Assessing Purpose-Extraction for Automated Corpora Annotations

Vincent Miller
Columbus State University
4225 University Ave, Columbus, GA 31907
miller_vincent@columbusstate.edu

Jesus R. Rijo Candelario
Mercer University
1501 Mercer University Dr, Macon, GA 31207
Jesus.Rafael.Rijo@live.mercer.edu

Dr. Lydia Ray
Columbus State University
4225 University Ave, Columbus, GA 31907
ray_lydia@columbusstate.edu

Alfredo J. Perez
University of Nebraska at Omaha
alfredoperez@unomaha.edu

2022 IEEE 19th International Conference on Mobile Ad Hoc and Smart Systems (MASS) | 978-1-6654-7180-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/MASS56207.2022.00107

***Abstract*—Privacy policies contain important information regarding the collection and use of user's data. As Internet of Things (IoT) devices have become popular during the last years, these policies have become important to protect IoT users from unwanted use of private data collected through them. However, IoT policies tend to be long thus discouraging users to read them. In this paper, we seek to create an automated and annotated corpus for IoT privacy policies through the use of natural language processing techniques. Our method extracts the purpose from privacy policies and allows users to quickly find the important information relevant to their data collection/use.**

***Keywords*—privacy policy, Internet of Things, natural language processing, purpose extraction**

I. Introduction

Privacy Policies are legal documents that disclose how a party collects, uses, manages, and shares information on a client or user's data. The type of information collected includes Private Identifiable Information. Many users do not read these policies simply because of their length and complexity. As the Internet-of-Things (IoT) becomes more prevalent in our lives, many privacy policies are not read, making users agree to use IoT devices which may expose not only their data, but aspects of their lives considered private.

The complexity and time required to read privacy policies have led to research automated privacy policy reading tools to find information relevant to users [1] and minimize the effort to comprehend them. In this work, we present a study of IoT privacy policies. Our contributions are as follows:

- We present an algorithm to crawl and find IoT privacy policies in the Internet
- We assess a purpose extraction approach for automatically creating a corpus containing annotations for IoT privacy policies
- We leverage the use of Natural Language Processing (NLP) to find the meaning of sentences in IoT policies and classify them into categories based on the sentence's purpose, allowing users to quickly find relevant information

II. Methods

In this section we describe our approach to automatically annotate IoT privacy policies. We divided this problem into three primary tasks:

- Implementing and running a web-crawler scheme to acquire IoT privacy policies.
- Implementing and assessing our implementation of the PurExt approach [2]
- Generating a corpus of annotated privacy policies

A. *Web Crawling and Preliminary Preprocessing*

We first compiled a list of IoT company names based on the public database provided by IoT ONE [3]. Recognizing that IoT privacy policies do not share a similar HTML structure, we chose to keep text contained within common HTML elements, such as the paragraph tag and list element tag.

B. *Sentence Classification*

We decided to create our own implementation in Python of the PurExt framework as described by Yang et al. [2]. In our Python implementation, we used the spaCy library which is an open source, industrial strength Natural Language Processing library (NLP) [4]. This allowed us to easily create a pipeline to tokenize, create part-of-speech (POS) tags, dependency parse, and named entity recognition (NER) labels for each word in the dataset.

PurExt classifies sentences into three categories including explicit sentences, implicit sentences, and other sentences. Explicit and implicit sentences involve sentences related to the data collection or use in privacy policies. Explicit sentences are syntactically based, whereas implicit sentences are semantically based but both have a syntactic structure [Fig 1] that PurExt considers for rule extraction.

$$Sentence_{explicit} = Noun_{purpose} + Verb_{link} \vee Verb_{contain} + Purpose$$

$$Sentence_{implicit} = Subject + Verb_{CoU} + Data\ Object + Purpose$$

$$Purpose\text{-}Aware\ Rule = \{Action, Data\ Object, Purpose\}$$

Fig. 1. Syntactic structures of purpose-aware rules, explicit sentences, and implicit sentences

2155-6814/22/$31.00 ©2022 IEEE
DOI 10.1109/MASS56207.2022.00107

C. *Rule Extraction and Corpus Creation*

Lastly, PurExt extracts privacy rules from the sentences. We organized and created our final corpus using the Pandas open source library which is a high performance tool with convenient data structures and tools for data analysis [5]. Each IoT privacy policy received its own Pandas dataframe (a table). Each entry contains the sentence type, the original sentence, action, data object, and the purpose. This allowed us to easily examine our data to find and correct any issues. We exported each data frame as a comma separated values (CSV) file. Furthermore, the file's name represents the company the data was extracted from. Our corpus has a total of 2134 rule extracted files, one for each processed IoT privacy policy.

III. Results

We present our results for the explicit and implicit extraction in table 1. We had surprisingly low explicit classification, even though the PurExt authors stated that their explicit classification approach is syntactically based. When we tested the database provided by Yang et al. with our implementation, it classified 112 out of the 120 explicit sentences correctly. These eight sentences follow a simple sentence structure and fail the PurExt's check to verify if the subject is modified by a complement containing at least one CoU verb. When we observed our data and examined potential reasons behind the low explicit classification, we inferred that privacy policies appear to have complex sentence structures. There are 8 types of complete sentences [6]. Therefore, we determined that the PurExt framework must perform additional checks regarding the syntactic structure of sentences to handle more complex sentences. We also had low implicit classification results. We expected low results once we tried to follow PurExt authors' approach of retraining the NER model. When we contacted PurExt, they did not share their annotated dataset for retraining of the NER model, which made replicating their work impossible. The implicit classification, although it has some syntactic structure, is mainly semantically based. Due to the legal definitions of words in privacy policies, the NER labels would need retraining to correctly label them.

TABLE I. Results

Parameter	Result
Total Explicit Sentences	65
Total Implicit Sentences	9948
Average Explicits per Policy	0.03
Average Implicits per Policy	4.66
Action Extraction Accuracy	100%
Data Extraction Accuracy	100%
Purpose Extraction Accuracy	100%

Because of the lack of access to skilled annotators, we were unable to retrain the NER model to improve these results. We also observed that library updates may have also played a key role in our results. In their work Yang et al. 's implementation, they used version 2.x of the spaCy library, whereas we used version 3.3 in our implementation.

IV. Conclusion and Future Work

Privacy policies stand as a core component toward helping users understand the primary implications of using any IoT device. Given the length and complexity of most IoT privacy policies, researchers will continue working towards making privacy policies more readable for most users. In this work we created a web crawling framework for IoT privacy policies and we implemented a purpose extraction tool based on NLP. In order to assess the quality of our implemented natural language processing framework, we curated a publicly available dataset of IoT privacy policies using our web crawling and purpose extraction framework. In future works, we plan to use or create a dataset with manual annotations to train the NER model of our purpose extraction implementation. Likewise, we plan to explore the syntactic structure to find ways to handle more complex sentences.

Acknowledgment

This material is based upon work supported by the National Science Foundation REU Program under Grant No. 1950416. Special thanks to Dr. Lydia Ray, Dr. Alfredo Perez, and Dr. Yesem Peker for their knowledge and mentorship during the Research Experience for Undergraduates program at Columbus State University.

References

[1] Liu, F., Wilson, F., Story, P., Zimmeck, S., and Sadeh, N. 2018. Towards Automatic Classification of Privacy Policy Text. Carnegie Mellon University, CMU‐ISR‐17‐118R, Institute for Software Research, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

[2] L. Yang, X. Chen, Y. Luo, X. Lan, L. Chen, "PurExt: Automated Extraction of the Purpose-Aware Rule from the Natural Language Privacy Policy in IoT", *Security and Communication Networks*, vol. 2021, Article ID 5552501, 11 pages, 2021. https://doi.org/10.1155/2021/5552501.

[3] *About IOT One - Leading Digitalization Consultancy*. IoT ONE. (n.d.). Retrieved June 10, 2022, from https://www.iotone.com/suppliers.

[4] "Spacy." Industrial-Strength Natural Language Processing in Python. https://spacy.io/.

[5] "Pandas." Pandas, https://pandas.pydata.org/.

[6] "A Definition of a Complete Sentence." *A Definition of a Complete Sentence*, https://www.csuohio.edu/writing-center/definition-complete-sentence.