Chase Madson

Final Project
Notice and Choice?: Automated Privacy Policy Analysis with BERT

Section 2 | DATASCI 266 | MIDS Program, UC Berkeley School of Information

1. Introduction

Online privacy law in the US is propped up by the free market model of Notice-and-Choice, through which corporations regularly conduct an Orwellian-amount of surveillance so long as they articulate their privacy practices somewhere deep within a bottomless and near-unreadable policy document. The core contradiction of Notice-and-Choice stems from the disparity between the volume of text needed to express a privacy policy and the attention span of users eager to access the service. Worse yet, profit-driven incentives spur online companies to exploit this power imbalance. The Federal Trade Commission – the chief regulatory body in this space – has long-acknowledged that Notice-and-Choice "are inadequate tools for protecting information privacy or security" (Cate 59-62).

Though this issue is fundamentally a political one to be addressed via a national US privacy law akin to the GDPR, privacy advocates have long sought technological solutions that can help even the scales in favor of regular citizens. In this paper, I explore a natural language processing approach to this problem space to attempt to flag privacy policies terms that give notice of the collection and use of personal data, along with terms explaining the user's choices pertaining to it.

2. Related Work

Though perhaps niche in the burgeoning and typically lucrative field of natural language processing, automated privacy policy analysis has been collecting a modest body of research thanks to the work of usable privacy advocates, AI practitioners, and legal academics who appreciate the importance of this problem space.

Schneider and Garnett (2011) developed the annotation tool ConsentCanvas built on unsupervised NLP methods to parse end-user license agreements and highlights concerning clauses. Members of the Usable Privacy Project published OPP-115: a corpus of online privacy policies that have been thoroughly annotated by legal experts (Wilson et al. 2016). This public research dataset addressed a dearth of well annotated privacy policy data (which is understandably expensive) at the level needed to conduct NLP. This data would be used (among others) by Harkous et al. (2018) to build a deep learning model as well as Nejad et al. (2020) to build a BERT model, with each setting benchmarks for automated privacy policy analysis. The analysis these papers perform focus on parsing and classifying parts of policy documents at a high level structure.

In this paper I build upon this field of research using the OPP-115 dataset, where I focus in on detecting specific language enabling the online service provider to collect, use, and share data collected on users, as well as any clauses that notify users of their ability to withdraw consent to these specific practices. An implementation of this model could serve to bolster both notice and choice. May be useful to build into a browser extension or some other application to assist people quickly learn about how their personal data may be used, and how they might opt out, when signing up for a service.

3. Methods

My analysis required data engineering to identify and process the appropriate labels for this task, which is demonstrated in the */notebooks/data_processing.ipynb* notebook. I chose to establish a baseline using a rules-based approach, similar to early work in the problem space that might have used a series of expert-developed regular expressions to match consistent language patterns.

3.1 Interpreting the OPP–115 Data Set

The OPP-115 dataset was developed by the Usable Privacy Policy Project with Carnegie Mellon University for the purposes of public research, teaching, and scholarship. The authors selected a sample of privacy policies from a representative slice of online companies, and then segmented the data into paragraph-length texts for use as units of analysis. Each policy was assigned to three law school graduate students for annotation (from a pool of 10 student annotators). The detailed annotations are meant to represent the full breadth and depth of the privacy practices included in these policy documents (e.g., annotation labels were designed to flag data retention, security, generic/introductory language, contact information, data collection and use).

To create meaningful labels that represent exploitative data collection and use practices, I aimed my attention towards text segments that include language annotated as "First Party Collection/Use" for my first label, and "Third Party Sharing/Collection" for my second. In particular I limited them to data collected, used, or shared for the purposes of advertising, marketing, potential mergers/acquisitions, or no reason given. For my third and forth labels, I chose segments containing language annotated as explaining choices and controls available to users: first party collection and use, and third party collection, use, and sharing, respectively.

3.2 Baseline Model

Prior to machine learning techniques, using a rules-based approach with regular expressions was a prominent way to detect subjects present in a class. My baseline model was to apply a few decently performing regular expression patterns to the test set.

3.3 BERT Model

For my attempt at applying modern natural language processing techniques, I decided to train a BERT model as a base, connect its [CLS] pooling token to a hidden layer with 128 nodes, apply a dropout at a rate of 50%, and connect that output to four binary classification output layers - one for each label.

I experienced some difficulty implementing a multi-label classification model in keras, given there was no coverage of downstream tasks with multiple output heads such as this. There are sparse examples available online that demonstrate how to implement this particular architecture using keras functional API, though I found one image recognition Kaggle notebook[1] which helped me complete my non-working code.

[1] https://www.kaggle.com/code/seraphwedd18/pe-detection-with-keras-model-creation/notebook

4. Results

The resulting model had a test set accuracy of ___, and a test set binary cross-entropy loss of ___. Results at the label-level are shown in Figure 1.

Label	Test BCE Loss	Test Accuracy	Baseline Accuracy
First Party Collection and Use	0.4343	0.8438	0.6406
Third Party Sharing	0.4507	0.8333	0.7619
Control over First Party Collection and Use	0.1576	0.9635	0.8940
Control over Third Party Sharing	0.0981	0.9809	0.9078

Figure 1: Label-level results of model on the test set

During my research, I discovered Legal-BERT - a family of domain-specific BERT models trained on a large corpus of legislation, court cases, and contracts across the English-speaking world (Cite: https://aclanthology.org/2020.findings-emnlp.261/). Unfortunately, after running experiments to test the performance of *legal-bert-uncased* against *base-bert-uncased*, I found no significant improvement in either loss or accuracy. Nonetheless, the legal-bert-uncased was chosen for the final model.

5. Discussion

I discovered a serious generalizability issue with the OPP-115 data set in how it defines “paragraph-length segments” that make up the unit of analysis for classification. There is no consistent definition for how segments should be parsed programmatically, as each policy writer may use their own distinct organization standard and formatting structure. Splitting segments by line breaks does not work for many instances, especially when information is presented as a list which leads to crucial contextual language (e.g., “does collect:” vs. “does not collect:”) being separated from relevant entities (e.g., “...> your biometric data…”). The authors of the study inserted delimiter strings in the files without documenting their process. It is not clear how these delimiter strings would be applied to new policy documents to make for consistent and meaningful “segments”. This introduces noise in the data that is difficult to detect, let alone remedy. Of course, this is a problem caused by idiosyncratic formatting conventions and the lack of standardization requirements in industry, and not simply an oversight of the authors.

Though considered to be in the domain of “natural language”, we should recognize how unnatural the legal jargon used in privacy policies truly is. In addition to its use of abstruse and jargony vernacular, privacy policies are plagued with a variety of formatting decisions that do not translate well in raw string format, such as a liberal use of lists and displaying information in tables. This makes for documents that are laborious to read for both humans and machines alike. Ultimately, these are contracts that serve the primary purpose of conferring the company with exhaustive legal coverage from the citizens their practices harm.

To further articulate the problems of using natural language processing techniques for automated privacy policy detection, these classification models are known to be vulnerable to manipulation. In a recent study, researchers Xu et al. (2022) demonstrated in their paper how transformer-based terms-of-service analysis systems are vulnerable to adversarial attacks. In particular, universal adversarial triggers and LM triggers could be used to “fairwash” a policy document to flip its otherwise “unfair” determination to “fair”, and it would not require direct access to the model.

6. Conclusion

This toy model may not perform at a high-confidence level. However, my takeaway from the literature is that there is strong potential for natural language processing techniques to help bridge the chasm between users and the corporations who use and abuse their data. These efforts are no substitute for structural change in the form of a comprehensive national privacy law, but could be used to confer citizens with a measure of genuine notice and choice.

7. References

Cate, Fred H. “The Limits of Notice and Choice.” *IEEE Security & Privacy*, vol. 8, no. 2, March-April 2010, pp. 59-62. *ieee.org*, https://ieeexplore.ieee.org/document/5439530.

Chalkidis, et al., Ilias. “LEGAL-BERT: The Muppets straight out of Law School.” *Findings of the Association for Computational Linguistics: EMNLP 2020*, no. Nov, 2020, pp. 2898-2904. *ACL Anthology website*, https://aclanthology.org/2020.findings-emnlp.261.

Harkous, et al., Hamza. “Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning.” *Proceedings of the 27th USENIX Security Symposium*, vol. 27th, no. Baltimore, MD, 2018, pp. 531-548. *Usenix website*, https://www.usenix.org/system/files/conference/usenixsecurity18/sec18-harkous.pdf.

Nejad, et al., Najmeh. “Establishing a strong baseline for privacy policy classification.” *IFIP International Conference on ICT*, 2020, pp. 370-383. *Springer website*, https://link.springer.com/chapter/10.1007/978-3-030-58201-2_25.

Schneider, Oliver, and Alex Garnett. “ConsentCanvas: Automatic Texturing for Improved Readability in End-User License Agreements.” *Proceedings of the ACL*, vol. 2011 Student Session, no. June, 2011, pp. 41-45. *ACL Anthology website*, https://aclanthology.org/P11-3008.

Wilson et al., Shomir. “The Creation and Analysis of a Website Privacy Policy Corpus.” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, vol. 1: Long Papers, no. Aug, 2016, pp. 1330-1340. *ACL Anthology website*, https://aclanthology.org/P16-1126.

Xu, et al., Shanshan. “Attack on Unfair ToS Clause Detection: A Case Study using Universal Adversarial Triggers.” *Arvix*, no. Nov, 2022. *Arxiv website*, https://arxiv.org/abs/2211.15556.