PLUE: Language Understanding Evaluation Benchmark for Privacy Policies in English

Jianfeng Chi[1,2] **Wasi Uddin Ahmad**[3*] **Yuan Tian**[3] **Kai-Wei Chang**[3]
[1]Meta AI, [2]University of Virginia, [3]University of California, Los Angeles
jianfengchi@meta.com,{wasiahmad,yuant,kwchang}@ucla.edu

Abstract

Privacy policies provide individuals with information about their rights and how their personal information is handled. Natural language understanding (NLU) technologies can support individuals and practitioners to understand better privacy practices described in lengthy and complex documents. However, existing efforts that use NLU technologies are limited by processing the language in a way exclusive to a single task focusing on certain privacy practices. To this end, we introduce the Privacy Policy Language Understanding Evaluation (PLUE) benchmark, a multi-task benchmark for evaluating the privacy policy language understanding across various tasks. We also collect a large corpus of privacy policies to enable privacy policy domain-specific language model pre-training. We evaluate several generic pre-trained language models and continue pre-training them on the collected corpus. We demonstrate that domain-specific continual pre-training offers performance improvements across all tasks. The code and models are released at https://github.com/JFChi/PLUE.

1 Introduction

Privacy policies are documents that outline how a company or organization collects, uses, shares, and protects individuals' personal information. Without a clear understanding of privacy policies, individuals may not know how their personal information is being used or who it is being shared with. The privacy violation might cause potential harm to them. However, privacy policies are lengthy and complex, prohibiting users from reading and understanding them in detail (Commission et al., 2012; Gluck et al., 2016; Marotta-Wurgler, 2015).

Various natural language understanding (NLU) technologies have recently been developed to understand privacy policies (Wilson et al., 2016a; Harkous et al., 2018; Ravichander et al., 2019; Ahmad et al., 2020; Parvez et al., 2022; Ahmad et al., 2021; Bui et al., 2021). These tasks focus on understanding specific privacy practices at different syntax or semantics levels and require significant effort for data annotations (e.g., domain experts). It is hard to develop generic pre-trained language models (e.g., BERT (Devlin et al., 2019)) with task-specific fine-tuning using limited annotated data. Besides, the unique characteristics of privacy policies, such as reasoning over ambiguity and vagueness, modality, and document structure (Ravichander et al., 2021), make it challenging to directly apply generic pre-trained language models to the privacy policy domain.

To address these problems and encourage research to develop NLU technologies in the privacy policy domain, we introduce the **P**rivacy Policy **L**anguage **U**nderstanding **E**valuation (PLUE) benchmark, to evaluate the privacy policy language understanding across six tasks, including text classification, question answering, semantic parsing, and named-entity recognition. PLUE also includes a pre-training privacy policy corpus that we crawl from the websites to enable privacy policy domain-specific language model pre-training. We use this corpus to pre-train BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), Electra (Clark et al., 2020), and SpanBERT (Joshi et al., 2020) and fine-tune them on the downstream tasks. We demonstrate that domain-specific continual pre-training offers performance improvements across all tasks. We will release the benchmark to assist natural language processing (NLP) researchers and practitioners in future exploration.

2 Policy Language Understanding Evaluation (PLUE) Benchmark

PLUE is centered on six English privacy policy language understanding tasks. The datasets and tasks are selected based on the following principles: (1) usefulness: the selected tasks can help practitioners

*Work done while at UCLA.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 352–365

Dataset	Task	Sub-domain	|Policy|	|Train|	|Dev|	|Test|	Metric
OPP-115	Classification	Websites	115	2,771	395	625	F1
APP-350	Classification	Mobile Apps	350	10,150	2,817	2,540	F1
PrivacyQA	QA	Mobile Apps	35	1,350	–	400	P / R / F1
PolicyQA	QA	Mobile Apps	115	17,056	3,809	4,152	F1 / EM
PolicyIE	Intent Classification Slot Filling	Websites Mobile Apps	31	4,209	–	1,041	F1 / EM
PI-Extract	NER	Websites	30	3,034	–	1,028	F1

Table 1: Statistics of the PLUE datasets and tasks.

in the domain quickly understand privacy practices without reading the whole privacy policy; (2) task diversity: the selected tasks focus on different semantic levels, e.g., words (phrases), sentences, and paragraphs; (3) task difficulty: the selected tasks should be adequately challenging for more room for improvement; (4) training efficiency: all tasks can be trainable on a single moderate GPU (e.g., GeForce GTX 1080 Ti) for no more than ten hours; (5) accessibility: all datasets are publicly available under licenses that allow usage and redistribution for research purposes.

2.1 Datasets and Tasks

PLUE includes six tasks in four categories. Table 1 presents an overview of the datasets and tasks within PLUE, and Table 4 in the Appendix gives an example for each task.

OPP-115 Wilson et al. (2016a) presented 115 Online Privacy Policies (OPP-115). The dataset comprises website privacy policies with text segments annotated with one or more privacy practices from ten categories (see Appendix A.1). We train a multi-label classifier to predict the privacy practices given a sentence from a policy document.

APP-350 Zimmeck et al. (2019) presented APP-350, a collection of mobile application privacy policies annotating what types of users' data mobile applications collect or share. Like OPP-115, each text segment in a policy document is annotated with zero or more privacy practices (listed in Appendix A.2). In total, there are 30 data-type-related classes in APP-350, and we assign one more class, `No_Mention`, to those text segments that do not pertain to such practices.

PrivacyQA Ravichander et al. (2019) proposed a question-answering dataset, PrivacyQA, comprised of 35 mobile application privacy policies. Given a question from a mobile application user and a sentence from a privacy policy, the task is to predict whether the sentence is relevant to the question. PrivacyQA includes unanswerable and subjective questions and formulates the QA task as a binary sentence classification task.

PolicyQA Ahmad et al. (2020) proposed a reading comprehension (Rajpurkar et al., 2016) style dataset, PolicyQA. The dataset is derived from OPP-115 annotations that include a set of fine-grained attributes and evidence text spans that support the annotations. Considering the annotated spans as the answer spans, PolicyQA generates diverse questions relating to the corresponding privacy practices and attributes. The task is to predict the answer text span given the corresponding text segment and question.

PolicyIE Ahmad et al. (2021) proposed a semantic parsing dataset composed of two tasks: intent classification and slot filling. Given a sentence in a privacy policy, the task is to predict the sentence's intent (i.e., privacy practice) and identify the semantic concepts associated with the privacy practice. Based on the role of the slots in privacy practices, PolicyIE groups them into type-I and type-II slots. In total, there are four intent labels and 14 type-I and four type-II slot labels. We individually train a text classifier and sequence taggers to perform intent classification and slot filling, respectively.

PI-Extract Bui et al. (2021) presented PI-Extract, a named-entity recognition (NER) dataset. It aims to identify what types of user data are (not) collected or shared mentioned in the privacy policies. It contains 4 types of named entities: `COLLECT`, `NOT_COLLECT`, `SHARE` and `NOT_SHARE`. Note that the named entities of different types may overlap. Thus, we report the results for collection-related and share-related entities, respectively.

2.2 Pre-training Corpus Collection

The existing pre-trained language models (PLMs) mostly use data from BooksCorpus (Zhu et al.,

2015) and English Wikipedia. Language models pre-trained on text from those sources might not perform well on the downstream privacy policy language understanding tasks, as privacy policies are composed of text written by domain experts (e.g., lawyers). Gururangan et al. (2020) suggested that adapting to the domain's unlabeled data (domain-adaptive pre-training) improves the performance of domain-specific tasks. Therefore, we collect a large privacy policy corpus for language model pre-training. In order to achieve broad coverage across privacy practices written in privacy policies (William, 2020; Ahmad et al., 2021), we collect the privacy policies from two sources: mobile application privacy policies and website privacy policies. Appendix B provides more details about how we collect these two types of privacy policies.

2.3 Models & Training

Baselines We benchmark pre-trained language models (PLMs), BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), SpanBERT (Joshi et al., 2020), Electra (Clark et al., 2020), and LEGAL-BERT (Chalkidis et al., 2020). We present the details of the PLMs in Appendix C.

Domain-specific Continual Pre-training In order to adapt PLMs to the privacy policy domain, we continue to train BERT, Electra, SpanBERT, and RoBERTa on the pre-training corpus described in Section 2.2. We refer to them as PP-BERT, PP-RoBERTa, PP-SpanBERT, and PP-Electra, respectively.[1] We present details in Appendix D.1.

Task-specific Fine-tuning We fine-tune PLMs for each PLUE task. We *only* tune the learning rate for each task, as we found in the preliminary experiments that model performances are highly sensitive to the learning rate. We present more details in Appendix D.2.

3 Experiment Results

Tables 2 and 3 present the results for all the experiment models for PLUE tasks. Rows 2-9 show the results of the base PLMs and their corresponding variants with privacy policy domain-specific continual pre-training. Similar to GLUE (Wang et al., 2019), we also provide the average scores of all PLUE tasks in the last column of Table 3. We observe that the language models (PP-BERT, PP-SpanBERT, PP-Electra, PP-RoBERTa) adapted to the privacy policy domain outperform the general language models consistently in all the tasks, and PP-RoBERTa performs the best among all base models in terms of the average scores of all PLUE tasks. In particular, PP-RoBERTa performs the best for OPP-115, APP-350, PrivacyQA,[2] and PI-Extract, among all base models. PP-BERT and PP-RoBERTa perform the best for PolicyQA; PP-Electra and PP-RoBERTa achieve the best performance for PolicyIE. In contrast, LEGAL-BERT (row 10) performs comparably or shows moderate improvements over BERT, indicating that pre-training on the general legal corpus does not necessarily help privacy policy language understanding.

It is interesting to see that continual pre-training of the language models using the privacy policy domain data benefits them differently. For example, in the text classification tasks (i.e., OPP-115 and APP-350), the performance difference between SpanBERT and PP-SpanBERT are most significant, while models using MLM (BERT and RoBERTa) already shows relatively high performance before continual pre-training and continual pre-training brings moderate gains to BERT and RoBERTa.

We further investigate the improvement of large variants of PLMs over base variants of PLMs on PLUE tasks. Since PP-RoBERTa$_{\text{BASE}}$ performs the best among all base models, we also continue pre-train RoBERTa$_{\text{LARGE}}$ (PP-RoBERTa$_{\text{LARGE}}$). As shown in the last five rows in Tables 2 and 3, the large pre-trained language models mostly outperform their base counterparts. Noticeably, PP-RoBERTa$_{\text{LARGE}}$ is the best-performing model in APP-350, PolicyQA, PI-Extract, and sub-tasks in PolicyIE, and it also achieves the highest average scores of all PLUE tasks among all models.

Lastly, even though domain-specific pre-training and large PLMs help boost the performance for all tasks, the performance of some tasks and datasets (e.g., APP-350, PrivacyQA, slot filling in PolicyIE) remains low, which indicates much potential for further work on NLP for the privacy policy domain.

4 Related Work

Privacy Policy Benchmarks The Usable Privacy Policy Project (Sadeh et al., 2013) is the most significant effort to date, resulting in a large pool of works (Wilson et al., 2016a,b; Sathyendra et al.,

[1] We continually pre-train only the base models to mitigate the environmental impact of our experiments, but our code supports continual pre-training of large PLMs too.

[2] Ravichander et al. (2019) reported 39.8% F1 score for BERT model; however, we are able to achieve 36.3%.

Models	|Model|	OPP-115	APP-350	PrivacyQA	PolicyQA	PI-Extract
		F1	F1	P / R / F1	F1 / EM	F1
Human	-	-	-	68.8 / 69.0 / 68.9	-	-
$\text{BERT}_{\text{BASE}}$	110M	75.3	59.6	44.6 / 35.9 / 36.3	55.1 / 27.7	63.7 / 54.6
$\text{Electra}_{\text{BASE}}$	110M	74.0	49.3	42.7 / 36.0 / 36.1	57.5 / 29.9	69.4 / 57.8
$\text{SpanBERT}_{\text{BASE}}$	110M	62.8	32.8	24.8 / 24.8 / 24.8	55.2 / 27.8	66.9 / 41.0
$\text{RoBERTa}_{\text{BASE}}$	124M	79.0	67.1	43.6 / 36.4 / 36.7	56.6 / 29.4	70.7 / 56.8
$\text{PP-BERT}_{\text{BASE}}$	110M	78.0	62.8	44.8 / 36.9 / 37.7	58.3 / 30.0	70.5 / 55.3
$\text{PP-Electra}_{\text{BASE}}$	110M	73.1	57.1	48.3 / 38.8 / 39.3	58.0 / 30.0	70.3 / 61.2
$\text{PP-SpanBERT}_{\text{BASE}}$	110M	78.1	61.9	43.4 / 36.4 / 36.8	55.8 / 27.5	65.5 / 50.8
$\text{PP-RoBERTa}_{\text{BASE}}$	124M	**80.2**	69.5	**49.8** / **40.1** / **40.9**	57.8 / 30.3	71.2 / 61.3
$\text{LEGAL-BERT}_{\text{BASE}}$	110M	76.0	57.4	45.6 / 37.6 / 38.2	55.1 / 27.7	69.1 / 51.1
$\text{BERT}_{\text{LARGE}}$	340M	79.3	71.2	43.8 / 35.4 / 36.1	56.6 / 28.7	68.1 / 54.8
$\text{Electra}_{\text{LARGE}}$	340M	78.7	41.5	46.6 / 42.1 / 40.5	60.7 / **33.2**	70.1 / 59.5
$\text{SpanBERT}_{\text{LARGE}}$	340M	79.4	66.0	45.2 / 36.5 / 37.3	58.2 / 30.8	68.2 / 50.8
$\text{RoBERTa}_{\text{LARGE}}$	355M	79.9	72.4	47.6 / 41.4 / 40.6	59.8 / 32.5	70.9 / 62.8
$\text{PP-RoBERTa}_{\text{LARGE}}$	355M	79.8	**74.5**	49.3 / 39.5 / 40.4	**61.1** / **33.2**	**71.6** / **66.9**

Table 2: Performance comparison of pre-trained models on text classification, question answering, and named entity recognition tasks. We fine-tune all the models three times with different seeds and report average performances. Human performances are reported from the respective works.

Models	|Model|	Intent Classification	Slot Filling				
			Type-I Slots		Type-II Slots		
		F1	F1	EM	F1	EM	**Avg**
Human	-	96.5	84.3	56.6	62.3	55.6	-
$\text{BERT}_{\text{BASE}}$	110M	73.7	55.2	19.7	34.7	29.8	48.2
$\text{Electra}_{\text{BASE}}$	110M	73.7	56.4	22.8	36.5	30.7	49.1
$\text{SpanBERT}_{\text{BASE}}$	110M	71.9	44.0	10.8	29.7	17.5	44.2
$\text{RoBERTa}_{\text{BASE}}$	110M	74.5	56.8	22.0	39.2	32.0	50.0
$\text{PP-BERT}_{\text{BASE}}$	110M	76.9	56.7	22.8	38.7	32.5	50.7
$\text{PP-Electra}_{\text{BASE}}$	110M	77.1	58.2	**24.1**	37.8	**32.9**	50.8
$\text{PP-SpanBERT}_{\text{BASE}}$	110M	75.0	54.1	19.8	33.6	26.7	48.4
$\text{PP-RoBERTa}_{\text{BASE}}$	110M	**78.1**	58.0	22.4	40.1	32.4	52.3
$\text{LEGAL-BERT}_{\text{BASE}}$	110M	72.6	53.8	19.5	36.1	29.7	48.6
$\text{BERT}_{\text{LARGE}}$	340M	75.5	56.8	23.0	38.4	32.2	50.0
$\text{Electra}_{\text{LARGE}}$	340M	75.6	57.9	24.0	39.6	32.4	50.2
$\text{SpanBERT}_{\text{LARGE}}$	340M	73.8	45.5	9.5	38.8	29.8	48.2
$\text{RoBERTa}_{\text{LARGE}}$	355M	77.6	58.4	22.9	41.4	32.7	52.9
$\text{PP-RoBERTa}_{\text{LARGE}}$	355M	77.7	**59.8**	23.9	**42.0**	32.3	**53.7**

Table 3: Performance comparison of pre-trained models on intent classification and slot filling tasks (PolicyIE) and average scores of all PLUE tasks. We fine-tune all the models three times with different seeds and report average performances. Human performances are reported from the respective works.

2016; Mysore Sathyendra et al., 2017; Bhatia and Breaux, 2015; Bhatia et al., 2016; Hosseini et al., 2016; Zimmeck et al., 2019) to facilitate the automation of privacy policy analysis. A wide range of NLP techniques have been explored accordingly (Liu et al., 2014; Ramanath et al., 2014; Wilson et al., 2016a; Harkous et al., 2018; Zimmeck et al., 2019; Shvartzshanider et al., 2018; Harkous et al., 2018; Ravichander et al., 2019; Ahmad et al., 2020; Bui et al., 2021; Ahmad et al., 2021).

Pre-trained Language Models In the last few years, NLP research has witnessed a radical change with the advent of PLMs like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019). PLMs achieved state-of-the-art results in many language understanding benchmarks. Consequently, PLMs have been developed for a wide range of domains, *e.g.,* scientific (Beltagy et al., 2019), medical (Lee et al., 2020; Rasmy et al., 2021; Alsentzer et al., 2019), legal (Chalkidis et al., 2020), and cybersecurity (Ranade et al., 2021; Bayer et al., 2022). This

work investigates the adaptation of PLMs to facilitate NLP research in the privacy policy domain.

5 Conclusion and Future work

Reliable aggregation of datasets and benchmarking foundation models on them facilitate future research. This work presents PLUE, a benchmark for training and evaluating new security and privacy policy models. PLUE will help researchers benchmark policy language understanding under a unified setup and facilitate reliable comparison.

PLUE also presents some challenges in language understanding evaluation for privacy policies. For example, the imbalance data issue for privacy practices is a major challenge in the PrivacyQA task (Parvez et al., 2022). Data efficiency is also a challenge for continual pre-training as the amount of unlabeled data is also small for this domain. Approaches such as (Qin et al., 2022) could be investigated to continually adapt LMs for the emerging data in this domain.

Limitations

The pre-training privacy policy corpus and the downstream task datasets are unlikely to contain toxic or biased content. Therefore, they should not magnify toxicity or bias in the pre-trained and fine-tuned models, although the models may exhibit such behavior due to their original pre-training. The pre-training and benchmark datasets are formed based on privacy policies crawled in the past; as a result, they could be outdated by now. This work focuses on the English language only, and the findings may not apply to other languages.

Ethics Statement

License The OPP-115 and APP-350 datasets are made available for research, teaching, and scholarship purposes only, with further parameters in the spirit of a Creative Commons Attribution-NonCommercial License (CC BY-NC). The PolicyQA and PI-Extract datasets are derived from OPP-115 datasets. The PrivacyQA and PolicyIE datasets are released under an MIT license. The pre-training corpus, MAPS Policies Dataset, is released under CC BY-NC. We strictly adhere to these licenses and will release the PLUE benchmark resources under CC BY-NC-SA 4.0.

Carbon Footprint We only use RoBERTa large models for continual training on the privacy policy domain to reduce the environmental impacts of training large models. The PP-BERT, PP-SpanBERT, PP-Electra, and PP-RoBERTa models were trained for 100k steps on Tesla V100 GPUs that took 1-2 days. Therefore, the training would emit only 9kg of carbon into the environment.[3] All fine-tuning experiments were very lightweight due to the small size of the datasets, resulting in approximately 12kg of carbon emission.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. This work was supported in part by National Science Foundation Grant OAC 2002985, OAC 1920462, and CNS 1943100, Google Research Award, CISCO Research Award, and Meta Research Award. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not necessarily reflect those of the US Government or NSF.

References

Wasi Ahmad, Jianfeng Chi, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang. 2021. Intent classification and slot filling for 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)*, pages 4402–4417, Online. Association for Computational Linguistics.

Wasi Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. PolicyQA: A reading comprehension dataset for privacy policies. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 743–749, Online. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Ryan Amos, Gunes Acar, Eli Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy policies over time: Curation and analysis of a million-document dataset. In *Proceedings of the Web Conference 2021*, WWW '21, page 2165–2176, New York, NY, USA. Association for Computing Machinery.

[3] Calculated using `https://mlco2.github.io/impact`, based on a total of 100 hours of training on Tesla V100 and Amazon Web Services as the provider.

Markus Bayer, Philipp Kuehn, Ramin Shanehsaz, and Christian Reuter. 2022. Cysecbert: A domain-adapted language model for the cybersecurity domain. *arXiv preprint arXiv:2212.02974*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Jaspreet Bhatia and Travis D Breaux. 2015. Towards an information type lexicon for privacy policies. In *2015 IEEE eighth international workshop on requirements engineering and law (RELAW)*, pages 19–24. IEEE.

Jaspreet Bhatia, Morgan C Evans, Sudarshan Wadkar, and Travis D Breaux. 2016. Automated extraction of regulated information types using hyponymy relations. In *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)*, pages 19–25. IEEE.

Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated extraction and presentation of data practices in privacy policies. *Proceedings on Privacy Enhancing Technologies*, 2021(2):88–110.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Federal Trade Commission et al. 2012. Protecting consumer privacy in an era of rapid change. *FTC report*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 2016. How short is too short? implications of length and framing on the effectiveness of privacy notices. In *Twelfth Symposium on Usable Privacy and Security ({SOUPS} 2016)*, pages 321–340.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} Security Symposium ({USENIX} Security 18)*, pages 531–548.

Mitra Bokaei Hosseini, Sudarshan Wadkar, Travis D Breaux, and Jianwei Niu. 2016. Lexical similarity of information type hypernyms, meronyms and synonyms in privacy policies. In *2016 AAAI Fall Symposium Series*.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Fei Liu, Rohan Ramanath, Norman Sadeh, and Noah A. Smith. 2014. A step towards usable privacy policy: Automatic alignment of privacy statements. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 884–894, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Florencia Marotta-Wurgler. 2015. Does "notice and choice" disclosure regulation work? an empirical study of privacy policies,". In *Michigan Law: Law and Economics Workshop*.

Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the provision of choices in privacy policy text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2774–2779, Copenhagen, Denmark. Association for Computational Linguistics.

Md Rizwan Parvez, Jianfeng Chi, Wasi Uddin Ahmad, Yuan Tian, and Kai-Wei Chang. 2022. Retrieval enhanced data augmentation for question answering on privacy policies. *arXiv preprint arXiv:2204.08952*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Yujia Qin, Jiajie Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. ELLE: Efficient lifelong pre-training for emerging data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2789–2810, Dublin, Ireland. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A. Smith. 2014. Unsupervised alignment of privacy policies using hidden Markov models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 605–610, Baltimore, Maryland. Association for Computational Linguistics.

Priyanka Ranade, Aritran Piplai, Anupam Joshi, and Tim Finin. 2021. Cybert: Contextualized embeddings for the cybersecurity domain. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 3334–3342. IEEE.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13.

Abhilasha Ravichander, Alan W Black, Thomas Norton, Shomir Wilson, and Norman Sadeh. 2021. Breaking down walls of text: How can NLP benefit consumer privacy? 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)*, pages 4125–4140, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Alan W Black, Shomir Wilson, Thomas Norton, and Norman Sadeh. 2019. Question answering for privacy policies: Combining computational and legal perspectives. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4947–4958, Hong Kong, China. Association for Computational Linguistics.

Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Aleecia M McDonald, Joel R Reidenberg, Noah A Smith, Fei Liu, N Cameron Russell, Florian Schaub, et al. 2013. The usable privacy policy project. *Technical report, Technical Report, CMU-ISR-13-119*.

Kanthashree Mysore Sathyendra, Florian Schaub, Shomir Wilson, and Norman Sadeh. 2016. Automatic extraction of opt-out choices from privacy policies. In *2016 AAAI Fall Symposium Series*.

Yan Shvartzshanider, Ananth Balashankar, Thomas Wies, and Lakshminarayanan Subramanian. 2018. RECIPE: Applying open domain question answering to privacy policies. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 71–77, Melbourne, Australia. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Henry William. 2020. Do web apps and mobile apps need separate privacy policies?

Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016a. 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)*, pages 1330–1340, Berlin, Germany. Association for Computational Linguistics.

Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A Smith, and Frederick Liu. 2016b. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International Conference on World Wide Web*, pages 133–143. International World Wide Web Conferences Steering Committee.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. Maps: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies*, 2019(3):66–86.

Supplementary Material: Appendices

A Dataset Details

A.1 OPP-115 Privacy Practices

1. First Party Collection/Use
2. Third Party Sharing/Collection
3. User Choice/Control
4. User Access, Edit, and Deletion
5. Data Retention
6. Data Security
7. Policy Change
8. Do Not Track
9. International and Specific Audiences
10. Other

A.2 APP-350 Privacy Practices

1. Contact
2. Contact_Address_Book
3. Contact_City
4. Contact_E_Mail_Address
5. Contact_Password
6. Contact_Phone_Number
7. Contact_Postal_Address
8. Contact_ZIP
9. Demographic
10. Demographic_Age
11. Demographic_Gender
12. Facebook_SSO
13. Identifier
14. Identifier_Ad_ID
15. Identifier_Cookie_or_similar_Tech
16. Identifier_Device_ID
17. Identifier_IMEI
18. Identifier_IMSI
19. Identifier_IP_Address
20. Identifier_MAC
21. Identifier_Mobile_Carrier
22. Identifier_SIM_Serial
23. Identifier_SSID_BSSID
24. Location
25. Location_Bluetooth
26. Location_Cell_Tower
27. Location_GPS
28. Location_IP_Address
29. Location_WiFi
30. SSO

B More Details of Pre-training Corpora

We use MAPS, the mobile application privacy policy corpus presented by Zimmeck et al. (2019). MAPS consists of the URLs of 441K mobile application privacy policies, which were collected from April to May 2018 from the Google Play store. We remove the duplicated URLs, crawl the privacy policy documents in HTML/PDF format, convert them to raw text format, and filter out the documents with noise (e.g., empty documents resulting from obsolete URLs). Finally, we ended up with 64K privacy policy documents. For website privacy policies, we use the Princeton-Leuven Longitudinal Corpus of Privacy Policies (Amos et al., 2021).[4] The Princeton-Leuven Longitudinal Corpus of Privacy Policies contains 130K website privacy policies spanning over two decades. We use the documents with the latest date and convert them (from markdown format) into text format. Combining these two corpora, we obtain our pre-training corpus with 332M words.

C Baseline Models

We benchmark a few pre-trained language models as baselines to facilitate future work.

BERT Devlin et al. (2019) proposed Transformer (Vaswani et al., 2017) based language model pre-trained on BooksCorpus and English Wikipedia data using masked language modeling (MLM) and next sentence prediction.

Electra Clark et al. (2020) pre-trains a generator and a discriminator on the same corpus as BERT, where the generator takes a masked text as input and is trained using the MLM objective. The discriminator takes the predictions from the generator and detects which tokens are replaced by the generator. After pre-training, the generator is discarded, and the discriminator is used as the language model for the downstream tasks.

SpanBERT Joshi et al. (2020) shares the same architecture and pre-training corpus as BERT but differs in the pre-training objectives. It extends BERT by masking contiguous spans instead of single tokens and training the span boundary representations to predict the masked spans.

RoBERTa Liu et al. (2019) presented a replication study of BERT pretraining where they showed that BERT was significantly undertrained and proposed RoBERTa that tunes key hyperparameters

[4] The corpus is publicly available at https://github.com/citp/privacy-policy-historical.

OPP-115	**Text:** *Secure Online Ordering For your security, we only store your credit card information if you choose to set up an authorized account with one of our Sites. In that case, it is stored on a secure computer in an encrypted format. If you do not set up an account, you will have to enter your credit card information each time you order. We understand that this may be a little inconvenient for you, but some customers appreciate the added security.* **Classes:** `Data Security`, `User Choice/Control`, `First Party Collection/Use`
APP-350	**Text:** *Our Use of Web Beacons and Analytics Services Microsoft web pages may contain electronic images known as web beacons (also called single-pixel gifs) that we use to help deliver cookies on our websites, count users who have visited those websites and deliver co-branded products. We also include web beacons in our promotional email messages or newsletters to determine whether you open and act on them.* **Classes:** `Contact_E_Mail_Address`, `Identifier_Cookie_or_similar_Tech`
PrivacyQA	**Sentence:** *We may collect and use information about your location (such as your country) or infer your approximate location based on your IP address in order to provide you with tailored educational experiences for your region, but we don't collect the precise geolocation of you or your device.* **Question:** *Does the app track my location?* **Answer:** `Relevant`
PolicyQA	**Text:** *Illini Media never shares personally identifiable information provided to us online in ways unrelated to the ones described above without allowing you to opt out or otherwise prohibit such unrelated uses. Google or any ad server may use information (not including your name, address, email address, or telephone number) about your visits to this and other websites in order to provide advertisements about goods and services of interest to you.* **Question:** Do you share my data with others? If yes, what is the type of data? **Answer:** *information (not including your name, address, email address or telephone number)*
PolicyIE	**Sentence:** *We may also use or display your username and icon or profile photo for marketing purposes or press releases.* **Intent:** `Data Collection/Usage` **Slots:** (1) *Data Collector: First Party Entity–We*, (2) *Action–use*, (3) *Data Provider: User–your*, (4) *Data Collected: User Online Activities/Profiles–username*, (5) *Data Collected: User Online Activities/Profiles–icon or profile photo*, (6) *Purpose: Advertising/Marketing–marketing purpose or press releases*.
PI-Extract	**Text:** *We may share aggregate demographic and usage information with our prospective and actual business partners, advertisers, and other third parties for any business purpose.* **Entities:** `SHARE` – *aggregate demographic and usage information*

Table 4: Examples from the tasks in PLUE.

and uses more training data to achieve remarkable performance improvements. Note that while BERT, Electra, and SpanBERT use the same vocabulary, RoBERTa uses a different vocabulary resulting in 15M more parameters in the model.

LEGAL-BERT Chalkidis et al. (2020) pre-trained BERT using 12 GB of the English text (over 351K documents) from several legal fields (e.g., contracts, legislation, court cases) scraped from publicly available resources. Since privacy policies serve as official documents to protect the company and consumers' privacy rights and might contain contents in response to privacy law (e.g., GDPR), we study LEGAL-BERT's effectiveness on the PLUE tasks.

D More Implementation Details

D.1 Domain-specific Continual Pre-training

Since BERT, Electra, and SpanBERT share the same model architectures, we use almost the same hyperparameters (e.g., learning rate, train steps, batch size) for them following the original papers. We scale down the train steps by the same factor, as the size of our pre-training corpus is roughly 1/10 the size of the pre-training corpus of BERT. We adhere to the guidelines outlined in Liu et al. (2019) to train RoBERTa with larger batch size, higher learning rate, and fewer train steps. Table 5 presents the training hyperparameters for PLMs.

D.2 Task-specific Fine-tuning

We fine-tune the models for each task using the Adam (Kingma and Ba, 2015) optimizer with a batch size of 32. We fine-tune the models on the QA tasks for 3 epochs and other tasks for 20 epochs and perform a grid search on the learning rate for each task with validation examples. We chose the learning rate for tasks without validation examples based on our findings from the tasks with validation examples. Table 6 lists the hyperparameters for all the downstream tasks.

In OPP-115 and APP-350, we compute the class weights (the class weights are inversely proportional to the occurrences of the classes) and apply them in fine-tuning, as we find out both datasets have the class-imbalance problem and using class weights brings gains to overall performance. We also report the human performances for PrivacyQA and PolicyIE from the original works.

D.3 Software Tools

To facilitate using PLUE, we release our implementation, which is built with Pytorch (Paszke et al., 2019) and the Huggingface transformers[5] package. Our implementation includes the continual pre-training of our baselines and the evaluation of any PLMs supported by the Huggingface transformers package on the PLUE benchmark tasks. In addition to PLUE datasets, we release the pre-training corpus and all data pre-processing scripts, including the pre-training corpus crawling scripts, to assist future research in this area.

[5] https://github.com/huggingface/transformers

	PP-BERT	PP-SpanBERT	PP-Electra	PP-RoBERTa
Learning Rate	1e-4	1e-4	1e-4	6e-4
Train Steps	100,000	100,000	100,000	12,500
batch Size	256	256	256	2048
Learning Rate Schedule	linear	polynomial_decay	linear	linear
# warm-up steps	1000	1000	1000	600
Optimizer	AdamW	AdamW	AdamW	AdamW

Table 5: Hyperparameters for pre-training language models.

	Text Classification	Question Answering	Semantic Parsing	NER
Dropout	0.1	0.1	0.1	0.1
Weight decay	0.0	0.0	0.0	0.0
Optimizer	AdamW	AdamW	AdamW	AdamW
Batch Size	32	32	32	32
Learning rate	[3e-4, 1e-4, 5e-5, 3e-5, 1e-5, 5e-6, 3e-6]			
Learning Rate Schedule	Linear	Linear	Linear	Linear
Warm-up Ratio	0.05	0.0	0.05	0.05
# epoch	20	3	20	20

Table 6: Hyperparameters for fine-tuning pre-trained language models on different PLUE tasks.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
In the section "Limitations"

☑ A2. Did you discuss any potential risks of your work?
In the sections "Limitations" and "Ethics Statement"

☑ A3. Do the abstract and introduction summarize the paper's main claims?
In the abstract and section 1 "Introduction"

☑ A4. Have you used AI writing assistants when working on this paper?
Yes, we use Grammarly and ChatGPT for assistance purely with the language of the paper (e.g., grammar error checking and paper paraphrasing). We mainly use them in the introduction.

B ☑ Did you use or create scientific artifacts?

In section 2, we describe the creation of our benchmark.

☑ B1. Did you cite the creators of artifacts you used?
In section 2, we cite the creators of the artifacts we used.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In the section "Ethics Statement."

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
In the section "Ethics Statement."

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2.

C ☑ Did you run computational experiments?

Section 3.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
In the section "Ethics Statement."

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 2.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 3.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
In appendix D.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.