DL4DS Legal Contract Dataset

Ethan Chang, Heng Chang, Josh Yip

March 2nd, 2025

Abstract

We aim to create a labeled dataset of contract clauses that will serve as a foundation for training and evaluating AI models in legal contract NLP. By curating and annotating contracts with a focus on clause categorization and summarization, our dataset will contribute to the development of AI-driven tools for contract analysis and verification.

Introduction

Understanding and analyzing legal contracts remains a complex task requiring specialized knowledge. With the increasing demand for AI-driven legal tech solutions, high-quality labeled datasets are crucial for training models that can accurately categorize and summarize contract clauses. Our project focuses on creating a dataset specifically designed for this purpose, providing a valuable resource for researchers and practitioners in legal NLP.

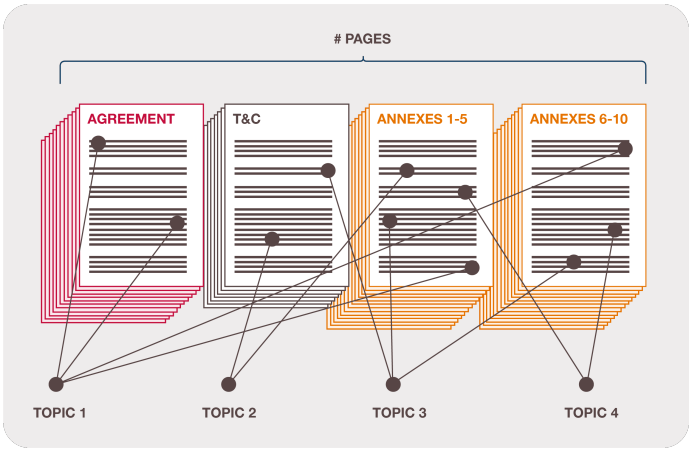

Figure 1: Contract document mapping visualization.

Related Work

Existing datasets such as the CUAD dataset by the Atticus Project and the Stanford ContractNLI dataset have paved the way for legal contract analysis in NLP. However, many existing datasets focus on broad contract types, leaving gaps in specific clause identification and domain-specific labeling. Our dataset will build upon these works while addressing these gaps, ensuring more comprehensive contract clause categorization.

Proposed Work

We will develop a structured pipeline for contract clause labeling, leveraging a combination of:

- Existing contract datasets (e.g., CUAD, ContractNLI)
- Manual annotation by legal experts
- Automated preprocessing and classification methods

This dataset will support model training for clause classification, summarization, and quality evaluation of contracts. The dataset creation process will be documented to ensure reproducibility and scalability.

Datasets

Our dataset will integrate data from:

- CUAD dataset (labeled contract clauses for NLP applications)
- Stanford ContractNLI dataset (legal contract entailment tasks)
- Other publicly available legal contract corpora
- Manually labeled contracts for our specific use case (to be defined)

We will ensure diversity in contract types and ensure annotations align with real-world contract analysis needs.

Evaluation

We will evaluate our dataset and labeling methodology based on:

- Inter-annotator agreement scores for manual labels
- Benchmarking against existing datasets in clause classification tasks

- Model performance on clause categorization and summarization tasks using our dataset

Success will be determined by achieving high consistency in annotations and improved model performance compared to existing models on existing and currated datasets.

Timeline

- Week 1-2: Research best practices in contract dataset creation, including fairness metrics, contract usefulness, and aggregation methods.
- Week 3-4: Investigate clause polarity (evaluating if a clause is beneficial or detrimental) and define quantifiable measures for fairness.
- Week 5-6: Explore classification approaches for contract clauses and analyze potential legal loopholes in existing datasets.
- Week 7-8: Collect and preprocess contract samples, ensuring a balanced representation across different contract types.
- Week 9-10: Perform k-means-based semi-automated labeling, clustering contract clauses based on semantic similarities to predefined categories. Also explore dynamic categorization with k-means.
- Week 11: Validate dataset quality through inter-annotator agreement and benchmark dataset performance using AI models.
- Week 12: Finalize dataset, prepare documentation, and determine dataset effectiveness with domain experts.

Conclusion

Our project aims to fill a crucial gap in legal NLP by creating a high-quality labeled dataset of contract clauses. By ensuring accurate categorization and annotation, we will contribute to the advancement of AI-driven contract analysis tools, ultimately fostering better legal transparency and compliance.

References

[1] The Atticus Project. *CUAD: Contract Understanding Atticus Dataset*. 2021.

[2] Stanford NLP Group. *ContractNLI: A Benchmark for Legal Contract Entailment*. 2022.

[3] WorldCC Foundation. *Contract Design Pattern Library*.