Using Natural Language Processing to Identify Unfair Clauses in Terms and Conditions Documents

Jonathan Sears
Tulane University / New Orleans, LA
jsears1@tuane.edu
Nicholas Radwin
Tulane / New Orleans, LA
nradwin@tulane.edu

Problem Overview

No one reads Terms and Conditions, but everyone agrees to them. Every online user is faced with these incredibly dense and complex documents as if they are expected to read them in their entirety before using a digital tool or app that they want to use. In reality, we just agree to the terms and conditions to use what's behind them, leaving the user in the dark from the ambiguous, unfair, and even potentially exploitative clauses that may not be in the user's best interest.

Our project is focused on bringing these buried, questionable clauses to light so that the user can understand what he is agreeing to at a glance. As we experiment with a variety of learning models, we are building a summarization tool that takes in terms and conditions as input and generates a page that bullets a list of either ambiguous, unfair, or exploitative clauses with a confidence score for how likely the model chose those sentences correctly. We also want the tool to highlight the sentences from the original text.

Data

Our pretrained model is trained on a data repository called PileOfLaw (link 1 below), which is full of 256GB of legal and administrative documents in the English language. Within PileOfLaw, there are two pretrained BERT models which differ only in seed. We decided to use the second model (link 2 below), as it is the same seed used in the original paper that used PileOfLaw; that way we can compare results to get better insight into how well any modifications we make are performing. https://huggingface.co/datasets/pile-of-law/pile-of-law https://huggingface.co/pile-of-law/legalbert-large-1.7M-2 To fine tune this model, we are using a dataset of 100 annotated terms of service documents. These documents come from the CLAUDETTE research paper, and are annotated with 5 different types of potential unfairness (see references), but for now, we are just going to combine all of these sub-labels into one label of fair or potentially unfair to simplify things. We may decide to revisit this in the future

Methods

What method or algorithm are you using? We have decided to use a transfer learning approach, by using a pretrained BERT model trained on the pile of law dataset as the head of our model, and then we will pass the pretrained word embeddings through a model of our own design to classify them as potentially fair or unfair. We feel that the model being pretrained on the pile of law dataset is extremely beneficial for us, as it will not only have a good baseline understanding of the english language, but it will also understand legal terms, which is almost an entirely different language itself. We will experiment with different kinds of neural network architectures using the BERT model embeddings to get the best possible results. As of this milestone, we are simply passing our pretrained embeddings through multiple connected linear layers, but that approach has led to some problems as you will see below. Since our problem is a binary classification problem, we will evaluate our results using accuracy and f1 scores, using an 80-10-10 train, test, validation split. Are you using an existing library to do so? Did you introduce any new variations to these methods? How will you evaluate the results? Which baselines will you compare against?

Preliminary Experiments and Results

State and evaluate your results up to the milestone.

Our current results are not great. Although we managed to get a working model, the model currently will only predict sentences to be lebeled "fair", which gives us a relatively high accuracy, as a majority of the sentenes in the training data are labeled fair, however it gives us an f1 score of 0, making our model essential useless in it's current form.

Related Work

Citations within the text appear in parentheses as or, if the author's name appears in the text itself, as .

1. CLAUDETTE https://arxiv.org/pdf/1805.01217v2.pdf

CLAUDETTE is an automated detector of potentially unfair clauses in terms and conditions, basically exactly what we are trying to build. The data they used was 50 annotated European terms of service contracts. It is worth noting they used European contracts, as different countries mean different laws, and different language used in the contracts. They used a wide variety of different ML approaches to this problem including, support vector machines, hidden markov models, LSTMs, CNNs, and different combinations of them, getting the best results (measured in F1-score) via the combined model.

2. Detecting and explaining unfairness in consumer contracts through memory networks https://link.springer.com/article/10.1007/s10506-021-09288-2

In this study, the researchers define unfair legal language in terms and conditions as language that causes "a significant imbalance in the parties' rights and obligations, to the detriment of the consumer, are deemed unfair by Consumer Law."

They note how pervasive this point out that despite substantive law in place, and despite the competence of enforcers, providers of online services still tend to use unfair and unlawful clauses in these documents."

Their system is trained on 100 labeled ToS documents and is based on splitting unfair clauses into 5 major label categories:

(i) liability exclusions and limitations

(ii) the provider's right to unilaterally remove consumer content from the service, including in-app purchases

(iii) the provider's right to unilaterally terminate the contract

(iv) the provider's right to unilaterally modify the contract and/or the service

(v) arbitration on disputes arising from the contract

The unfair language/rationale pairing output was most accurate when the researchers used a memory-augmented neural network in combination with "strong supervision," meaning that they fed the expertly written rationales for specific unfair labels into their model.

3. LexGLUE: A Benchmark Dataset for Legal Language Understanding in English chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://aclanthology.org/2022.acl-long.297.pdf

The LexGLUE paper was written to create a benchmark for legal language understanding. It is a compilation of datasets, pretrained models and different tests and evaluation metrics performed on those datasets and models for different tasks. We think it will be useful as both a potential source and potential benchmark for our model.

4. Automatic semantics extraction in law documents https://dl.acm.org/doi/10.1145/1165485.1165506

This study aimed to classify arguments from legislative texts using the multiclass support vector machine (MSVM) algorithm.

Stemming (representing the morphology/roots of words) improved accuracy and feature selection (restricted vocabulary) provided a work around for possible overfitting.

Overall, the study showcases MSVM's effectiveness in extracting what arguments are actually being made in legal documents, and helps us with coming up with a method for efficient legal document analysis.

5. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset https://arxiv.org/abs/2104.08671

This paper presents an argument for domain specific pretraining and tests it's hypothesis on a dataset of their own creation: CaseHOLD. They compare the performance of wide domain BERT models trained separately on a corpus of legal documents (smaller, specific domain) and another model trained on wikipedia and google books (wider domain). They found that pretraining on the smaller corpus of legal documents was more effective for legal classification tasks despite their being much less data available. This leads us to believe that if we decide to use a pretrained model, we will likely use one that is domain specific to legal documents.

Since our problem is a classification problem, we think an 80-20 cross validation approach with metrics like our accuracy, precision, recall, and F1 score should suffice. However, the scale in which we perform these evaluation metrics may vary, as we may want to measure them on a document level, and on a sentence by sentence level. We can use pretrained models, as well as a simple logistic regression and naïve Bayes models as baseline metrics to compare our model too.

Division of Labor

We will continue to divide the work as equally as possible as we experiment with our model trying to get it to perform well in its classification of unfair language. One thing we plan on dividing is different model approaches; for example, Jonathan might test how an LSTM performs on this task, while Nick tests an RNN. In the end we will highlight our best performing model and focus our report on its inner workings and results.

Timeline

Moving forward, we will work together on the project at least twice a week. We live together so we should be able to find time most days to work on it to finish by the due date–April 30, 2024.

@mischendersonkrass2022pileoflaw, url = https://arxiv.org/abs/2207.00220, author = Henderson, Peter and Krass, Mark S. and Zheng, Lucia and Guha, Neel and Manning, Christopher D. and Jurafsky, Dan and Ho, Daniel E., title = Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset, publisher = arXiv, year = 2022 @inproceedingsdrawzeski-etal-2021-corpus, address = Punta Cana, Dominican Republic, author = Drawzeski, Kasper and Galassi, Andrea and Jablonowska, Agnieszka and Lagioia, Francesca and Lippi, Marco and Micklitz, Hans Wolfgang and Sartor, Giovanni and Tagiuri, Giacomo and Torroni, Paolo, booktitle = Proceedings of the Natural Legal Language Processing Workshop 2021, doi = 10.18653/v1/2021.nllp-1.1, month = nov, pages = 1–8, publisher = Association for Computational Linguistics, title = A Corpus for Multilingual Analysis of Online Terms of Service, url = https://aclanthology.org/2021.nllp-1.1, year = 2021 @mischendersonkrass2022pileoflaw, url = https://arxiv.org/abs/2207.00220, author = Henderson, Peter and Krass, Mark S. and Zheng, Lucia and Guha, Neel and Manning, Christopher D. and Jurafsky, Dan and Ho, Daniel

E., title = Pile of Law: Learning Responsible Data Filtering from the Law and a 256GB Open-Source Legal Dataset, publisher = arXiv, year = 2022