Evaluating the impact of incorporating 'legalese' definitions and abstractive summarization on the categorization of legal cases

Daniela Cortes Bermudez, Shiu Tin Ivan Ko, Huiyun Zhang, Henry Han

School of Computer Science and Engineering
Laboratory of Data Science and AI Innovations
Baylor University
Waco, Texas 76798, USA

E-mail: Henry_Han@baylor.edu

July 24, 2023

Overview

- Introduction
  - Terminology
- Data
- Methods
- Process Framework
- Metrics
- Results
  - Interpretation
  - Discussion
- Conclusions

Introduction

"Legalese"

Introduction

"Legalese"

- This domain-specific terminology is characterized by lengthy, wordy, and complex sentence structure

Introduction

"Legalese"

- This domain-specific terminology is characterized by lengthy, wordy, and complex sentence structure

Plain English

Introduction

"Legalese"

- This domain-specific terminology is characterized by lengthy, wordy, and complex sentence structure

Plain English

- Clear, straightforward, and concise language that avoids inflated vocabulary and complex sentence structure

Introduction

"Legalese"

- This domain-specific terminology is characterized by lengthy, wordy, and complex sentence structure

Plain English

- Clear, straightforward, and concise language that avoids inflated vocabulary and complex sentence structure

Case Holding

Introduction

"Legalese"

- This domain-specific terminology is characterized by lengthy, wordy, and complex sentence structure

Plain English

- Clear, straightforward, and concise language that avoids inflated vocabulary and complex sentence structure

Case Holding

- Final decision the court reached on a case

Introduction

- We are interested in comparing text classification models exploring whether the classification of a case statement to its holding is affected by:
  - Data Processing
  - Model Stacking
- Through exploring this, we want to see if the change in the data affects information fidelity
  - To assess information fidelity, we ask:
    - "Does model stacking affect classification performance?"
    - "Does performance change with pretraining?"

Data

- **BillSum**

Data

- **BillSum**
  - Benchmark dataset for legal document summarization

Data

- **BillSum**
  - Benchmark dataset for legal document summarization
  - Contains 22,218 reference summaries from both US Congressional and California State bills

Data

- **BillSum**
  - Benchmark dataset for legal document summarization
  - Contains 22,218 reference summaries from both US Congressional and California State bills
- **"Legalese" Glossary**

Data

- **BillSum**
  - Benchmark dataset for legal document summarization
  - Contains 22,218 reference summaries from both US Congressional and California State bills
- **"Legalese" Glossary**
  - We compiled a legal glossary of terms and definitions from the United States Courts' Glossary of Legal Terms

Data

- **BillSum**
  - Benchmark dataset for legal document summarization
  - Contains 22,218 reference summaries from both US Congressional and California State bills
- **"Legalese" Glossary**
  - We compiled a legal glossary of terms and definitions from the United States Courts' Glossary of Legal Terms
    - Only one definition per observation

Data

- **BillSum**
  - Benchmark dataset for legal document summarization
  - Contains 22,218 reference summaries from both US Congressional and California State bills
- **"Legalese" Glossary**
  - We compiled a legal glossary of terms and definitions from the United States Courts' Glossary of Legal Terms
    - Only one definition per observation
    - Latin terms were translated

Data

- **BillSum**
  - Benchmark dataset for legal document summarization
  - Contains 22,218 reference summaries from both US Congressional and California State bills
- **”Legalese” Glossary**
  - We compiled a legal glossary of terms and definitions from the United States Courts’ Glossary of Legal Terms
    - Only one definition per observation
    - Latin terms were translated
    - If a term had two or more definitions, all were included

Data

- **BillSum**
  - Benchmark dataset for legal document summarization
  - Contains 22,218 reference summaries from both US Congressional and California State bills
- **"Legalese" Glossary**
  - We compiled a legal glossary of terms and definitions from the United States Courts' Glossary of Legal Terms
    - Only one definition per observation
    - Latin terms were translated
    - If a term had two or more definitions, all were included
- **CaseHOLD**

Data

- **BillSum**
  - Benchmark dataset for legal document summarization
  - Contains 22,218 reference summaries from both US Congressional and California State bills
- **"Legalese" Glossary**
  - We compiled a legal glossary of terms and definitions from the United States Courts' Glossary of Legal Terms
    - Only one definition per observation
    - Latin terms were translated
    - If a term had two or more definitions, all were included
- **CaseHOLD**
  - Benchmark dataset with over 53,000 multiple-choice questions

Data

- **BillSum**
  - Benchmark dataset for legal document summarization
  - Contains 22,218 reference summaries from both US Congressional and California State bills
- **"Legalese" Glossary**
  - We compiled a legal glossary of terms and definitions from the United States Courts' Glossary of Legal Terms
    - Only one definition per observation
    - Latin terms were translated
    - If a term had two or more definitions, all were included
- **CaseHOLD**
  - Benchmark dataset with over 53,000 multiple-choice questions
    - Each observation consists of a cited case and five case holding options

Methodology

- **Abstractive Summarization**

Methodology

- **Abstractive Summarization**
  - Used Google's FLAN-T5 Base Model

Methodology

- **Abstractive Summarization**
  - Used Google's FLAN-T5 Base Model
    - This model was trained for summarization and fined-tuned for legal data using the BillSum Corpus

Methodology

- **Abstractive Summarization**
  - Used Google's FLAN-T5 Base Model
    - This model was trained for summarization and fined-tuned for legal data using the BillSum Corpus

- **"Legalese" Definition Mapping**

Methodology

- **Abstractive Summarization**
  - Used Google's FLAN-T5 Base Model
    - This model was trained for summarization and fined-tuned for legal data using the BillSum Corpus
- **"Legalese" Definition Mapping**
  - The mapping returns a modified observation, where each term matching a term in the "legalese" glossary is replaced with its definition

Methodology

- **Abstractive Summarization**
  - Used Google's FLAN-T5 Base Model
    - This model was trained for summarization and fined-tuned for legal data using the BillSum Corpus
- **"Legalese" Definition Mapping**
  - The mapping returns a modified observation, where each term matching a term in the "legalese" glossary is replaced with its definition
- **Case-Holding Classification**

Methodology

- **Abstractive Summarization**
  - Used Google's FLAN-T5 Base Model
    - This model was trained for summarization and fined-tuned for legal data using the BillSum Corpus
- **"Legalese" Definition Mapping**
  - The mapping returns a modified observation, where each term matching a term in the "legalese" glossary is replaced with its definition
- **Case-Holding Classification**
  - We did our baseline classification with our testing dataset on three base models: BERT, LegalBERT, and GPT2

Methodology

- **Abstractive Summarization**
  - Used Google's FLAN-T5 Base Model
    - This model was trained for summarization and fined-tuned for legal data using the BillSum Corpus
- **"Legalese" Definition Mapping**
  - The mapping returns a modified observation, where each term matching a term in the "legalese" glossary is replaced with its definition
- **Case-Holding Classification**
  - We did our baseline classification with our testing dataset on three base models: BERT, LegalBERT, and GPT2
    - Testing dataset is 5,000 observations from the CaseHOLD Dataset

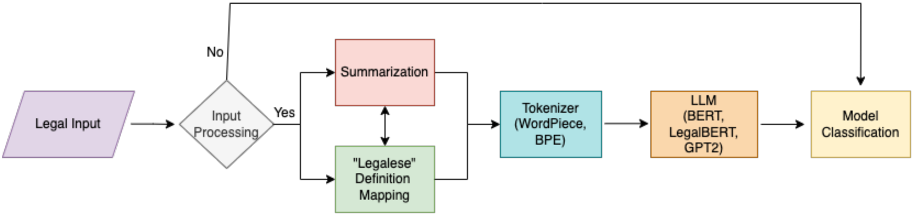
Methodology

- **Abstractive Summarization**
  - Used Google's FLAN-T5 Base Model
    - This model was trained for summarization and fined-tuned for legal data using the BillSum Corpus
- **"Legalese" Definition Mapping**
  - The mapping returns a modified observation, where each term matching a term in the "legalese" glossary is replaced with its definition
- **Case-Holding Classification**
  - We did our baseline classification with our testing dataset on three base models: BERT, LegalBERT, and GPT2
    - Testing dataset is 5,000 observations from the CaseHOLD Dataset
  - We classified two versions of each model: Base and "Gen1"

Methodology

- **Abstractive Summarization**
  - Used Google's FLAN-T5 Base Model
    - This model was trained for summarization and fined-tuned for legal data using the BillSum Corpus
- **"Legalese" Definition Mapping**
  - The mapping returns a modified observation, where each term matching a term in the "legalese" glossary is replaced with its definition
- **Case-Holding Classification**
  - We did our baseline classification with our testing dataset on three base models: BERT, LegalBERT, and GPT2
    - Testing dataset is 5,000 observations from the CaseHOLD Dataset
  - We classified two versions of each model: Base and "Gen1"
    - The **"Gen1"** model is our model after pretraining with 60,000 observations

Process Framework



- A total of 30 trials were run
  - $3\, models * 2\, generations * 5\, inputs$
- Each trial went through a 5-fold-cross validation
- **Tokenizers**
  - GPT2 uses Byte-Pair Encoding (BPE)
  - BERT-based models use WordPiece

Metrics

D-Index

- The diagnostic index is a novel machine-learning evaluation method that detects small performance differences between models and provides a comprehensive evaluation by taking into account data imbalance
- **Scoring range:** $1.1699 \rightarrow 2$

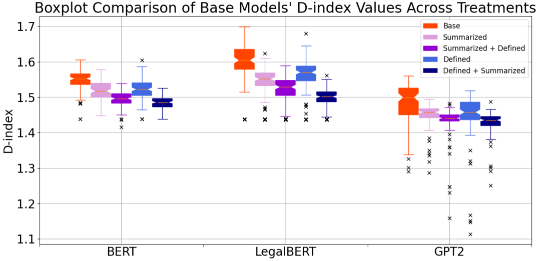
- Three metrics were chosen for the evaluation:
  - Accuracy
  - F1-Score
  - D-Index

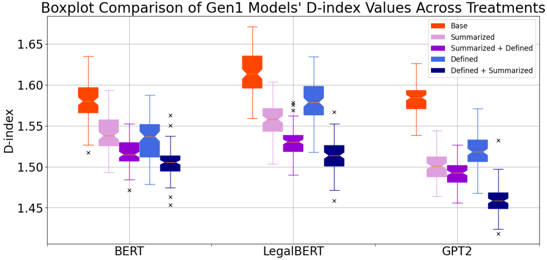
Results

Table 1. Results of the accuracy, F1-score, and D-index metrics for the base BERT, LegalBERT, and GPT2 models for each processed dataset, both for pre-trained and not pretrained models. Gen1 = Pretrained model; D = Defined; S = Summarized; S + D = Summarized + Defined; D + S = Defined + Summarized

		BERT		Legal BERT		GPT2	
		Base	Gen1	Base	Gen1	Base	Gen1
Acc	Original	**0.809 ± 0.011**	**0.824 ± 0.008**	**0.836 ± 0.014**	**0.838 ± 0.007**	**0.782 ± 0.035**	**0.818 ± 0.008**
	D	0.801 ± 0.014	0.812 ± 0.011	0.824 ± 0.009	0.829 ± 0.009	0.754 ± 0.066	0.803 ± 0.011
	D + S	0.799 ± 0.013	0.803 ± 0.010	0.807 ± 0.008	0.807 ± 0.012	0.763 ± 0.044	0.785 ± 0.014
	S	0.805 ± 0.010	0.810 ± 0.011	0.820 ± 0.009	0.819 ± 0.010	0.769 ± 0.036	0.793 ± 0.010
	S + D	0.800 ± 0.015	0.807 ± 0.009	0.814 ± 0.009	0.814 ± 0.006	0.760 ± 0.053	0.793 ± 0.010
F1	Original	**0.374 ± 0.084**	**0.424 ± 0.045**	**0.440 ± 0.134**	**0.481 ± 0.044**	**0.264 ± 0.141**	**0.440 ± 0.031**
	D	0.323 ± 0.091	0.336 ± 0.055	0.369 ± 0.124	0.418 ± 0.051	0.242 ± 0.107	0.321 ± 0.054
	D + S	0.225 ± 0.074	0.280 ± 0.049	0.227 ± 0.092	0.292 ± 0.046	0.161 ± 0.094	0.198 ± 0.055
	S	0.310 ± 0.069	0.357 ± 0.048	0.340 ± 0.112	0.379 ± 0.041	0.208 ± 0.095	0.293 ± 0.045
	S + D	0.269 ± 0.067	0.304 ± 0.044	0.278 ± 0.122	0.322 ± 0.044	0.182 ± 0.093	0.269 ± 0.048
D-Index	Original	**1.550 ± 0.030**	**1.580 ± 0.023**	**1.601 ± 0.058**	**1.616 ± 0.025**	**1.487 ± 0.053**	**1.583 ± 0.017**
	D	1.522 ± 0.030	1.534 ± 0.025	1.561 ± 0.046	1.581 ± 0.026	1.444 ± 0.073	1.520 ± 0.024
	D + S	1.484 ± 0.019	1.506 ± 0.017	1.495 ± 0.025	1.513 ± 0.020	1.427 ± 0.036	1.459 ± 0.017
	S	1.518 ± 0.029	1.518 ± 0.029	1.546 ± 0.040	1.556 ± 0.021	1.449 ± 0.035	1.500 ± 0.017
	S + D	1.496 ± 0.023	1.517 ± 0.017	1.519 ± 0.039	1.530 ± 0.017	1.430 ± 0.049	1.491 ± 0.014

Results


Boxplot Comparison of Base Models' D-index Values Across Treatments
D-index
1.7
1.6
1.5
1.4
1.3
1.2
1.1
BERT
LegalBERT
GPT2
Base
Summarized
Summarized + Defined
Defined
Defined + Summarized


Boxplot Comparison of Gen1 Models' D-index Values Across Treatments
D-index
1.65
1.60
1.55
1.50
1.45
BERT
LegalBERT
GPT2
Base
Summarized
Summarized + Defined
Defined
Defined + Summarized

Conclusions

- Classification on the original dataset outperforms all other inputs

Conclusions

- Classification on the original dataset outperforms all other inputs
- The Gen1 model results show that pretraining results in a performance boost on all inputs

Conclusions

- Classification on the original dataset outperforms all other inputs
- The Gen1 model results show that pretraining results in a performance boost on all inputs
  - However, model pretraining is expensive both environmentally and financially for the legal domain

Conclusions

- Classification on the original dataset outperforms all other inputs
- The Gen1 model results show that pretraining results in a performance boost on all inputs
  - However, model pretraining is expensive both environmentally and financially for the legal domain
- The "legalese" definition mapping is limited by the amount of data available and the lack of benchmark datasets to do so

Conclusions

- Classification on the original dataset outperforms all other inputs
- The Gen1 model results show that pretraining results in a performance boost on all inputs
  - However, model pretraining is expensive both environmentally and financially for the legal domain
- The "legalese" definition mapping is limited by the amount of data available and the lack of benchmark datasets to do so
- There is very limited legal summarization data available

Conclusions

- Classification on the original dataset outperforms all other inputs
- The Gen1 model results show that pretraining results in a performance boost on all inputs
  - However, model pretraining is expensive both environmentally and financially for the legal domain
- The "legalese" definition mapping is limited by the amount of data available and the lack of benchmark datasets to do so
- There is very limited legal summarization data available
  - Creating summaries is costly and time-consuming, restricting the capabilities of summarization model fine-tuning

Conclusions

- Classification on the original dataset outperforms all other inputs
- The Gen1 model results show that pretraining results in a performance boost on all inputs
  - However, model pretraining is expensive both environmentally and financially for the legal domain
- The "legalese" definition mapping is limited by the amount of data available and the lack of benchmark datasets to do so
- There is very limited legal summarization data available
  - Creating summaries is costly and time-consuming, restricting the capabilities of summarization model fine-tuning
- Order of data processing methods affects the performance