

The average length of a word for an article after text-cleaning is 7 letters. It seems a very big value, but there are 2 reasons for that.

- Most of the stopwords are short (and they have been deleted)
- Agreements are written with formal language which usually consists of longer words

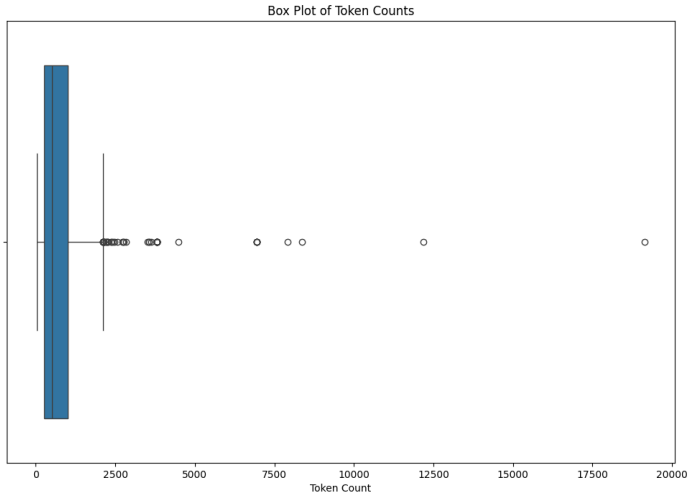

Figure 5: Number of words boxplot

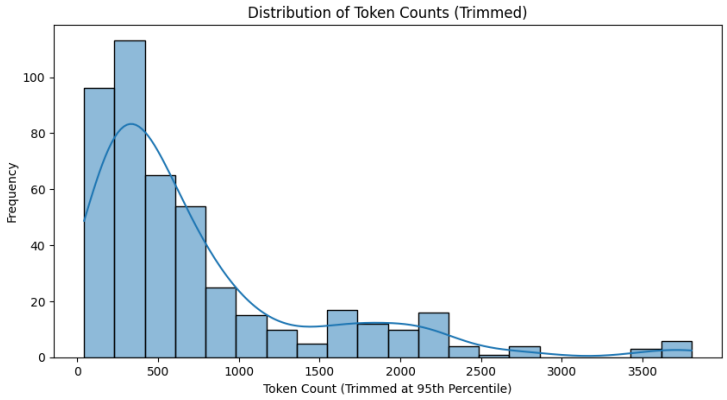

Figure 6: Trimmed number of words

6 Final solutions

In this section, we describe the methods used for each of the tasks. Sample results and LLM prompts can be found in the appendix A

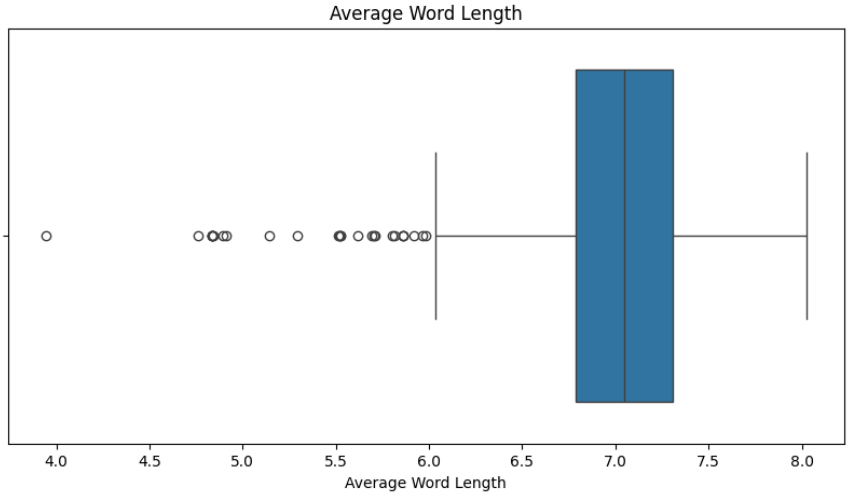

Figure 7: Caption

6.1 Identification of areas of cooperation

We extracted cooperation areas from international agreements by employing the Meta-Llama-3-8B-Instruct large language model and classifying the text of the agreements into predefined categories. These categories included ”Economic Development and Trade,” ”Education and Academic Exchange,” ”Cultural Exchange,” ”Tourism Promotion,” ”Environmental Protection and Sustainability,” ”Infrastructure Development and Urban Planning,” ”Public Health and Social Services,” ”Technology and Innovation,” ”Disaster Preparedness and Emergency Management,” ”Good Governance and Administrative Cooperation,” and ”Agriculture.”

To handle the often lengthy and complex agreements, we cropped the documents to 4096 characters, ensuring the model could process them effectively without exceeding its context limitations. Such length turned out to be enough to extract the most important areas from the document, as they were often mentioned at the beginning and many arguments fit in this size. The language model was guided by a detailed prompt, which provided clear instructions and definitions for each category. This allowed the model to analyze the text carefully and classify it into one or more relevant areas based on explicit mentions, specific projects, or clear associations found in the content. The model was also asked to add details to each category, ensuring more specific information is provided.

The outputs from each chunk were aggregated to produce a comprehensive set of cooperation areas for each agreement with specific details.

Most of the obtained results are in the correct format and seem to be valid, con-

taining both main area(s) and additional details. Only 5% agreements are missing the results. Expert validation of the details is required.

In Figure 8 top 10 main areas of cooperation can be seen.

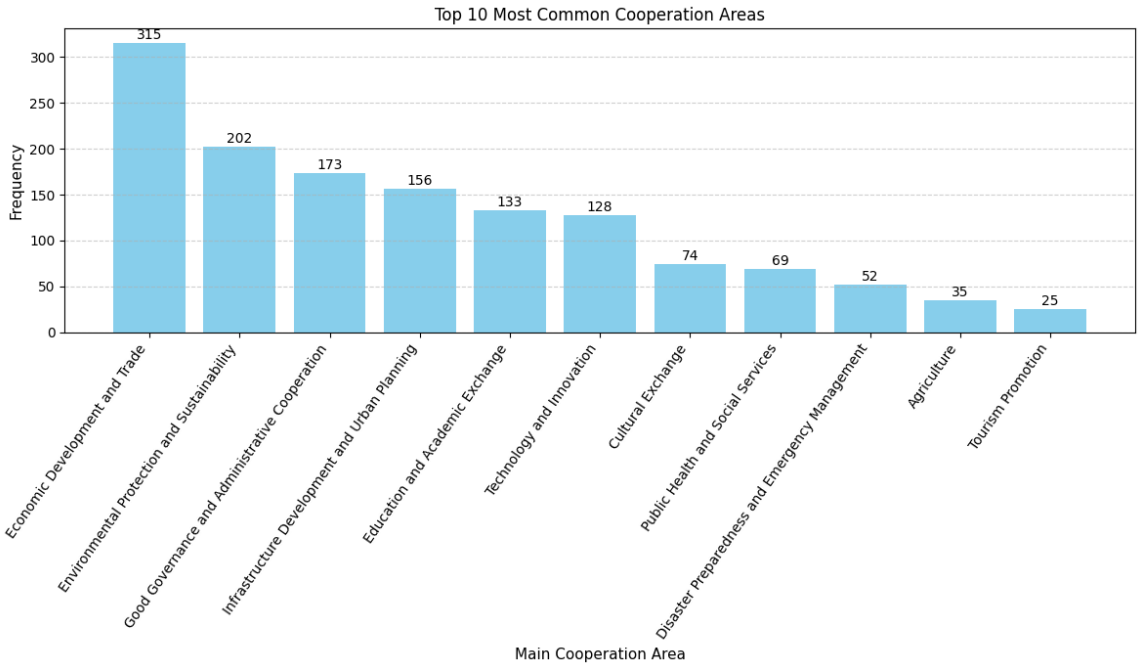

Figure 8: Most common cooperation areas mentioned in agreements

6.2 Identification of the parties involved

We extracted contracting parties from international agreements using the Meta-Llama-3-8B-Instruct model. The analysis focused mostly on the beginning portions of each agreement, as these typically contain key phrases indicating the contracting parties, such as ”This agreement is between,” ”entered into by,” or ”signed by.”

To process the text effectively, each document was divided into chunks, though in most cases, only the first chunk was analyzed. The Llama model was provided with a detailed prompt that instructed it to identify and extract the full names of contracting parties, ensuring compound names (e.g., ”Jiangsu Province, Republic of China”) were treated as single entities. It also ensured institutions (e.g., ”Bank of England”) were not confused with sovereign entities (e.g., ”United Kingdom”) and appended country names for non-U.S. entities for clarity.

The model returned results as a Python list of contracting parties for each analyzed chunk. This focused and streamlined approach allowed for accurate identification of contracting parties, leveraging Llama's advanced language capabilities while maintaining efficiency.

Most of the results are very good and in the correct format. Only 2% of data is missing. However, some of the extracted parties are invalid e.g., [State, Members] does not provide any information. Manual inspection found very little of such cases, but expert validation is required.

In Figure 9 we can see countries most often involved in the agreements.

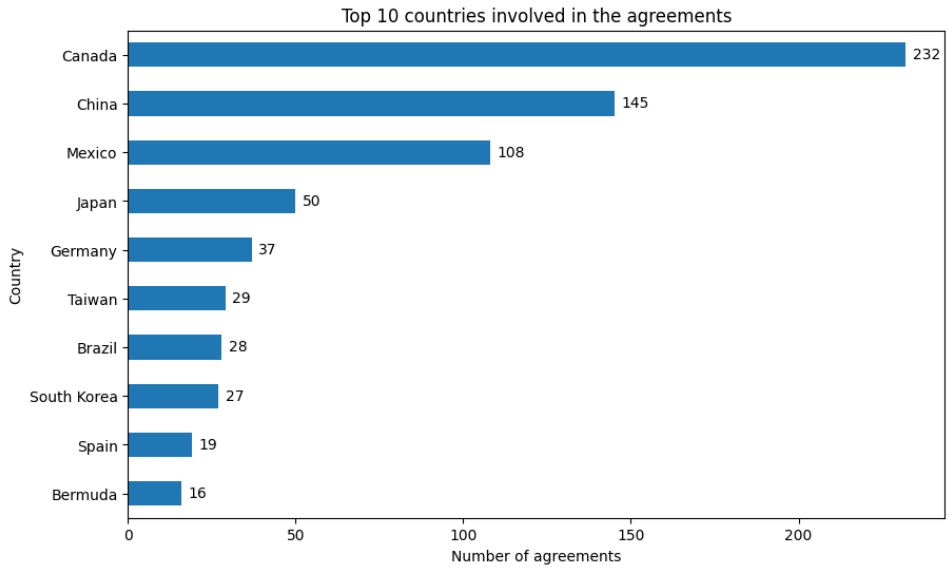

Figure 9: Countries (besides the USA) most often involved in the agreements

6.3 Identification of types of agreements

The purpose of this task was to classify agreements based on the type of agreement. Possible options were a Memorandum of Understanding, Sister Cities Agreement, Trade Agreement, Cultural Exchange Agreement, Technology Transfer Agreement, Educational Cooperation Agreement, Bilateral Cooperation Agreement, Partnership Agreement, and Investment Agreement. For this task, we used a "facebook/bart-large-mnli" model from the transformer library. Although this model is not very heavy and complex, since the type of agreement is usually stated directly at the beginning of the document the model didn't have many problems specifying the

type. This implementation yielded divided agreements into several subgroups for each type of agreement. The most common type by far is a ”Memorandum of Understanding”. One thing that is worth mentioning is that some agreements didn't specify any special type of agreement. In this case we introduced a special option ”Normal agreement”.

6.4 Identification of percentage of Sister Cities patronage

The same as above, we used a ”facebook/bart-large-mnli” model from the transformer library for the zero-shot-classification task. This model yielded 17% of agreements under the Sister Cities International Patronage.

6.5 Identification of international organizations

The identification of international organizations in agreements was performed using a multi-step process that combined the Flair NER English OntoNotes-Large model for entity extraction and the Meta-Llama-3-8B-Instruct model for validation and refinement. The text of each agreement was tokenized into sentences, and the Flair model was applied to detect entities tagged as ”ORG,” producing an initial list of potential organization names. These names often included incomplete, ambiguous, or irrelevant entries.

To refine this initial output, the extracted entities were passed to the Llama model, along with a list of contracting parties previously identified in the agreements. The model was prompted to verify the validity of each organization name, correct any incomplete or ambiguous entries, and exclude entities that overlapped with the contracting parties, even when the names differed semantically, such as ”UNESCO” and ”United Nations Educational, Scientific and Cultural Organization.” Additionally, it removed irrelevant or generic terms and prioritized accurate and complete organization names.

The Llama model returned a Python list of validated organization names, which were further processed using fuzzy string matching to deduplicate similar entities. A similarity threshold of 95% was applied to filter out overlapping entities while retaining distinct ones.

The results are of mixed quality. 23% of the agreements didn't return any additional organizations besides the mentioned parties. Some of the extracted organizations are too general e.g., the United States or the Ministry of Economy (of which country?) some of them are unknown acronyms. The exact numbers are hard to estimate. Manual inspection revealed most of them seem valid but expert evaluation is required. Additional processing could be helpful and eliminate some of the errors.

6.6 Determination of the terms of validity for each agreement

To extract the term of validity from international agreements, we used a combined approach involving the Meta-Llama-3-8B-Instruct model and the Flair NER English OntoNotes-Large tagger. The term validity refers to either a specific expiration date or a relative duration (e.g., "valid for three years from the signing date"). To ensure clarity and avoid confusion between the term of validity and the signing date, we also extracted the signing date, as it is often necessary for interpreting relative terms of validity.

The process began by dividing each document into overlapping chunks, enabling the Flair NER model to detect sentences mentioning dates. Sentences containing date mentions were extracted and passed as context to the Llama model, along with a detailed prompt. The prompt instructed the model to differentiate between the signing date (e.g., "signed on January 1, 2020") and the term of validity (e.g., "expires five years after signing"). This step addressed the common issue where the model conflated these two dates.

The Llama model returned results in a structured JSON format with two fields: "signing_date" and "validity_date". The signing date was critical for interpreting relative terms like "valid for three years from signing." This method ensured consistent and accurate extraction of both the signing date and term of validity, even for complex or ambiguous phrasing.

The results are unfortunately not great. Only 23% of agreements have a signing date only 13% of agreements have a validity date and only 3% have both. Unfortunately, such a huge amount of missing data has a detrimental impact on the usefulness of the results. Bad quality of extracted texts, weak performance of Flair, or incorrect prompt are all possible causes of such results.

6.7 Identification of length of each agreement

The task was to count all the words from each agreement. Because during data preprocessing we distilled words, we just counted them without using any external model. Most agreements are about few hundred word in length. There is a small percentage of agreements that are longer, with a length of a few thousand words. 6

6.8 Determination of the conditions for extending each agreement

First we are extracting all sentences mentioned following keywords: terminate, extend, renew, withdraw, remain in effect, lengthen, prolong, revoke, stop, discontinue. With this given context we use Phi to identify agreement extension conditions, detecting whether renewals occur automatically or require explicit decisions. The result is not satisfying, as most clouses that are retrieved contains sentences

such as *"Any of the Participants may at any time, withdraw from this MoU by providing a written notice to the other Participant thirty 30 days in advance."* or *"Either Side may terminate the cooperation under this MOC after 45 days written notice to the other Side.*. These sentences focus on describing the conditions for termination rather than explicitly stating how the agreement's renewal process works. Despite this lack of clear information about the renewal mechanism, the phi model categorized it as requiring a decision to renew *By decision*.

6.9 Analysis of the frequency of recurring clauses in the agreements

Finding the most frequent clauses have been made during EDA section. We distilled the most popular words: authority, state, party, shall, information, agreement; the most popular bigrams: united state, confidential information, memorandum understanding; and trigrams: united state America, Illinois et al. Similar phrases are also common for quadgrams and quintgrams. They consist of: state California united state, California united state America, California air resource board, state California united state America, create legally binding right obligation.

6.10 Identification of the partners with whom the agreements tend to be more detailed, how text is detailed.

As we didn't find any accurate metrics for measuring detail in the formal agreements, we've merged this task with the next one, which is responsible for the evaluation of implementation for such an agreement. We assume that when the agreement provides a given specific plan for the realization of the cooperation plan, we can treat it as a detailed agreement.

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

Firstly we used Phi to identify detailed cooperation plans within agreements, specifically focusing on detecting organized activities like meetings, collaborative events, and joint initiatives and evaluating whether these planned objectives were successfully implemented. Based on feedback from the expert group indicating that our initial approach didn't align with their intended goals, we simplified our classification method. We adopted a straightforward prompt that asks: "*Does this agreement include specific provisions for evaluating its implementation? Answer ONLY with "Yes" if monitoring/evaluation processes are described, otherwise "No"*." While this approach is more straightforward, it has a limitation: we can no longer assess how the phi model approaches or reasons about the problem.

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

The Blackstone Named Entity Recognition (NER) model, pre-trained on 70,000 UK legal documents, was implemented to detect references to external legal documents. This model was selected due to its training in formal legal language similar to that found in paradiplomacy agreements. Specifically, we've implemented Blackstone's Named Entity Recognition (NER) capability, specifically utilizing its CITATION entity detection for identifying legal document references. Currently, the system detects references in a specific format containing both the date and legal document code. This method was not developed further as after discussion with professors, following prelimnary results were not good enough.

References

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. Flair: An easy-to-use framework for state-of-the-art nlp. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, pages 54–59, 2019.

[2] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. Legal-bert: The muppets straight out of law school, 2020.

[3] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[4] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: Decoding-enhanced bert with disentangled attention, 2021.

[5] Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review, 2021.

[6] Daniel Martin Katz and et al. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*, 2023.

[7] Microsoft. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

[8] Amin Sleimi et al. An automated framework for the extraction of semantic legal metadata from legal texts. *arXiv preprint arXiv:2001.11245*, 2020.

[9] Joanna Wojciechowska, Mateusz Odrowaz-Sypniewski, Maria W. Smigielska, Igor Kaminski, Emilia Wiśnios, Bartosz Pieliński, and Hanna Schreiber. Deep dive into the language of international relations: Nlp-based analysis of unesco's summary records. In *Proceedings of the 3rd Workshop on Computational Linguistics for the Political and Social Sciences*, pages 75–87. Association for Computational Linguistics, 2023.

[10] Anna Wróblewska, Bartosz Pieliński, Karolina Seweryn, Sylwia Sysko-Romańczuk, Karol Saputa, Aleksandra Wichrowska, and Hanna Schreiber. Automating the analysis of institutional design in international agreements. In *Computational Science – ICCS 2023*, pages 59–73. Springer, 2023.

A Example results

Cooperation Area	**Details**
Economic Development and Trade	Promoting trade and investment opportunities through joint marketing activities, data interchange, market studies, modernization, and improvements.
Public Health and Social Services	Collaborative approach to public health emergencies, emergency assistance capacity, integrated surveillance, early notification.
Environmental Protection and Sustainability	Conservation efforts for the Mexican wolf species.
Disaster Preparedness and Emergency Management	Coordinating efforts for wildland fire prevention, presuppression, and control.
Infrastructure Development and Urban Planning	Establishing a cooperative initiative for road safety and highway infrastructure.

Table 1: Examples of Cooperation Areas with Details

Party 1	**Party 2**
State of Alabama, United States	Jiangsu Province, People's Republic of China
State of Arkansas, United States	Dong Nai Province People's Committee, Vietnam
Ministry of Energy of the United Mexican States, Mexico	State of California, United States
State of Delaware, USA	Miyagi Prefecture, Japan
State	Members

Table 2: Examples of Agreement Parties

Name of the file	Entities
1Undated.pdf	Northwest, Pacific Northwest Border Health Alliance, State of Washington Department of Health, Province of Saskatchewan Ministry of Health, Province of British Columbia Ministry of Health Services, Province of Alberta Ministry of Health and Wellness, State of Oregon Department of Human Services, State of Alaska Department of Health and Social Services, State of Montana Department of Public Health and Human Services, State of Idaho Department of Health and Welfare, Yukon Department of Health and Social Services
Indiana_7.pdf	State of Karnataka Government, India, Indiana State Government, United States of America
1June29.pdf	State of Arkansas, Socialist Republic of Vietnam, Dong Nai Province People's Committee, United States of America

Table 3: Examples of other entities involved in the agreement

Signing Date	Validity Date
2019-06-26	Valid for 5 years from signing date
2001-06-01	Not specified
Not specified	Not specified
Not specified	Valid for five years after signing
Not specified	Commitment to reduce tropical deforestation and protect global climate system until at least 2020 (specifically mentioned)

Table 4: Examples of Signing and Validity Dates extracted from agreements

International Organizations
United Nations, Global Island Partnership (GLISPA), International Green Island Forum (IGIF)
FINMA, Allianz Insurance Group
Partnership
Ohio, CIPA, TDB, MOFCOM
International Council on Clean Transportation, ZEV Alliance

Table 5: Examples of International Organizations extracted from agreements

Signing Date	Validity Date
2019-06-26	Valid for 5 years from signing date
2001-06-01	Not specified
Not specified	Not specified
Not specified	Valid for five years after signing
Not specified	Commitment to reduce tropical deforestation and protect global climate system until at least 2020 (specifically mentioned)

Table 6: Examples of Signing and Validity Dates extracted from agreements

B LLM Prompts

```
"You are an AI assistant trained to classify text "
"from international agreements into predefined "
"cooperation areas. Your task is to analyze a "
"provided document chunk and identify which of the "
"following predefined areas it relates to:\n"
"['Economic Development and Trade', "
"'Education and Academic Exchange', "
"'Cultural Exchange', 'Tourism Promotion', "
"'Environmental Protection and Sustainability', "
"'Infrastructure Development and Urban Planning', "
"'Public Health and Social Services', "
"'Technology and Innovation', "
"'Disaster Preparedness and Emergency Management', "
"'Good Governance and Administrative Cooperation', "
"'Agriculture'].\n\n Instructions:\n"
"1. Read the document chunk carefully and classify "
"it into one or more of the predefined areas, but "
"only include areas that are explicitly relevant.\n"
"2. If a text refers to a specific project, "
"activity, or concept, ensure it is "
"classified into the correct area(s).\n"
"3. For each selected area, provide a short "
"description of why it was chosen, "
"adding this in parentheses next to the area name. "
"For example: ['Economic Development and Trade "
"(Promoting trade and investment opportunities)', "
"'Education and Academic Exchange "
"(Establishing student exchange programs)', "
"'Environmental Protection and Sustainability "
"(Sharing expertise in waste "
"management and pollution control)'].\n"
"4. Avoid over-classifying. Only choose areas that "
"clearly align with the content of the text.\n"
"5. If no areas apply, return an empty list [] "
"without explanation.\n\n Document Chunk:\n"
f"{text_chunk}\n\n Output:\n"
"Return a Python list of the relevant areas with "
"descriptions in parentheses as explained. "
"If no areas apply, return an empty list []"
```

Listing 1: LLama prompt for identifying cooperation areas

```
"You are an AI assistant trained to assist law experts in
↪  extracting contracting parties "
"from international agreements. Your task is to extract the
↪  full names of the contracting "
"parties, ensuring compound names are treated as single
↪  entities. Pay attention to political "
"divisions, institutions, and geographic qualifiers (e.g.,
↪  'Jiangsu Province, Republic of China'). "
"Recognize that institutions (e.g., 'Bank of England')
↪  should not be confused with geographic or "
"sovereign entities (e.g., 'United Kingdom'). For entities
↪  from outside the United States, add "
"the name of the country they are from after a comma.
↪  Analyze phrases like 'This agreement is "
"between,' 'entered into by,' or 'signed by' to identify the
↪  involved parties. Always treat each "
"party as a distinct and complete entity. Output strictly as
↪  a Python list of party names, e.g., "
"['Party 1', 'Party 2', 'Party 3']. If no relevant details
↪  are found, return strictly an empty "
"list []. Example Input: 'This agreement is between the
↪  State of Alabama, United States and the "
"Province of Ninh Thuan.' Example Output: ['State of
↪  Alabama, United States', 'Province of Ninh "
"Thuan, Vietnam']. Example Input: 'This contract is signed
↪  by the Bank of England and the "
"European Investment Bank.' Example Output: ['Bank of
↪  England, United Kingdom', 'European "
"Investment Bank, European Union']. Example Input: 'This
↪  agreement involves Free and Sovereign "
"State of Nuevo Len and the State of California, United
↪  States.' Example Output: ['State of "
"Nuevo Len, Mexico', 'State of California, United States'].
↪  Under no circumstances should you "
"output any other format than the one specified above.
↪  Agreement Chunk:\n"
f"{text_chunk}"
```

Listing 2: LLama prompt for extracting contracting parties from the agreements.

```
"You are an AI assistant specializing in extracting and
↪  verifying valid organization and "
"institution names from text. Your task is to review a list
↪  of potential organization names "
"extracted using an automated method. These names may be
↪  broken, incomplete, invalid, or "
"irrelevant. Your job is to:\n"
"1. Verify the validity of each organization name and
↪  correct incomplete or broken names using "
"your internal knowledge.\n"
"2. Remove irrelevant, invalid, or generic terms that are
↪  not real organizations.\n"
"3. Compare the extracted entities with the provided list of
↪  previously identified parties, and "
"exclude any entities that refer to the same organization
↪  even if their names are not an exact "
"match. Use semantic similarity and common abbreviations to
↪  identify matches (e.g., 'Government "
"of the State of California' and 'Government of California'
↪  or 'UNESCO' and 'United Nations "
"Educational, Scientific and Cultural Organization').\n"
"4. Prioritize providing full and accurate names of
↪  organizations where possible. If a name is "
"ambiguous or incomplete, use your knowledge to infer the
↪  correct full name.\n"
"5. Return a final Python list of verified and corrected
↪  full organization names, ensuring no "
"duplicates.\n\n"
"Details to process:\n"
f"Extracted Entities: {entities}\n"
f"Previously Identified Parties to Ignore:
↪  {excluded_parties}\n\n"
"Carefully review the extracted entities and ensure all
↪  names are valid, accurate, and complete. "
"Exclude any entity that matches or overlaps with those in
↪  the 'Previously Identified Parties to "
"Ignore' section, even if their names differ slightly. Use
↪  your internal knowledge and reasoning "
"to identify and exclude duplicates. Return only the final
↪  Python list of verified organization "
"names, with no additional text or explanation."
```

Listing 3: LLama prompt for extracting organization names extracted from text.

```
"You are an AI assistant trained to extract critical
↪  information about international agreements. "
"Your task is to identify the signing date and the
↪  expiration or validity date of the agreement. "
"The signing date refers to the date when the agreement was
↪  signed. The expiration or validity date "
"refers to the specific date or duration until which the
↪  agreement remains in effect.\n\n"
"Details to extract:\n"
"- Signing Date: Look for phrases such as 'signed on [date]'
↪  or similar references. If no signing date "
"is mentioned, return 'Not specified'.\n"
"- Validity Date: Identify the specific date or relative
↪  duration (e.g., 'valid for five years from the signing
↪  date' "
"or 'expires in December 2025'). If no validity date is
↪  mentioned, return 'Not specified'.\n\n"
"Return the extracted information strictly in the following
↪  JSON format:\n"
"{\n"
"  \"signing_date\": \"[DD-MM-YYYY]\" or \"Not
↪  specified\",\n"
"  \"validity_date\": \"[DD-MM-YYYY]\", \"[Relative term]\",
↪  or \"Not specified\"\n"
"}\n\n"
"Examples:\n"
"1. Input: 'This agreement was signed on March 1, 2000 and
↪  is valid until March 1, 2005.'\n"
"   Output: {\n"
"     \"signing_date\": \"01-03-2000\",\n"
"     \"validity_date\": \"01-03-2005\"\n"
"   }\n\n"
"2. Input: 'The agreement will expire five years after
↪  signing but does not specify the signing date.'\n"
"   Output: {\n"
"     \"signing_date\": \"Not specified\",\n"
"     \"validity_date\": \"Valid for five years after
↪  signing\"\n"
"   }\n\n"
"3. Input: 'This agreement has no specific expiration date
↪  and was signed on January 15, 2010.'\n"
"   Output: {\n"
"     \"signing_date\": \"15-01-2010\",\n"
"     \"validity_date\": \"Not specified\"\n"
"   }\n\n"
f"Possible date mentions: {date_sentences}"
```

Listing 4: Prompt for extracting signing and validity dates from the agreements.

```
"You will be delivered with the agreement between
"United States and other foreign state or city."
"Determine if the extension of agreement will be automatic
↪  renewed without further decision-making or require
↪  active approval from the parties involved?"
"If any steps for approval are describied within document,
↪  please extract them. Do not infer or recommend any
↪  actions. Please write it in concise form."
"Agreement text:"
```

Listing 5: Prompt for checking if agreement was prolonged by decision or automatic.

```
"""<|user|>
Analyze this contractual clause regarding agreement
↪  duration:
Does renewal happen AUTOMATICALLY or REQUIRE SPECIFIC
↪  APPROVAL?
Answer strictly as:
"Extension: [Automatic/By decision]
Reason: [2-5 word reason from text]"

Context:
{context}
<|end|>
<|assistant|>"""
```

Listing 6: 2nd Prompt for checking if agreement was prolonged by decision or automatic.