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

Abstract

Objective This project aims to utilize Legal-BERT to assess privacy risks in Terms and Conditions (T&C) texts, helping users understand the implications of these terms on their privacy.

Motivation: Legal language is one of the most complex 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 T&C when she signed up for a Disney+ trial 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 on the privacy risks of T&C texts. Our work aims to explore AI explainability (XAI) and transparency in often deliberately obfuscated T&C statements, and to provide consumer-friendly AI outputs. Our baseline findings suggest that unfinetuned models perform rather poorly on legal tasks: zero-shot learning with Meta's Llama 2 with 7B achieved 28.30% accuracy five post-August 2023 T&C statements. However, fine-tuning Legal-BERT with LoRA shows promise for significantly improving performance.

1 Key Information to include

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

2 Approach

Main approach. The project contains two distinct NLP tasks.

- **Task 1: Individual Clause Classification**: Classify each individual clause in T&C documents into 4 categories: "Very Bad", "Bad", "Neutral", or "Good".
- **Task 2: Overall Document Scoring** : Provide an overall privacy risk score (from A to E) for entire T&C documents.

Model Choice: We aim to use Legal-BERT [1] as our base model. Legal-BERT has the advantage of being pretrained on a diverse domain-specific corpus comprising of pieces from EU legislation, UK legislation and US court cases. This pretrained model has been shown to perform well on several downstream legal tasks (e.g. predicting the outcome of a court case given a text describing case's facts).

Fine-Tuning Strategy: For both individual T&C clause classification and overall T&C classification tasks, we plan to fine-tune LEGAL-BERT to improve accuracy for both tasks. However, our dataset of annotated T&C documents contains the annotated statements of approximately 450 websites. Therefore, we plan to use LoRA [2] instead of full fine-tuning. Due to the small size of our dataset, full fine-tuning runs the risk of overfitting, without significant performance benefits compared to LoRA, which minimizes overfitting by only training a small subset of parameters and is much more computationally efficient.

$$\max_{\Theta, |\Theta| \ll |\Phi_0|} \sum_{(x,y) \in Z} \sum_{t=1}^{|y|} \log \left(p_{\Phi_0 + \Delta\Phi(\Theta)}(y_t \mid x, y_{<t}) \right) \tag{1}$$

LoRA fine-tuning maximization problem

This objective function represents optimizing a small subset of Legal-BERT's parameters via LoRA while maintaining generalization across unseen privacy clauses.

Baseline. Our baseline consists of zero-shot learning with Meta's Llama 2 model [3] with 7B parameters. The prompt is: "Rate the privacy invasiveness of the following clause with one word (very bad, bad, neutral, or good): insert clause". Accuracy is our metric for this task. Since Llama 2 was trained between Jan. and Jul. 2023, our baseline is evaluated on 5 T&Cs from after Aug. 2023 (Tuta, Free Music Archive, Mozilla, Discord and Steam), ensuring we are working with "truly" unseen data (previous T&Cs were in the training corpus for Llama 2). We also evaluate our baseline on Facebook's annotated T&Cs.

	LLaMA 7B
Accuracy on 5 post 08/2023 T&C	28.30%
Accuracy on Facebook T&C	21.74%

Table 1: Accuracy of the LLaMA 7B model

3 Experiments

Data. We use a community-annotated dataset of T&C statements of 450 companies, which qualified attorney volunteers have analyzed and labeled. This dataset shows overall T&C fairness rating, as well as clause-level annotations. We have written code to acquire the original T&C documents, overall document and individual clause ratings. However, these need to be adequately processed and cleaned in order to start finetuning our language model (e.g. removal of HTLM tags, of special characters, of disclaimers etc). We will then conduct a class imbalance analysis to examine whether the categories "very bad" and "bad" are significantly underrepresented in our dataset, and analyze its potential impact on model performance. Depending on the extent of the class imbalance, we may leverage data augmentation to equalize the class proportions in our training corpus.

Evaluation method. For both the T&C document privacy risk rating task, and the individual clause privacy risk rating task, our primary metric is accuracy. In comparison, for the related classification of reporting whether clauses in German T&Cs [4] are legally "void" or "valid", the authors reported reaching 90% accuracy.

Baseline results analysis. The zero-shot results were less good than anticipated. Several explanations are possible: the lack of legal domain training (Llama 2 was not explicitly trained on legal text), the ambiguity of legal language (legal clauses may require context which zero-shot prompting does not provide) and the potential limitation of our specific prompt (need for better prompt engineering).

Prompt engineering. For our baseline, we only conducted zero-shot learning. We plan to expand on this by conducting few-shot learning experiments where we provide the model with several examples of legal clauses and their annotations to help it understand the task better. We will pay close attention to the impact of few-shot prompt sample size on classification accuracy. Finally, we will experiment with prompt engineering and evaluate its impact on our results.

Finetuning Legal-BERT. We will finetune the Legal-BERT model using LoRA as described above. We split our dataset into 80% for training, 10% for our dev set to tune hyperparameters such as learning rate and number of training steps, and 10% for our test set to evaluate performance. We will then compare the performance with zero-shot learning, few-shot learning and prompt engineering using Llama 2.

4 Future work and timeline

- **Milestone 1 Data Collection, Cleaning and Analysis (by end of February)** Complete data collection/cleansing, including text cleaning, annotation validation & class imbalance analysis.
- **Milestone 2: Further Baseline Experiments (first week of March)** Conduct few-shot experiment using Llama 2, evaluating the sensitivity of performance to sample size, and then experiment the effect of prompt engineering on performance.
- **Milestone 3: Legal-BERT Fine-tuning (March 15)** Fine-tune Legal-BERT on our training corpus LoRA, for both overall document annotation and individual legal clause rating tasks.
- **Milestone 4: Analysis of the Legal-BERT model** Explore the attention weights of the Legal-BERT model before and after finetuning to our legal clause annotations task to figure out what exactly helped it perform better for the classification.

References

[1] 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.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. 2021.

[3] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. 2023.

[4] 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.