I Read but Don't Agree: Privacy Policy Benchmarking using Machine Learning and the EU GDPR

Welderufael B. Tesfay
Goethe University Frankfurt
Frankfurt am Main, Germany
welderufael.tesfay@m-chair.de

Peter Hofmann
Goethe University Frankfurt
Frankfurt am Main, Germany
peter_hofmann91@gmx.de

Toru Nakamura
KDDI Research Inc.
Tokyo, Japan
tr-nakamura@kddi-research.jp

Shinsaku Kiyomoto
KDDI Research Inc.
Tokyo, Japan
kiyomoto@kddi-research.jp

Jetzabel Serna
Goethe University Frankfurt
Frankfurt am Main, Germany
jetzabel.serna@m-chair.de

ABSTRACT

With the continuing growth of the Internet landscape, users share large amount of personal, sometimes, privacy sensitive data. When doing so, often, users have little or no clear knowledge about what service providers do with the trails of personal data they leave on the Internet. While regulations impose rather strict requirements that service providers should abide by, the defacto approach seems to be communicating data processing practices through privacy policies. However, privacy policies are long and complex for users to read and understand, thus failing their mere objective of informing users about the promised data processing behaviors of service providers. To address this pertinent issue, we propose a machine learning based approach to summarize the rather long privacy policy into short and condensed notes following a risk-based approach and using the European Union (EU) General Data Protection Regulation (GDPR) aspects as assessment criteria. The results are promising and indicate that our tool can summarize lengthy privacy policies in a short period of time, thus supporting users to take informed decisions regarding their information disclosure behaviors.

KEYWORDS

privacy; privacy policy; data protection regulation

ACM Reference Format:

Welderufael B. Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. I Read but Don't Agree: Privacy Policy Benchmarking using Machine Learning and the EU GDPR. In *WWW '18 Companion: The 2018 Web Conference Companion, April 23-27, 2018, Lyon, France*. ACM, New York, NY, USA, 4 pages. <https://doi.org/10.1145/3184558.3186969>

1 INTRODUCTION

Nowadays, The number of websites in the world wide web has reached over a billion [9]. These websites are accessed by almost half of the world's population on daily basis. This has resulted in Internet services fundamentally changing the way we interact with each other. These on-line interactions also result in a huge amount

of personal digital trails, which is being collected and processed by service providers, quite often with users being unaware of what data is being collected for what purpose. In the end, users are left without control of their own data, creating a huge information asymmetry in contrast to the service providers who have full access. To address this issue, regulatory bodies have devised compliance guidelines and responsibilities for service providers that aim to preserve user's informational self-determination rights. As such, privacy policies have emerged to be the de facto transparency boards that service providers use to communicate their information processing practices.

While privacy policies also function as legal agreements between service providers and their users [12], a number of studies have shown that these documents suffer from bad readability indexes [1, 8, 10]. In effect, users rarely read these documents, mainly due to their complexity, use of technical jargons and excessive length. Reading privacy policies is not only cumbersome, but it also consumes a lot of time that would cause tremendous harm to the economy. McDonald and Cranor [6] showed in 2008 that if a user was to read the privacy policy of every service she visits on the Internet, she would, on average, need 244 hours annually which is slightly more than half of the average time a user would spend on the Internet by then. This becomes even more complicated and time consuming in today's Internet services infrastructure, where the number of on-line services has more than doubled since then and data is also shared with "third parties" which have their own different privacy policies [3]. E.g., an excerpt from CNN's¹ rather long privacy policy states: "Some of our Sites contain links to or integrations with other sites such as Facebook, Twitter, LinkedIn, etc., whose information practices may be different than ours. *Visitors should consult the other sites' privacy notices* as we have no control over information that is submitted to, or collected by, these third parties."

While the Internet continues to grow, service providers and other data processors are continuously being challenged to abide by data protection regulations. One such regulation is the newly adopted European Union (EU) General Data Protection Regulation (GDPR) [4]. The GDPR takes, among other things, transparency and user consent as core components wherein the responsibilities of the data

This paper is published under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW '18 Companion, April 23-27, 2018, Lyon, France

© 2018 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC BY 4.0 License.

ACM ISBN 978-1-4503-5640-4/18/04.

<https://doi.org/10.1145/3184558.3186969>

¹<http://edition.cnn.com/2015/01/06/world/privacy-policy/index.html> Accessed Jan 2018

Figure 1: Conceptual grouping of privacy aspects

processor and the rights of the user are specified. Especially, article 13 emphasizes on the way information that must be provided to the data subject at the time their personal data is collected has to be presented. Similarly, article 12 states that information must be provided in *"a concise, transparent, intelligible and easily accessible form, using clear and plain language, in particular for any information addressed specifically to a child"*. Our motivation in this work is, thus, mainly influenced by the GDPR and newly emerging techniques of machine learning. We aim to support Internet users by simplifying privacy policies' readability. We take the requirements stated in the GDPR as a basis and provide a privacy policy benchmarking tool that considers privacy aspects and through the use of risk indicators we draw users' attention to the most relevant parts of the privacy policy. To achieve this, we introduce a categorization of data based on the interpretation of the GDPR and propose a risk based approach to privacy policy benchmarking. We use supervised machine learning techniques to classify the lengthy privacy policies and employ a three scale risk based categorization. As a result we simplify the interpretation of the privacy policy and draw users attention to the proposed risk indicators.

2 METHODOLOGY AND IMPLEMENTATION

2.1 Privacy Aspect and Risk Assessment Criteria

The privacy aspects of the EU GDPR needed to be interpreted before they could be used as assessment criteria for privacy policies. While the EU GDPR is the legal boundary of privacy policies, it does not reflect the wording and formulation that is commonly used. Research works such as Costante *et al.* [2] have defined their privacy aspects mainly based on regulations, however, others used

empirical techniques like quantitative user studies [14] or qualitative expert interviews [13]. For the definition of classification categories, we adapted the privacy aspects of the EU GDPR based on combinations of definitions of the aspects (cf. Figure 1). Article 9 is specifically used to assess how service providers promise to treat specially privacy sensitive information types.

Once the main categories were defined, a set of application privacy policies were analyzed by a group of privacy experts, as a result, three risk levels for each of the categories were determined.

2.2 Privacy Policy Corpus

Supervised machine learning techniques take as an input a training data set that must be manually labelled beforehand, so that afterwards the algorithm can generalize and classify further input data of similar nature. Thus, in order to build up the 'privacy corpus', 45 privacy policies were collected taking into consideration the 45 most accessed websites in Europe according to the results provided by Alexa.com.

In order to perform the manual labeling task, 35 participants were recruited: 68,57% male and 31,43% female, all of them with at least bachelor degree, and ages ranging between 21 and 29 years old. They were provided with a description of the task, detailed instructions and explanations for each of the eleven categories and their a potential risk level (including descriptive examples). Finally, to facilitate the labeling task, they were provided with an annotation form with drop-downs and text passages describing the categories. On average, each participant extracted 1266 text passages from a single privacy policy, resulting on 443 text passages assigned to a privacy category and classified with a risk level.

2.3 Preprocessing

PrivacyGuide applies stop words removal, stemming, and tokenization as preprocessing measures before extracting the features using TF-IDF.

2.4 Classification

The choice of a perfect Machine Learning algorithm for a specific task remains unanswered by the research community. Related studies have therefore focused on creating algorithm portfolios from which they choose the best performing ones [5]. Since, the problem space is intuitively a classification one, we defined a multi-class approach (i.e., eleven privacy aspects and three associated risks levels for each of the categories), and four well known classifiers namely: Naive Bayes, Support Vector Machine (SVM), Decision Tree (DT), and Random Forest that fulfill this criteria were implemented (cf. Figure 3). The source code can also be found in a Github repository². Certain privacy sentences may also fall into more than one category, thus we also implemented the Binary Relevance multi-label classification technique. We then compared the results using established evaluation metrics, i.e., precision, recall and finally f-measure as weighted harmonic average of both.

²<https://github.com/InfiniteCoding/PrivacyGuide>

Figure 2: PrivacyGuide Screen shots (a) Before evaluation of policy, (b) Applying to sample privacy policy.

2.5 Privacy Policy Evaluation Presentation

Two information levels have been considered: a high-level information (by default), comprised of indicators representing each privacy aspect (category) and the resulting evaluation (risk level); and, a detailed information level, providing the privacy aspect description, meaning of the risk level, and finally, the privacy policy fragment associated to the privacy aspects and its evaluation.

3 RESULTS AND SUMMARY

Before building the final model, we ran experiments on the four algorithms (cf. Section 2.4). While Naive Bayes and SVM seem to be robust in the available training and testing data, Decision Tree and Random Forest showed low precision. This could be due to both the size of the training data and presence of some "noise" in it.

Naive Bayes and SVM have relatively similar performance. However, SVM takes more time compared to Naive Bayes, thus, the final model and the results indicated in Table 1 and Figure 2 (b) are based on the Naive Bayes classifier. Furthermore, Naive Bayes performed even better in the risk classification phase resulting in average f-measure of 90% across the aspects.

We also observe that, in contrast to the reports in [6] as to how long it would take an average user to finish reading a privacy policy, PrivacyGuide summarizes the privacy policies in a short period of time, helping users save a significant amount of time. Nevertheless, the time that is needed for loading a privacy policy can be affected by a variety of external reasons like: the websites' response time, the Internet connection, the bandwidth, the amount of content that is displayed on the same page and many more. Further observations show that despite the requirements in the GDPR for data breach notifications, our assessments showed that none of the Internet services we analyzed have plans for privacy breach notification (cf. Table 1 data breach notification results).

In summary, the benefits of the PrivacyGuide results are two fold: 1) supporting users in assessing or interpreting privacy policies; 2) providing actionable insights for service providers with respect

Table 1: Classification performance

Privacy Aspect	Recall (%)	Precision (%)	F1 (%)
Data collection	79	78	79
Protection of children	91	1.0	95
Third-party sharing	76	66	71
Data security	71	88	78
Data retention	68	59	63
Data aggregation	64	63	64
Control of data	63	79	70
Privacy settings	69	60	64
Account deletion	58	68	63
Policy changes	87	91	89
Data breach notification	0	0	0

to their privacy policies. Moreover, it can also trigger discussions among data protection officers, lawyers, privacy lobbyists, and other interested entities.

4 DEMONSTRATION, LIMITATIONS AND FUTURE WORKS

Demonstration: During the demo session, we will showcase how the tool can extract privacy policy content from a given Uniform Resource Locator (URL). Furthermore, we will also demonstrate how the privacy policy is benchmarked using machine learning based on the EU GDPR aspects as criteria. The tool presents a risk-based, condensed and easy to understand summary of lengthy privacy policies in short period of time. We use user-friendly icons and colors to indicate the associated risk and draw user's attention. A screen cast of the tool is available on Youtube³. The tool is tested to work on Macbook Pro OS X Yosemite 2.3 GHz Intel Core i7, and Microsoft Surface Pro 3 device with Windows 10 Pro OS, 8GB RAM,

³<https://www.youtube.com/watch?v=JJC53PlfQIk&feature=youtu.be>

Figure 3: Process flows in the prototype development

Intel Core i5-4300U CPU @ 1.90 GHz (Boost: 2.50 Ghz) and ADSL 2+ Internet connection. It requires a stable Internet connection to extract the privacy policy from target web services.

Limitations: While this paper takes important steps towards building a stable tool that summarizes lengthy privacy policies in a short period of time, we also acknowledge that there are certain limitations, particularly: 1) PrivacyGuide works for privacy policies in English language only; 2) at its current phase, it works as stand alone tool, which in itself might have privacy gains, but it also presents some challenges with respect to model updating; 3) PrivacyGuide provides important insights about service provider's promises, however, as with any machine learning model, there are false positives and false negatives in the classification process. Thus, the tool should be taken as supportive but non-binding input for decisions.

Future works: We envision to enhance the prediction accuracy of PrivacyGuide by enlarging the privacy policy corpus and also experimenting with machine learning techniques that can better grasp the semantics of the privacy policy text, e.g., using word embeddings and deep learning techniques. Furthermore, we also plan to run usability and user adoption studies. Last but not least, we envisage to build a holistic information privacy enhancing framework composed of (1) PrivacyGuide; (2) default privacy preference setting prediction engine [7]; and (3) unintended information disclosure detection and quantification component [11].

ACKNOWLEDGMENTS

The authors would like to thank the text annotation task participants. The work is partially supported by the H2020 EU Project CREDENTIAL under Grant No.: 653454.

REFERENCES

- [1] Rochelle A Cadogan. 2011. An imbalance of power: the readability of internet privacy policies. *Journal of Business & Economics Research (JBER)* 2, 3 (2011).
- [2] Elisa Costante, Yuanhao Sun, Milan Petković, and Jerry den Hartog. 2012. A machine learning solution to assess privacy policy completeness: (short paper). In *Proceedings of the 2012 ACM workshop on Privacy in the electronic society*. ACM, 91–96.
- [3] Lorrie Faith Cranor. 2012. Necessary but not sufficient: Standardized mechanisms for privacy notice and choice. *J. on Telecomm. & High Tech. L.* 10 (2012), 273.
- [4] EU. 2016. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (2016).
- [5] Lars Kotthoff, Ian P Gent, and Ian Miguel. 2011. A preliminary evaluation of machine learning in algorithm selection for search problems. In *Fourth Annual Symposium on Combinatorial Search*.
- [6] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. *ISJLP* 4 (2008), 543.
- [7] Toru Nakamura, Shinsaku Kiyomoto, Welderufael B Tesfay, and Jetzabel Serna. 2016. Easing the Burden of Setting Privacy Preferences: A Machine Learning Approach. In *International Conference on Information Systems Security and Privacy*. Springer, 44–63.
- [8] Robert W Proctor, M Athar Ali, and Kim-Phuong L Vu. 2008. Examining usability of web privacy policies. *Intl. Journal of Human-Computer Interaction* 24, 3 (2008), 307–328.
- [9] Internet Live Stats. 2017. Total number of websites. <http://www.internetlivestats.com/total-number-of-websites/>. (2017). Accessed: December 2017.
- [10] Ali Sunyaev, Tobias Dehling, Patrick L Taylor, and Kenneth D Mandl. 2014. Availability and quality of mobile health app privacy policies. *Journal of the American Medical Informatics Association* 22, e1 (2014), e28–e33.
- [11] Welderufael B Tesfay and Jetzabel Serna-Olvera. 2016. Towards User-Centered Privacy Risk Detection and Quantification Framework. In *New Technologies, Mobility and Security (NTMS), 2016 8th IFIP International Conference on*. IEEE, 1–5.
- [12] 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. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *ACL (1)*.
- [13] Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. [n. d.]. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. ([n. d.]).
- [14] Sebastian Zimmeck and Steven M Bellovin. 2014. Privee: An Architecture for Automatically Analyzing Web Privacy Policies.. In *USENIX Security Symposium*. 1–16.