Towards Measuring Risk Factors in Privacy Policies

Najmeh Mousavi Nejad*
Fraunhofer IAIS & University of Bonn
Sankt Agustin, Germany
nejad@cs.uni-bonn.de

Damien Graux
Fraunhofer IAIS
Sankt Agustin, Germany
damien.graux@iais.fraunhofer.de

Diego Collarana
Fraunhofer IAIS
Sankt Agustin, Germany
diego.collarana.vargas@iais.fraunhofer.de

ABSTRACT

The ubiquitous availability of online services and mobile apps results in a rapid proliferation of contractual agreements in the form of privacy policies. Despite the importance of such consent forms, the majority of users tend to ignore them due to their content length and complexity. In this study, we propose a hybrid approach which measures a privacy policy's risk factor applying both supervised deep learning and rule-based information extraction. Benefiting from an annotated dataset of 115 privacy policies, a deep learning component is first able to predict high-level categories for each paragraph. Then, a rule-based module extracts pre-defined attributes and their values, based on high-level classes. Finally, a privacy policy's risk factor is computed based on these attribute values.

KEYWORDS

privacy policy, deep learning, rule-based information extraction, risk factor

1 INTRODUCTION

In the current digital era, almost everyone is exposed to accepting contractual agreements in the form of privacy policies. However, the majority of people skip privacy policies due to their length and complex terminology. According to a recent survey, from 543 university students, only 26% did not choose the 'quick join' routine, while joining a factious social network and unsurprisingly, their average reading time was only 73 seconds [3]. Moreover, some users attempt to apply self protection methods to ensure that their personal data are not being misused, e.g., either by providing false information or installing a privacy protection software [1].

To assist end-users with consciously agreeing to the conditions, we can apply Natural Language Processing (NLP) and Information Extraction (IE) to present a privacy policy in a structured view. Our approach applies supervised deep learning using an annotated dataset (named OPP-115), to assign high-level classes to a privacy policy's paragraphs. Then, according to predicted classes, we define hand-coded rules based on experts annotations, to extract attributes values from each paragraph. Finally, having detailed information for each paragraph, a risk measurement function computes a risk factor based on extracted information. Consequently, a user could choose to stop using a website, if the predicted risk score is high.

OPP-115 is a widely-used dataset in the context of privacy policy analysis [6]. It contains in-depth annotations for 115 privacy policies at paragraph level and each paragraph was annotated by 3 experts. There are two types of annotations: high-level classes which define 10 data practice categories; and low-level attributes which include mandatory and optional attributes. For instance, the high-level class *First Party Collection/Use* has 3 attributes: *Collection Mode (explicit or implicit)*, *Information Type (financial, health, contact, location, etc.)* and *Purpose (advertising, marketing, analytics, legal requirement, etc.)*.

The approach proposed in this paper, is built upon on our previous effort, which exploits OPP-115 and deep learning to solve a multi-label classification problem. We feed privacy policy's paragraphs along with the predicted classes into a rule-based IE component and retrieve attribute values. The rules are defined based on OPP-115 low-level annotations. Finally, all predicted categories and extracted information are passed into a risk measurement module and a risk factor will be computed based on hand-coded rules.

The paper is divided into the following sections: in Section 2, we provide an overview of existing effort on measuring risks in privacy policies; Section 3 presents our proposed approach and our evaluation scheme; and finally Section 4 will conclude this paper.

2 RELATED WORK

In light of the, now enforced EU-wide, General Data Protection Regulation (GDPR) [5], there has been an increasing interest towards privacy policy analysis as this new set of regulations increases the constrains for companies holding customers data. Here, we provide a brief overview of studies that specifically addressed risk levels in privacy policies.

`Polisis` is an online service for automatic analysis of privacy policies [2]. Along with classification and structured presentations of privacy policies, it assigns privacy icons which are based on the *Disconnect*[1] icons. These icons include *Expected Use*, *Expected Collection*, *Precise Location*, *Data Retention* and *Children Privacy*. For instance, *Data Retention* color assignments are: Green for retention periods of less than a year; Yellow, when the retention period is longer than one year; and Red, when there is no data retention policy provided. `Polisis` benefits from OPP-115 and employs supervised machine learning to extract high-level categories (in the above example, *Data Retention*) and attribute values of each category (e.g., *Retention Period* in this case). Finally, based on retrieved attribute values and heuristic rules, privacy icons along with their colors are produced. Currently, `Polisis`'s interface can only generate a limited set of privacy icons. In future, we intend

In: Proceedings of the Workshop on Artificial Intelligence and the Administrative State (AIAS 2019), June 17, 2019, Montreal, QC, Canada.

Published at http://ceur-ws.org

[1] https://disconnect.me/

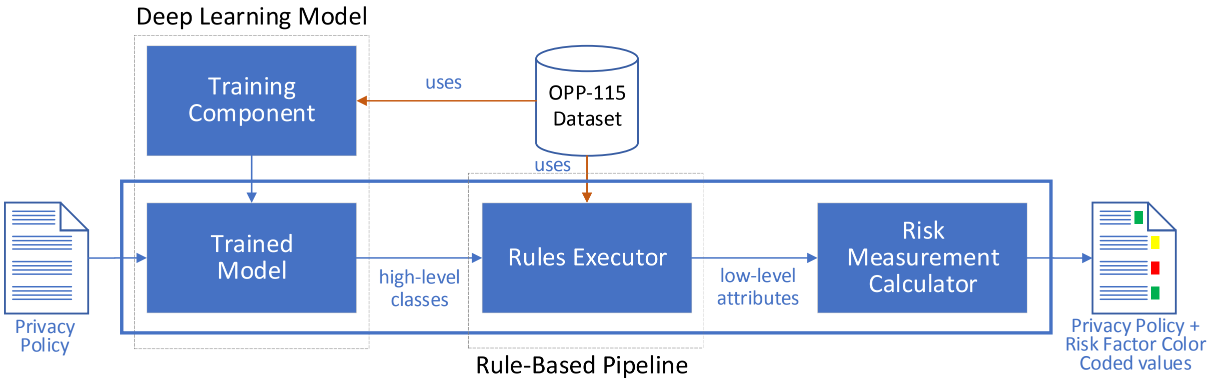

Figure 1: General Architecture.

to further analyze privacy icons and extend them with the help of legal experts.

PrivacyCheck is an approach for automatic summarization of privacy policies using data mining [7]. It answers 10 pre-defined questions concerning privacy and security of users' data and is also available as a Chrome browser extension. In order to train the model, a corpus containing 400 privacy policies was compiled and 7 privacy experts manually assigned risk levels (Green, Yellow, Red) to the 10 factors. First, a pre-processing step finds those paragraphs that have at least one keyword related to one of 10 factors. The methodology of selecting keywords was largely manual. Then, the selected paragraphs will be sent to a data mining server where 11 data mining models were trained, one for checking if the corresponding page is a privacy policy and one each for the 10 questions. The authors claim that on average, 60% of the times, PrivacyCheck finds the correct risk level. The limitation of PrivacyCheck is its lack of Inter Annotator Agreement (IAA) for the annotators. According to the paper, the quality control was performed by assigning each policy to two team members. However, only 15% of privacy policies were compared and their discrepancies were resolved which makes the training dataset less reliable.

PrivacyGuide is another summarization tool inspired by GDPR that classifies a privacy policy into 11 categories using NLP and machine learning and further measures the associated risk level of each class [4]. Similar to previous studies, PrivacyGuide uses the three-level scale risk based on classification (i.e. Green, Yellow, Red). The 11 criteria and their associated risk levels were defined by GDPR experts. Based on these criteria, a privacy corpus was compiled with the help of 35 university students. Each participant assigned a privacy category to text snippets and classified them with a risk level. The author reported that the weighted average accuracy is 74% for classifying a privacy policy into one of the 11 classes and the accuracy of risk level detection is 90%. Although the results were encouraging, the dataset was not annotated by experts which is a fundamental criterion in legal text processing and analysis.

3 PROPOSED APPROACH

In this section, we provide details of our approach for measuring a privacy policy's risk factor. Our proposed method leverages OPP-115 annotated dataset for training and evaluation [6]. As discussed earlier, OPP-115 high-level annotations are divided into 10 classes:

(1) *First Party Collection/Use*: how and why the information is collected.
(2) *Third Party Sharing/Collection*: how the information may be used or collected by third parties.
(3) *User Choice/Control*: choices and controls available to to users.
(4) *User Access/Edit/Deletion*: if users can modify their information and how.
(5) *Data Retention*: how long the information is stored.
(6) *Data Security*: how is users' data secured.
(7) *Policy Change*: if the service provider will change their policy and how the users are informed.
(8) *Do Not Track*: if and how Do Not Track signals[2] is honored.
(9) *International/Specific Audiences*: practices that target a specific group of users (e.g., children, Europeans, etc.)
(10) *Other*: additional practices not covered by the other categories.

In addition, each high-level category includes low-level attribute annotations. For instance, *Data Retention* category is further annotated with its attributes, which are: *Retention Period*, *Retention Purpose* and *Information Type*. The annotators provided either one or several values for each attribute along with the span of text based on which they have chosen that specific value(s). In the above example, *Retention Period* may have one of the following values: *stated period*, *limited*, *indefinitely* or *unspecified*.

Figure 1 shows the architecture of our proposed approach which consists of three main components: 1) a deep learning module is trained to predict high-level classes of a policy's paragraphs; 2) a rule-based pipeline in which the rules are defined based on low-level attribute annotations of OPP-115; and 3) a risk measurement function that assigns risk icons along with their corresponding colors (green, yellow, red), according to extracted information.

Following conventional ML practices, in the deep learning component, dataset splits are randomly partitioned into a ratio of 3:1:1 for training, validation and testing respectively; while maintaining

[2] https://en.wikipedia.org/wiki/Do_Not_Track

Table 1: Sample rules for extracting values of Retention Period from Data Retention Category.

Rule	Value	Sample
`[delete/remove][Token]*[after][number][day/month/year]`	Stated Period	1. We remove the entirety of the IP address after 6 months. 2. All stored IP addresses, except the account creation IP address, are deleted after 90 days.
`[not][Token]*[delete/remove]`	Indefinitely	The posts and content you made will not be automatically deleted as part of the account removal process.
`[store/keep/retain/maintain][Token]*[indefinitely]`	Indefinitely	1.This data is generally retained indefinitely. 2. The information we collect for statistical analysis and technical improvements is maintained indefinitely.
`[store/keep/retain/maintain][Token]*[as long as][Token]+`	Limited	1. We will retain your information for as long as your account is active or as needed to provide you services. 2. We will retain your personal information while you have an account and thereafter for as long as we need it for purposes not prohibited by applicable laws
`If not one of the above conditions`	Unspecified	1. We receive and store certain types of information whenever you interact with us. 2. The personal information collected about you through our online applications and in our communications with you is stored in our internal database.

a stratified set of labels. We further decomposed the *Other* category into its attributes: *Introductory/Generic*, *Privacy Contact Information* and *Practice Not Covered*. Therefore, considering that a paragraph in the dataset may be labeled with more than one category, we face a multi-label classification problem with 12 classes. The implementation of the ML component is completed and we achieve 79% micro-average for F1.

The high-level predicted classes are passed to the rule-based component where low-level attribute values will be extracted. The definition of rules are based on experts annotations in OPP-115 dataset. We intend to use 60% of low-level annotations for defining the rules, 20% for validating the defined rules and the remaining 20% for the final test. Table 1 shows some sample rules for finding values of *Retention Period* attribute in *Data Retention* category. We found our rules definitions based on experts annotations. As shown in the table, the rules definition use the knowledge about high-level categories predicted by the deep learning component.

Having information about attribute values, the risk measurement module is able to assign appropriate risk icons along with their corresponding colors. As a proof-of-concept, we will found our risk measurement rules on *Disconnect* icons. Aforementioned in literature review, the *Disconnect Data Retention* color assignment are as follows: Green for retention period <= 12 months; Yellow, for retention period > 12 months; and Red, when there is no data retention policy provided. Algorithm 1 shows our interpretation of *Data Retention* icon. It is worth to mention that our interpretation is based on the available annotations from OPP-115 dataset. Hence, it is not the only representation that can be built from *Disconnect* icons and others may adopt their own understanding.

Algorithm 1 Sketch of risk measurement algorithm

```
Require: predicted high-level category, extracted attribute values
for all paragraphs in the privacy policy do
    $category$ ← predicted high-level category
    if $category$ ∈ Data Retention then
        $RetentionPeriod$ ← extracted retention period
        if $RetentionPeriod$ ∈ (Stated Period, Limited) then
            $DataRetentionIcon$ ← Green
        else if $RetentionPeriod$ ∈ Indefinitely then
            $DataRetentionIcon$ ← Yellow
        else
            $DataRetentionIcon$ ← Red
        end if
    end if
    if $category$ ∈ First Party Collection/Use then ...
    end if
end for
Ensure: risk icons and their corresponding colors
```

For the evaluation of our approach, we intend to generate risk factors according to OPP-115 experts annotations and use it as a goldstandard. We believe the final error will be close to sum of error rate in the deep learning module (predicting high-level classes) and the error which is caused due to incomplete set of rules in rule executor component. Considering the fact that we are now able to predict the correct high-level classes with 79% F1, with the careful definition of rules for extracting attribute values, it is predicted to gain a reasonable accuracy at the end of our pipeline.

4 CONCLUSION

In this study, we proposed the application of Deep Learning models and Rule-Based Information Extraction to automatically present a structured view of risk factors in privacy policies. In particular, we presented a hybrid approach that takes advantage of the dataset OPP-115. We state that this approach is of paramount importance to support users to consciously agree with terms and conditions of online services. As next steps, we plan to implement the proposed architecture and run empirical evaluations to validate the presented hypothesis, i.e, users will be more motivated to read privacy policies when a color-coded structured view is presented to them.

REFERENCES

[1] A. Acquisti and J. Grossklags. Privacy and rationality in individual decision making. *IEEE Security and Privacy*, 3(1):26–33, Jan. 2005.

[2] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. *CoRR*, abs/1802.02561, 2018.

[3] J. A. Obar and A. Oeldorf-Hirsch. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society*, pages 1–20, 2018.

[4] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna. Privacyguide: Towards an implementation of the eu gdpr on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*, IWSPA '18, pages 15–21, New York, NY, USA, 2018. ACM.

[5] P. Voigt and A. Von dem Bussche. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 2017.

[6] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, 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)*, volume 1, pages 1330–1340, 2016.

[7] R. N. Zaeem, R. L. German, and K. S. Barber. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Trans. Internet Technol.*, 18(4):53:1–53:18, Aug. 2018.