Calpric: Inclusive and Fine-grain Labeling of Privacy Policies with Crowdsourcing and Active Learning

Wenjun Qiu, David Lie, and Lisa Austin, *University of Toronto*

https://www.usenix.org/conference/usenixsecurity23/presentation/qiu

This artifact appendix is included in the Artifact Appendices to the Proceedings of the 32nd USENIX Security Symposium and appends to the paper of the same name that appears in the Proceedings of the 32nd USENIX Security Symposium.

August 9–11, 2023 • Anaheim, CA, USA

978-1-939133-37-3

Open access to the Artifact Appendices to the Proceedings of the 32nd USENIX Security Symposium is sponsored by USENIX.

USENIX'23 Artifact Appendix: Calpric: Inclusive and Fine-grained Labeling of Privacy Policies with Crowdsourcing and Active Learning

Wenjun Qiu David Lie Lisa Austin
University of Toronto

A Artifact Appendix

A.1 Abstract

The Calpric project leverages active learning and crowdsourcing techniques to address the challenge of the cost of training accurate deep learning models on privacy policies. In this artifact appendix, we describe the two artifacts available to the public: the Calpric Privacy Policy Corpus (CPPS) and the source codes for Calpric's major components.

We provide the source codes for Calpric as a reference, including the category models and the action models, as well as the privacy policy-based embedding PriBERT. We do not include a full pipeline test for Calpric as the complete system involves additional manual setup and accessing costs, such as the AWS account used to actively crowdsource training labels.

A.2 Description & Requirements

The CPPS data set includes privacy policy segment labels covering 9 data categories (contact, device, location, health, financial, demographic, survey, social media and personally identifiable information) with 3 data actions (collect/use, share, and store). For clarity purposes, duplicated labels have been removed, resulting in a total of 12,585 labels.

A.2.1 Security, privacy, and ethical concerns

The use of human annotators was approved by our institutional review board (IRB). We do not include any personal identifiable information in the publicly accessible dataset.

A.2.2 How to access

The artifacts are accessible via the Calpric GitHub page: https://github.com/dlgroupuoft/Calpric.

A.2.3 Hardware dependencies

The CPPS dataset and PriBERT embedding do not require any specific hardware feature. The source codes support both CPU and GPU processing. Depending on the size of the active querying pool, the required memory size may vary.

A.2.4 Software dependencies

To use the CPPS dataset, no other software dependencies are needed except the Python standard library and the *csv* package.

For the source code example, the following Python packages are required along with the Python standard library: *re, langdetect, numpy, pandas, keras, modAL and tensorflow*.

A.2.5 Benchmarks

Data required by the artifacts: CPPS.

A.3 Set-up

A.3.1 Installation

No other installation is required other than the software dependencies described above.

A.3.2 Basic Test

Run functionality_checks.py. The expected output from the CPPS check should be:
number of health labels: {'health': 1796}

A.4 Version

Based on the LaTeX template for Artifact Evaluation V20220926. Submission, reviewing and badging methodology followed for the evaluation of this artifact can be found at https://secartifacts.github.io/usenixsec2023/.