Building a Long-Text Privacy Policy Corpus with Multi-Class Labels

Abstract

This work introduces a new hand-coded dataset for the interpretation of privacy policies. The dataset captures the contents of 162 privacy policies, including documents they incorporate by reference, on 64 dimensions that map onto commonly found terms and applicable legal rules. The coding approach is designed to capture complexities inherent to the task of legal interpretation that are not present in current privacy policy datasets. These include addressing textual ambiguity, indeterminate meaning, interdependent clauses, contractual silence, and the effect of legal defaults.

This paper also introduces the suite of open-source, online tools we developed to build the dataset. The tools are explicitly designed to allow non-technical domain experts to create similar datasets.

1. Introduction

Privacy policies are important legal documents; they govern how firms collect, use, share, and secure personal information. They have become a prime target for automated interpretation. Despite their importance, they are rarely read. Privacy policies are long, complex, and require legal expertise to understand. At the same time, many privacy policies are publicly available and map to a consistent set of well-defined legal questions. This presents an opportunity to represent policy content in a reasonably consistent and somewhat objective manner against which automated interpreters can be tuned and measured.

Recent advancements in machine learning, especially the introduction of the large language model (LLM), increased interest in automating long-text interpretation. Practical legal use of NLP has expanded from term- and clause-level classification to nuanced interpretation of long bodies of legal text. Legal interpretation presents an especially challenging interpretative task. Privacy policies are a stereotypical legal document, as they are drafted by experts and include domain-specific vocabulary and interpretation (Zheng et al., 2021; Mellinkoff, 2004; Mertz, 2007); they can contain inconsistencies and be susceptible to multiple valid interpretations (Reidenberg et al., 2016); and they often contain interdependent clauses—sometime spread across multiple documents—whose meaning is best understood when read wholesale. Like other legal texts, privacy policies must be interpreted in the context of applicable legal rules, which can define terms and provide guidance on issues not explicitly addressed in the text.

Current privacy policy datasets either offer high-granularity labels for short samples of policy text, or low-granularity classification of longer text. These approaches may not capture many domain-specific aspects of legal interpretation that are relevant to the expanding range of automated legal tasks. For example, neither approach accounts for how documents that are "incorporated by reference" may affect the way a policy restricts (or doesn't restrict) the ways in which a company can use user data. As legal interpretation increasingly becomes the target of automation, new datasets are needed. This paper aims to help address that need.

We provide a legal dataset of labeled online privacy policies coded by legally-trained experts. It contains 162 privacy policies along with the documents they incorporate by reference, including Terms of Use, Cookie Policies, California Consumers Privacy Act (CCPA) disclosures, and documents pertaining to compliance with the European Union's General Data Protection Regulation (GDPR). Our coding accounts for the ways in which applicable legal rules and referenced documents can affect the meaning of terms. It also tracks relevant legal terms and reflects aspects of the legal interpretation task, including accounting for ambiguity and reasonable disagreements, and interpreting silence.

2. Related Work

Prior work building datasets for privacy policies mostly focuses on expert annotation or classification of short text. (**?**Lippi et al., 2019; Bui et al., 2021; Ahmad et al., 2021). Some investigation has also looked into crowd-sourcing annotation (Wilson et al., 2018). One privacy-policy-adjacent dataset involving classification of longer legal text labels the content of cookie banner disclosures with the stated purposes for data collection (Santos et al., 2021). In addition to annotated datasets, there are large-scale compilations of privacy policies scraped from the Internet and Internet Archive(Amos et al., 2021; Srinath et al., 2021).

Perhaps the most widely-used privacy policy dataset is the

OPP-115 dataeset introduced by Wilson et. al. in 2016. OPP contains 115 privacy policies that were annotated paragraph-by-paragraph to identify phrases related to 36 data practices grouped into 10 categories. The OPP dataset was used to train prominent tools used to pick out specific clauses from privacy policies (Harkous et al., 2018; Mousavi Nejad et al., 2020). It has also been used to generate related datasets, either by transforming its annotations for use in a new task like question-answering or GDPR compliance (Poplavska et al., 2020; Ahmad et al., 2020), or as an input into composite legal-task benchmarks like LEGALBENCH and PRIVACYGLUE (Guha et al., 2023; Chalkidis et al., 2022). The OPP taxonomy scheme has also been used to organize other privacy-related datasets (Ravichander et al., 2019). Another notable privacy policy dataset—the *unfair-TOS* dataset introduced by Lippi et. al.—annotates "potentially unfair" clauses in privacy practices and is also incorporated into some composite benchmarks, including the privacy-policy-specific PRIVACYGLUE benchmark (Shankar et al., 2023).

Benchmarking legal AIs goes beyond the traditional metrics-and-datasets approach. Alternative evaluation approaches include having NLP systems take the bar exam (Bommarito II & Katz, 2022; Katz et al., 2024) (though some have questioned the efficacy of that evaluation approach (Martínez, 2024)), grading LLM-generated law school exam answers (Choi et al., 2021), and measuring how law student performance is affected by LLM use (Choi & Schwarcz, 2023).

3. Dataset Preparation

Our labeling approach differs from prior work in two key ways. First, by classifying the entire body of relevant text, we capture interplay between related clauses that isn't included in short-text annotation. This provides a dataset that is more representative of the holistic used in real-life legal analysis. Second, we capture relevant ambiguity in text by isolating the causes of disagreement and low confidence by expert coders, and tracking when those disagreements and instances of low confidence are caused by interpretation of policy text.

As part of this project, we built a suite of online tools that automates schema preparation, our coding methodology, and the process of consolidating the resulting dataset into a shareable format. These tools are deliberately designed to require no experience coding or working with ML datasets. We make them available online at (website redacted for anonymous submission) and as an open-source library. In a companion law review article to this technical paper, we discuss why datasets prepared by legal experts are important to the effective development and deployment of LLMs in legal domains. These tools are intended to help lower the barrier to the creation of similar expert-labeled datasets to the one we present here.

Questions	64	
Categories	11	
Total Coders	18	
Coders per Policy	2+	1
Policies	88	74
Paragraphs	29,359	27,372
Words	937,943	977,364
Highlight Annotations	14,811	7,218
Policy Classifications	11,718	4,733
Confidence Scores	9,608	4,464

Table 1. Summary statistics on the corpus at time of submission to GenLaw.

4. Dataset Contents

The dataset comprises annotations for 162 documents. At time of submission to GenLaw, 88 have been coded by at least two coders and 74 were coded by a single coder; we are on track to have a second set of labels for those policies by the end of the summer. For each document and question, the dataset contains classifications selected by each coder, along with a list of sentences the coder marked as relevant to answering the question and their self-reported confidence ranked on a Likert scale. The dataset includes 64 variables motivated by 11 legal categories. For a subset of documents and questions, the dataset also includes the amount of time each coder spent answering the question. Coders were all upper-level law students who had completed coursework covering the relevant topics in contract law. Table 1 provides additional descriptive statistics about the dataset.

5. Results

Using our dataset, we evaluate models on their ability to perform two tasks.

The first task ("holistic classification") is a multi-classification task that uses the entire policy as input: given our questions and a policy from our dataset, select the most likely answers for each question. Comparing coding from flagship commercial models with our dataset, we observe that BCE loss varies significantly between questions and topics. Across all questions, we find that Claude-3-Haiku has an average BCE loss of .292, and GPT-4 has an average loss of .212, suggesting that there is still room for improvement in the automated interpretation of long legal documents.

The second task ("highlight prediction") is an annotation task that targets individual paragraphs: given a question from our dataset and a paragraph from one of the policies in our sample, predict whether a coder marked that paragraph as relevant to answering the question. We report performance statistics for several popular open models against this benchmark.

References

Ahmad, W. U., Chi, J., Tian, Y., and Chang, K.-W. Policyqa: A reading comprehension dataset for privacy policies, 2020.

Ahmad, W. U., Chi, J., Le, T., Norton, T., Tian, Y., and Chang, K.-W. Intent classification and slot filling for privacy policies, 2021.

Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A., and Mayer, J. Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021*, pp. 2165–2176, 2021.

Bommarito II, M. and Katz, D. M. Gpt takes the bar exam. *arXiv preprint arXiv:2212.14402*, 2022.

Bui, D., Shin, K. G., Choi, J.-M., and Shin, J. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2021(2):88–110, 2021. doi: doi:10.2478/popets-2021-0019. URL https://doi.org/10.2478/popets-2021-0019.

Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D. M., and Aletras, N. Lexglue: A benchmark dataset for legal language understanding in english, 2022.

Choi, J. H. and Schwarcz, D. Ai assistance in legal analysis: An empirical study. *Available at SSRN 4539836*, 2023.

Choi, J. H., Hickman, K. E., Monahan, A. B., and Schwarcz, D. Chatgpt goes to law school. *J. Legal Educ.*, 71:387, 2021.

Davis, K. E. and Marotta-Wurgler, F. Contracting for personal data. *NYUL Rev.*, 94:662, 2019.

Guha, N., Nyarko, J., Ho, D. E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D. N., Zambrano, D., Talisman, D., Hoque, E., Surani, F., Fagan, F., Sarfaty, G., Dickinson, G. M., Porat, H., Hegland, J., Wu, J., Nudell, J., Niklaus, J., Nay, J., Choi, J. H., Tobia, K., Hagan, M., Ma, M., Livermore, M., Rasumov-Rahe, N., Holzenberger, N., Kolt, N., Henderson, P., Rehaag, S., Goel, S., Gao, S., Williams, S., Gandhi, S., Zur, T., Iyer, V., and Li, Z. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, 2023.

Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G., and Aberer, K. Polisis: Automated analysis and presentation of privacy policies using deep learning, 2018.

Hosseini, M. B., Heaps, J., Slavin, R., Niu, J., and Breaux, T. Ambiguity and generality in natural language privacy policies. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pp. 70–81. IEEE, 2021.

IAB Tech Lab. IAB Content Taxonomy 3.0, January 2024. URL https://iabtechlab.com/standards/content-taxonomy/.

Katz, D. M., Bommarito, M. J., Gao, S., and Arredondo, P. Gpt-4 passes the bar exam. *Philosophical Transactions of the Royal Society A*, 382(2270):20230254, 2024.

Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.-W., Sartor, G., and Torroni, P. Claudette: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27:117–139, 2019.

Marotta-Wurgler, F. Understanding Privacy Policies: Content, Self-Regulation, and Markets. SSRN Scholarly Paper ID 2736513, Rochester, NY, January 2016a. URL https://papers.ssrn.com/abstract=2736513.

Marotta-Wurgler, F. Self-regulation and competition in privacy policies. *The Journal of Legal Studies*, 45(S2): S13–S39, 2016b.

Marotta-Wurgler, F. Understanding privacy policies: Content, self-regulation, and markets. *NYU Law and Economics Research Paper*, (16-18), 2016c.

Martínez, E. Re-evaluating gpt-4's bar exam performance. *Artificial Intelligence and Law*, pp. 1–24, 2024.

Mellinkoff, D. *The language of the law*. Wipf and Stock Publishers, 2004.

Mertz, E. *The language of law school: learning to" think like a lawyer"*. Oxford University Press, USA, 2007.

Mousavi Nejad, N., Jabat, P., Nedelchev, R., Scerri, S., and Graux, D. Establishing a strong baseline for privacy policy classification. In *ICT Systems Security and Privacy Protection: 35th IFIP TC 11 International Conference, SEC 2020, Maribor, Slovenia, September 21–23, 2020, Proceedings 35*, pp. 370–383. Springer, 2020.

Müller, R., Kornblith, S., and Hinton, G. When does label smoothing help?, 2020.

Poplavska, E., Norton, T. B., Wilson, S., and Sadeh, N. From prescription to description: Mapping the gdpr to a privacy policy corpus annotation scheme. In *Legal Knowledge and Information Systems-JURIX 2020: 33rd Annual Conference*, 2020.

Ravichander, A., Black, A. W., Wilson, S., Norton, T., and Sadeh, N. Question answering for privacy policies: Combining computational and legal perspectives, 2019.

Reidenberg, J. R., Bhatia, J., Breaux, T. D., and Norton, T. B. Ambiguity in privacy policies and the impact of regulation. *The Journal of Legal Studies*, 45(S2):S163–S190, 2016.

Santos, C., Rossi, A., Sanchez Chamorro, L., Bongard-Blanchy, K., and Abu-Salma, R. Cookie banners, what's the purpose? analyzing cookie banner text through a legal lens. In *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society*, pp. 187–194, 2021.

Shankar, A., Waldis, A., Bless, C., Andueza Rodriguez, M., and Mazzola, L. Privacyglue: A benchmark dataset for general language understanding in privacy policies. *Applied Sciences*, 13(6):3701, 2023.

Srinath, M., Wilson, S., and Giles, C. L. Privacy at scale: Introducing the privaseer corpus of web privacy policies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.532. URL http://dx.doi.org/10.18653/v1/2021.acl-long.532.

Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathyendra, K. M., Russell, N. C., et al. The creation and analysis of a website privacy policy corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1330–1340, 2016.

Wilson, S., Schaub, F., Liu, F., Sathyendra, K. M., Smullen, D., Zimmeck, S., Ramanath, R., Story, P., Liu, F., Sadeh, N., and Smith, N. A. Analyzing privacy policies at scale: From crowdsourcing to automated annotations. *ACM Trans. Web*, 13(1), dec 2018. ISSN 1559-1131. doi: 10.1145/3230665. URL https://doi.org/10.1145/3230665.

Zheng, L., Guha, N., Anderson, B. R., Henderson, P., and Ho, D. E. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pp. 159–168, 2021.