The Case for a GDPR-specific Annotated Dataset of Privacy Policies

Matthias Gallé, Athena Christofi and **Hady Elsahar**

Abstract

In this position paper we analyse the pros and contras for the need of a dataset of privacy policies, annotated with GDPR-specific elements. We revise existing related data-sets and provide an analysis of how they could be augmented in order to facilitate machine-learning techniques to assess privacy policies with respect to their compliance or not to GDPR.

Introduction

Natural language is the *de facto* channel of communicating terms of use for the use of a digital or physical service. However, it is well known that their acceptance constitutes “the biggest lie on the Internet”, as they are very rarely read; largely due do their length, complexity and vagueness [Massey et al., 2013, Liu et al., 2016, Lebanoff and Liu, 2018]. It seems like an obvious application for natural language processing techniques: analyse the policies with a trained model that extracts the important information and present them in a succinct, unambigous and personalised way. It is nevertheless not always clear what exact information should be extracted, what is relevant to the current legislation and how this should be done. Only recently were there attempts to formalise this through annotations, which not only detail the information to be extracted, but also allow the creation of data-driven algorithms that could then be applied to non-annotated policies [Wilson et al., 2016, Tesfay et al., 2018].

In this paper we revise those data-sets in light of the recent introduction of the General Data Protection Regulation (GDPR) in the European Union, and call out four potential problems that the use of those data-set can have when training models that would help address the new requirements of that regulation.

Existing Datasets

There are many new opportunities that would become possible with the automatic analysis of privacy policies. General trends could be considered, as well as quantifying different practices across time, geographies and industries. New services could exist, such as aggregators which could take into account requirements of users and help them find companies that fulfill those requirements. For the owners of those companies, compliance with new regulations would be much easier to obtain and, of course, the work of Data Protection Agencies would be enormously simplified.

One proposal to achieve this is to directly stop communicating privacy policies through natural language, and rely purely on some formal language [Kelley et al., 2009]. This proposal has the appealing characteristics that it can be consumed easily by machines and be transformed directly into a structured data-base. Despite potential interest for the user, it is clear that this approach has not been embraced by the market.

Instead, several natural-language processing techniques have been proposed to analyze privacy policies, and extract relevant elements. Initially, and probably due to a lack of annotated data, those techniques have been unsupervised, trying to provide insights into the completeness of the policies and comparing them between themselves [Ramanath et al., 2014, Costante et al., 2012].

The opportunities for a (semi-)automatic processing took a significative leap with the release of an annotated dataset. OPP-115 [Wilson et al., 2016] – an outcome of the UsablePrivacy project – is an annotated dataset of 115 privacy policies. This dataset has opened up applications like Pribot, a question-answering system addressing privacy policies [Harkous et al., 2018]. The annotation ontology was carefully designed, covering many aspects. As an example, it has a very broad coverage of the purposes that the personal data can be used (Advertising, Analytics, Legal requirement, Marketing, Perform service, Service operation and security, Others, Unspecified) which are very relevant to GDPR legislation.

A more recent dataset concerns European companies [Tesfay et al., 2018], although it is much smaller (45 privacy policies). Moreover, the annotation is shallower and only covers first-level categories (so, a phrase can be annotated as belonging to “data collection”, but not its purpose or duration of retention).

Datasets for GDPR

The introduction of the General Data Protection Regulation (GDPR) caused significant changes to the practice of digital companies holing personal data, as well as the way they communicate with their clients.

GDPR element	OPP-115	Tesfay et al. [2018]
Information on the company (including DPO)	X	X
Type of personal data	✓	X
Purpose of processing	✓	X
Storage Period	✓	✓
Transfer 3rd country	(✓)	(✓)
Source of personal data	X	X
Rights to withdraw, object, etc	(✓)	(✓)
Automated decision making	X	X
Cookie policy	(✓)	X

Table 1: Overview of GDPR elements and their presence in existing data-sets. Both datasets annotate transfer to third parties, not necessarily if those are in other countries. Source of personal data has to be specified only if obtained indirectly (Art. 14). Checkmarks in parentheses means that this aspect is only covered partially or indirectly.

GDPR

The GDPR entered into force in May 2018 and aims at providing a unified framework of the treatment of personal data for EU citizens. It compels service providers to specify what personal data is collected, for which purpose and if it is shared with third parties. It also provides a series of rights to individuals, such as right to object, to access the data, to request its erasure, etc. Purely automated decision-making is treated separately (Article 22): in case user's personal data is subject to such a process, including profiling, it has to be clearly disclosed in the privacy policy.

Throughout the rest of this paper, all articles and recitals refer to the text of the GDPR.[1]

Existing data-sets in the light of the GDPR

It would be tempting to just use algorithms trained on existing data-sets to assess compliance with the GDPR. In this section we argue that any such use should consider at least four considerations, each one of which risks to introduce some major bias:

1. Impact of new elements
2. Impact of multi-linguality
3. Impact of domain shift due to the type of companies
4. Impact of domain shift due to adaptation to the GDPR

Impact of new elements Analysis of privacy policies targeting the GDPR has some very specific elements which are not addressed by current datasets. We compared the requirements of the GDPR with the annotations of the two data-sets, and provide an overview in Table 1.

In first place, the GDPR requires companies to appoint a Data Protection Officer (DPO) whose contact details have to be provided. None of the existing data-sets provide such annotations, although arguably this could be done with standard NLP pre-trained models (named entity recognizers). Only OPP-115 provides an taxonomy of types of personal data, or different types of the purposes of their processing. The GDPR is very clear that those purposes have to be specified, and provides special provisions for so-called "sensitive data". Recognizing the transparency of borders on the Internet, special care is taken in the GDPR for transfer to third countries (not in the EU). This is only very partially covered by existing data-sets that only have labels for transfer to third parties, without special treatment of where they operate. Art. 14 specifies information that have to be provided to the user if personal data was not obtained directly from her. Such requirements are not covered in existing data-sets

The GDPR provides a series of rights to the individual, including the right to:

- withdraw consent (Art. 7)
- object, including direct marketing (Art. 21)
- access (Art. 15)
- rectification (Art. 16)
- erasure (Art. 17)
- restrict processing (Art. 18)
- data portability (Recital 68)
- lodge complaint (Art. 77)

OPP-115 and Tesfay et al. [2018] have only one relevant label ("User Choice","Control of Data" respectively) which is very generic. None of the existing data-sets provide annotations for the special case that the personal data is subject to automated processing.

While mentioned only once (Recital 30), the GDPR states clearly that cookies can be considered as personal data as it can be used to identify the individual. As such, it falls under the same considerations as other personal data, which is need of explicit consent, disclosure of purpose, lifespan as well as means to opt-out in the future. OPP-115 considers cookies as one type of personal data, but only if that is present in privacy policies. More detailed information is presented most often in a separate cookie policy, which is not considered.

Impact of multi-linguality The European Union has 24 official languages. This becomes particularly an issue when considering smaller companies (see next point), as most

[1] https://gdpr-info.eu/

micro-enterprises only provide services in their local language. Multilingual natural language processing is a very broad topic, and an open area of research which is beyond the current proposal. This might not be needed, as it could be that the formulaic legal language can be translated through automatic translation system (trained, for instance, on the widely used EuroParl corpus [Koehn, 2005]). However, the effect of a privacy law in a highly multi-lingual region like Europe makes the need much more stringent of a data-set available in at least two languages.

Domain-shift 1: type of companies The GDPR is a law at the EU-level, and applies to any company doing business with EU citizens. While large, digital companies are obviously affected by that law it also affects any other business including SMEs and micro-enterprises. In Europe, small and medium-size enterprise employ together 90 million people, representing 2/3 of the EU-28 workforce. While Wilson et al. [2016] takes care of sampling across different domains, all the privacy policies are from popular websites (based on their alexa.com ranking). Similarly, Tesfay et al. [2018] selects European websites, but focuses on the 45 most accessed websites (again, following their alexa.com ranking)

Domain-shift 2: impact of GDPR It became obvious to any Internet user that a major modification happened in May 2018 when they received notifications mail of updates to privacy policies of subscribed services. Nowadays, most policies are written having in mind the new legislation. The impact of such a shift in language, with the appearance of new terminology and more details creates a so-called domain-shift in the distribution of the data. Current machine-learning tools are famously very sensitive to such shifts, and the impact on applying algorithms trained on pre-GDPR data to pos-GDPR policies should at least be measured.

Conclusion

The GDPR is meant to change the way personal data is processed by European and multi-national companies doing business in Europe. As such, natural language processing techniques have much to offer, in particular to facilitate small enterprises to be compliant. However, current techniques of machine learning rely heavily on the existence of annotated data-set. We have revised existing data-sets which could be useful for that task, and highlighted four considerations which should be taken into account before applying algorithms trained on those data-sets. The introduction of new GDPR elements could be addressed by annotating only those missing elements. However, the impact of analysing policies in different languages, as well as the domain shift due to the type of companies whose policies were annotated and the fact that those policies changed substantially after the new legislation should at least be measured.

If the analysis of one or all of those considerations shows that there is a substantial impact on the results, there will be need for a new GDPR-specific dataset.

Acknowledgment The work has received funding from the European Unions Horizon 2020 research and innovation programme under grant agreement No. 786741.

References

Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. A machine learning solution to assess privacy policy completeness:(short paper). In *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*, pages 91–96. ACM, 2012.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. *arXiv preprint arXiv:1802.02561*, 2018.

Patrick Gage Kelley, Joanna Bresee, Lorrie Faith Cranor, and Robert W. Reeder. A "nutrition label" for privacy. In *Proceedings of the 5th Symposium on Usable Privacy and Security*, SOUPS '09, pages 4:1–4:12, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-736-3. doi: 10.1145/1572532.1572538. URL `http://doi.acm.org/10.1145/1572532.1572538`.

Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

Logan Lebanoff and Fei Liu. Automatic detection of vague words and sentences in privacy policies. *arXiv preprint arXiv:1808.06219*, 2018.

Fei Liu, Nicole Lee Fella, and Kexin Liao. Modeling language vagueness in privacy policies using deep neural networks. In *2016 AAAI Fall Symposium Series*, 2016.

Aaron K Massey, Jacob Eisenstein, Annie I Antón, and Peter P Swire. Automated text mining for requirements analysis of policy documents. In *2013 IEEE 21st International Requirements Engineering Conference (RE)*, pages 4–13. IEEE, 2013.

Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A Smith. 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)*, volume 2, pages 605–610, 2014.

Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. I read but don't agree: Privacy policy benchmarking using machine learning and the eu gdpr. In *Companion of the The Web Conference 2018 on The Web Conference 2018*, pages 163–166. International World Wide Web Conferences Steering Committee, 2018.

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, 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.