Lost in Legalese: NLP for Privacy Risk Detection

Stanford CS224N Custom Project

Ray Hu
Department of Computer Science
Stanford University
rayhu@stanford.edu

Benjamin Ward
Institute for Computational and Mathematical Engineering
Stanford University
wardb@stanford.edu

Basant Khalil
Department of Computer Science
Stanford University
bkhalil@stanford.edu

1 Project Information

- **Project type**: custom project.
- **Mentor**: Jing Huang (hij@stanford.edu) will be the project mentor.

2 Research paper summary

Title	NLP for Consumer Protection: Battling Illegal Clauses in German Terms and Conditions in Online Shopping
Venue	Proceedings of the 1st Workshop on NLP for Positive Impact
Year	2021
URL	https://aclanthology.org/2021.nlp4posimpact-1.10/

Table 1: Presentation of Related Research Paper [1]

Background. In 2014, global retail e-commerce sales were approximately 1.3 trillion dollars. In 2023, this had grown to 5.8 trillion dollars, and is projected to surpass 8 trillion by 2027. Whenever consumers place an order, they often agree to Terms and Conditions (T&C) statements. These are complex legal documents, which often contain unfavourable, and sometimes legally unenforceable clauses. This paper aims to show how NLP can be used to help protect consumers from inadvertedly signing potentially adverse legal documents. More specifically, this paper strives to identify "void clauses" online shopping consumer contracts, that is clauses that are legally invalid under German consumer protection laws.

Summary of contributions. This paper developed an NLP-based application that legally assesses clauses in German T&C from German online shops under EU jurisdiction. More specifically, they showed that by finetuning a pretrained German BERT model, they were able to achieve 90% accuracy (and 90% precision and recall) in the detection of void clauses. The first contribution of this work is the gathering of a dataset consisting of 1,186 legal clauses. The data consists of 24 entire T&C documents from German online shops (totalling 968 clauses), and 218 individual clauses (2/3 of which were specifically selected by legal experts as specifically relevant to their everyday work). The majority of the clauses were legally labeled as "void" or "valid" by two experts, with the remaining clauses labeled "void" after being classified as so following successful legal proceedings. The second contribution of this work is explaining their methodology for finetuning the BERT language model for the identification of legally void clauses, in particular with regards to the hyperparameter optimization process for this specific task with a relatively small dataset (here of size around 1,000).

Limitations and discussion. A first limitation of this work is that it is pertaining to German T%C statements, and the model built was finetuned from a pretrained German-language BERT. It might have been interesting to see if an English BERT model, finetuned in a similar way, would have achieved similar levels of performance. A second limitation, noted by the authors of the paper, is that the model may be using the "type" of clause when predicting whether a clause is void or not. For example, in the paper's dataset, 0 out of the 21 clauses on intellectual property were void, whereas 126 of the 305 clauses on payment were void. This fact, that certain clause types in our dataset were virtually never void, whereas other were much more likely to be, may have positively inflated the results. From a Bayesian point of view, this is not necessarily an issue, but it also suggests that the model may be analyzing themes beyond purely legal language.

Why this paper? We chose this paper for two reasons. First, it was the most elaborate paper we found relating specifically to the use of NLP to analyze legally dubious T&C statements. Most others papers applying NLP to legal language have downstream tasks which pertain to broader legal documents such as Supreme Court decisions, like in [2]. Second, the paper by Field et al. highlighted some of the challenges in their data collection methodology, as well as the specificities of finetuning a BERT model for this specific task. We believe that the approach for finetuning the German BERT model that is described in this paper can serve as inspiration for the finetuning of our own English language model for the analysis of English online T&C statements.

Wider research context. This paper is a contribution to the overall field of applying NLP to legal language. Legal language is particularly hard for computers to model because of its highly specialized vocabulary and domain-specific terminology, its ambiguity (lega language is often deliberately vague to allow for flexible interpretations), and well as its complex structure (often using long, nested and convoluted sentences). Understanding legal language also requires understanding the inherent "if-then" logic built inside a given legal clause. In addition, the labeled legal text data is very rare, since it often requires costly expert legal annotation (or annotation from previous court rulings). This paper [1] by Field et al. specifically, examines whether individual German T&C clauses are legally void or valid. They showed that of the keys to achieving high performance with legal language was to finetune a pretrained BERT model on a wider range of parameters than usually indicated, to account for the particularities of legal languages and the lower quantity of data. This is confirmed by other related work such as LEGAL-BERT [3], which also showed that using an expanded grid search when finetuning BERT for legal end-tasks had a significant impact on performance. Other works such as [2] have also showed that similar BERT-based models achieved high performance on the classification of Supreme Court decisions into 15 classes (80% accuracy). Interestingly, the best performance was reached by combining windowing techniques with the LEGAL-BERT model mentioned above. The specific technique used was Stride-64, which uses overlapping sliding windows to process long texts, which is the case for legal texts which often go beyond the maximum token limit for BERT (512).

3 Project description (1-2 pages)

Background. Legal language is one of the most complex and sophisticated forms of natural language. A tragic 2024 case brought this issue to public attention: a woman passed away at Disney Park, and the company argued that her family had waived their right to sue because she had accepted the terms and conditions (T&C) when she signed up for a trial of Disney+ stream back in 2019. The legal clause, buried in fine print, favored the company, highlighting how consumers often unknowingly forfeit their rights. Our project focuses on the analysis of T&C documents, and thus aligns with ongoing research in legal text processing, document summarization, and explainable AI. Our work aims to contribute to social good by exploring AI explainability (XAI) and transparency in corporate policies. AI is often seen as a "black box" in legal applications, and our work could help increase transparency in automated decision-making. Many T&C documents are deliberately obfuscated, so our model could provide consumer-friendly, interpretable AI outputs. If successful, our project could pressure companies to write more consumer-friendly T&C by exposing unfair practices, and incentivize companies to adopt better legal practices.

Goal. Our project aims to develop an NLP model that can summarize and rate T&C, making them easier for consumers to understand. Our project will investigate the effectiveness of transformer-based models in summarizing and rating legal T&C for consumer comprehension. Specifically, we aim to

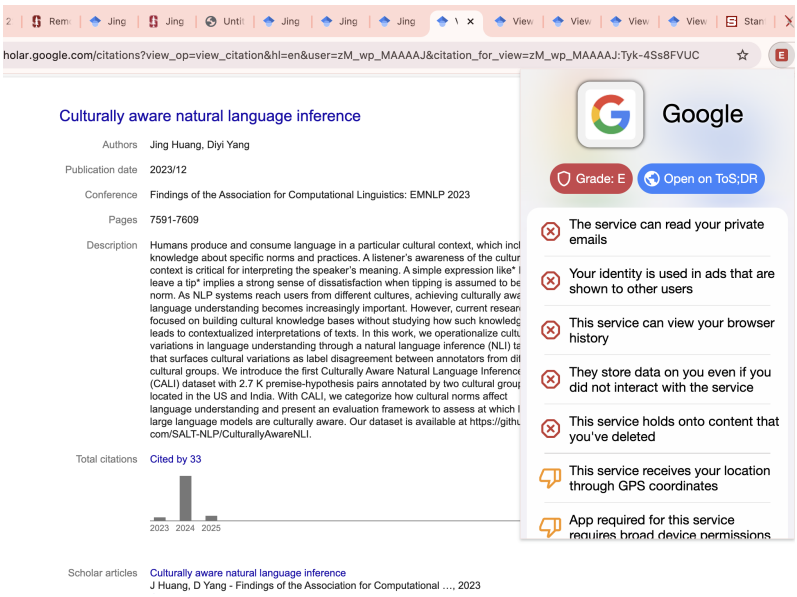

Figure 1: Community Annotated T&C rating and summaries

answer:
1. Can pre-trained language models effectively summarize legal T&C while preserving critical legal meanings?
2. How well do our models detect fairness and bias in T&C documents that match the annotations by attorneys?
3. Can context-aware NLP techniques, inspired by culturally aware language models, improve legal text interpretation and fairness assessment?

We will fine-tune transformer-based models to generate structured, explainable, and user-friendly summaries. Additionally, we will evaluate whether contextual embeddings or retrieval-augmented generation (RAG) improves interpretability and bias detection. This aligns with the objectives and results from [1] where the authors fine-tuned a pretrained BERT model to classify individual legal clauses as legally void or valid. In addition, the insights from this paper, specifically about the need for expanding the grid for the hyperparameter search will be particularly useful for our task.

Task. Our task is to automate the summarization and fairness evaluation of T&C documents using natural language processing (NLP) techniques. Specifically, we aim to:
1. Summarize lengthy and complex T&C documents into concise, user-friendly explanations while preserving key legal implications.
2. Assign a fairness rating to T&C based on consumer rights protections, transparency, and legal risk, using a dataset of attorney-annotated contracts.

An illustration can be found in figure 1.

Data. We will use a community-annotated dataset of T&C of 500 companies, which qualified attorney vonlunteers have analyzed and labeled. This dataset was previously used to develop a browser extension that show T&C fairness rating. It also includes clause-level annotations. Google T&C is ranked as grade E as it includes many clauses that don't respect users' privacy. On the contrary, DuckDuckGo has an excellent rating.

To prepare the dataset for fine-tuning transformer models, we will:
1. Tokenization & Cleaning: Standardize legal text we acquired from the company. Remove the date/version mismatches.
2. Sentence Segmentation: Break long legal clauses into manageable units for NLP models.
3. Label Normalization: Align fairness ratings with a numerical scale for supervised learning.
4. Augmentation (if needed): Expand the dataset using semi-supervised techniques like zero-shot

prompting from GPT models to generate additional annotations.

Methods. Our approach involves fine-tuning transformer-based models for legal text summarization and fairness evaluation. We will explore and fine-tune pre-trained sequence-to-sequence models to evaluate the best outcome. We will use Hugging Face transformers to download models. For our specific task, we will train on our annotated dataset, optimizing for text summarization quality and fairness classification accuracy. In addition, we will implement our legal text pre-processing pipeline, including clause segmentation, tokenization, and explainability analysis. Finally, we will compare our model against existing summarization models (e.g., GPT-4, Claude) to assess performance gains.

Baselines. We will compare our T&C summarization model with Pre-trained Transformer Models such as OpenAI ChatGPT 4o, Claude, etc. We will also compare our classifier with the pre-trained classifiers such as LEGAL-BERT and the human expert annotations.

Evaluation. For our summarization task, we will explore and use various evaluation methods, including n-gram-based semantic similarity (e.g. BLEU). For our multiclass classification task (grading the T&C from A to E), we plan to use ROC-AUC approaches.

Limitations and discussion. The project is based on a community-annotated dataset of T&C analysis of 500 companies' websites. The company's T&C may have updated, but the community annotation probably cannot catch up with it, resulting the wrong annotation. Also, the annotations are all in English, potentially hurt the accuracy of other languages. Despite the limitations, we are still confident the project is meaningful and will contribute to user-right awareness and push the companies to take more social responsibilities.

Ethical Challenges Our project presents ethical challenges related to algorithmic bias in legal fairness assessments and misinterpreting AI-generated summaries. We will perform bias audits by cross-checking AI ratings with the community and asking for feedback. We will also put a legal disclaimer that AI summaries are informational, not legal advice, and recommend consulting a lawyer for critical decisions.

References

[1] Anjalie Field, Shrimai Prabhumoye, Maarten Sap, Zhijing Jin, Jieyu Zhao, and Chris Brockett. NLP for consumer protection: Battling illegal clauses in German terms and conditions in online shopping. In *Association for Computational Linguistics (ACL)*, 2021.

[2] Shubham Vatsal, Adam Meyers, and John E. Ortega. Classification of US Supreme Court cases using BERT-based techniques. In Ruslan Mitkov and Galia Angelova, editors, *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, Varna, Bulgaria, September 2023. INCOMA Ltd., Shoumen, Bulgaria.

[3] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online, November 2020. Association for Computational Linguistics.