Automated Privacy Policy Analysis

Subjects: Computer Science, Interdisciplinary Applications
Contributor: Veronika Belcheva, Tatiana Ermakova, Benjamin Fabian

Privacy policies are the main method for informing Internet users of how their data are collected and shared. Automated privacy policy analysis, including machine learning methods, has grown in popularity during the last decade. The main goal is to grant users a better understanding of how their data are used and help them make informed decisions regarding their privacy.

Keywords: privacy policy ; machine learning ; automated privacy policy analysis ; natural language processing

1. Introduction

Natural language privacy policies serve as the primary means of disclosing data practices to consumers, providing them with crucial information about what data are collected, analyzed, and how they will be kept private and secure. By reading these policies, users can enhance their awareness of data privacy and better manage the risks associated with extensive data collection. However, for privacy policies to be genuinely useful, they must be easily comprehensible to the majority of users. Lengthy and vague policies fail to effectively inform the average user, rendering them ineffective in ensuring data privacy awareness.

Privacy policies are often excessive in length, requiring a substantial amount of time to read through. Estimates show that the average Internet user would spend around 400 h per year reading all encountered privacy terms [1]. This time investment may deter users from thoroughly reviewing policies, leading them to hurriedly click the “I agree” button without fully understanding the implications.

Addressing the significance of readability and privacy regulations, such as General Data Protection Regulation (GDPR), mandate that privacy policies should be concise, easy to understand, and written in plain language. Additionally, the California Consumer Privacy Act (CCPA) emphasizes the need to present policies in a clear and straightforward manner, avoiding technical or legal jargon.

To enhance clarity and conciseness, the GDPR guidelines recommend the use of active voice instead of passive voice in writing [2]. The active voice directs the reader’s attention to the performer of the action, reducing ambiguity and making the text more straightforward.

Additionally, policies become less comprehensible due to ambiguity, which occurs when a statement lacks clarity and can be interpreted in multiple ways. The use of imprecise language in a privacy policy hinders the clear communication of the website’s actual data practices. The presence of language qualifiers like “may”, “might”, “some”, and “often” contributes to ambiguity, as noted by the European Commission’s GDPR guidelines [2]. Recent research suggests an increasing use of terms such as “may include” and “may collect” in privacy policies, which may result in policies becoming more ambiguous over time [3].

2. Automated Privacy Policy Analysis

2.1. Privacy Policy Datasets

Various privacy policy datasets have been made accessible to researchers (see **Table 1**), with the Usable Privacy Policy Project [4] playing a significant role in this regard. Their OPP-115 corpus [5] contains annotated segments from 115 website privacy policies, enabling advanced machine learning research and automated analysis. Another dataset from the same project is the OptOutChoice-2020 corpus [6], which includes privacy policy sentences with labeled opt-out choices types. PolicyIE [7] offers a more recent dataset with annotated data practices, including intent classification and slot filling, based on 31 web and mobile app privacy policies. Nokhbeh Zaeem and Barber [8] created a corpus of over 100,000 privacy policies, categorized into 15 website categories, utilizing the DMOZ directory. PrivaSeer [9] is a privacy policy

dataset and search engine containing approximately 1.4 million website privacy policies. It was built using web crawls from 2019 and 2020, utilizing URLs from “Common Crawl” and the “Free Company Dataset”. Finally, Amos et al. [3] released the Princeton-Leuven Longitudinal Corpus of Privacy Policies, a large-scale longitudinal corpus spanning two decades, consisting of one million privacy policy snapshots from around 130,000 websites, enabling the study of trends and changes over time.

Table 1. Publicly available privacy policy datasets.

Dataset	# Policies	# Websites	Timeframe	Labeling
OPP-115	**115**	**115**	**2015**	**Yes**
OptOutChoice-2020	**236**	**236**	**-**	**Yes**
PolicyIE	**400**	**400 (websites + apps)**	**2019**	**Yes**
DMOZ-based Corpus	**117,502**	**-**	**2020**	**No**
PrivaSeer	**1,005,380**	**995,475**	**2019**	**No**
Princeton-Leuven Corpus	**910,546**	**108,499**	**1997–2019**	**No**

2.2. Classification and Information Extraction

Classification and information extraction from privacy policies have been widely explored using machine learning techniques. Kaur et al. [10] employed unsupervised methods such as Latent Dirichlet Allocation (LDA) and term frequency to analyze keywords and content in 2000 privacy policies. Supervised learning approaches have also been utilized, including classifiers trained on the OPP-115 dataset. Audich et al. [11] compared the performance of supervised and unsupervised algorithms to label policy segments, while Kumar et al. [12] trained privacy-specific word embeddings for improved results. Deep learning models like CNN, BERT, and XLNET have further enhanced their classification performance [13][14][15]. Bui et al. [16] tackled the extraction of personal data objects and actions using a BLSTM model with contextual word embeddings. Alabduljabbar et al. [17][18] proposed a pipeline called TLDR for the automatic categorization and highlighting of policy segments, enhancing user comprehension. Extracting opt-out choices from privacy policies has also been studied [6][19][20]. In the field of summarization, Keymanesh et al. [21] introduced a domain-guided approach for privacy policy summarization, focusing on labeling privacy topics and extracting the riskiest content. Several studies have worked on developing automated privacy policy question-answering assistants [22][23][24].

Furthermore, the PrivacyGLUE [25] benchmark was proposed to address the lack of comprehensive benchmarks specifically designed for privacy policies. The benchmark includes the performance evaluations of transformer language models and emphasizes the importance of in-domain pre-training for privacy policies.

2.3. Privacy Policy Applications for Enhancing Users’ Comprehension

Applications enhancing the comprehension of privacy policies have been developed to provide users with useful and visually appealing presentations of policy information. PrivacyGuide [26] employs a two-step multi-class approach, identifying relevant privacy aspects and predicting risk levels using a trained model on a labeled dataset. The user interface utilizes colored icons to indicate risk levels. Polisis [27][28] combines a summarizing tool, policy comparison tool, and chatbot. The query system employs neural network classifiers trained on the OPP-115 dataset and privacy-specific language models. PrivacyCheck is a browser extension that extracts 10 privacy factors and displays their risk levels through icons and text snippets [29][30][31][32]. Opt-Out Easy is another browser extension that utilizes the OptOutChoice-2020 dataset to identify and present opt-out choices to users during web browsing [6][33].

2.4. Regulatory Impact

User research has also focused on evaluating privacy policies for regulatory compliance, particularly in response to the implementation of General Data Protection Regulation (GDPR) in Europe. The tool Claudette detects unfair clauses and evaluates privacy policy compliance with GDPR [34][35]. KnIGHT (“Know your rIGHTs”) utilizes semantic text matching to map policy sentences to GDPR paragraphs [36]. Cejas et al. [37] and Qamar et al. [38] leveraged NLP and supervised machine learning to identify GDPR-relevant information in policies and assess their compliance. Similarly, Sánchez et al. [39] used manual annotations and machine learning to tag policies based on GDPR goals, offering both aggregated scores and fine-grained ratings for better understanding. Degeling et al. [40] and Linden et al. [41] examined the effects of GDPR on privacy policies through longitudinal analysis, observing updates and changes in policy length and disclosures. Zaeem

and Barber [42] compared pre- and post-GDPR policies using PrivacyCheck, highlighting deficiencies in transparency and explicit data processing disclosures. Libert [43] developed an automated approach to audit third-party data sharing in privacy policies.

2.5. Comprehensibility of Privacy Policies

Studies on privacy policy comprehensibility have examined deficiencies in readability, revealing that privacy policies are difficult to read and demonstrating correlations between readability measures [44][45]. Furthermore, researchers have examined the changes in length and readability of privacy policies over time [1][3].

Other scholars have studied ambiguous content in privacy policies. Kaur et al. [10] and Srinath et al. [9] analyzed the use of ambiguous words in a corpus of 2000 policies. Furthermore, Kotal et al. [46] studied the ambiguity in the OPP-115 dataset and showed that ambiguity negatively affects the ability to automatically evaluate privacy policies. Srinath et al. [9] reported on privacy policy length and the use of vague words in their PrivaSeer corpus of policies. Lebanoff and Liu [47] investigated the detection of vague words and sentences using deep neural networks.

2.6. Mobile Applications

The research community has also examined privacy policies in the context of mobile applications, establishing several corpora of mobile app privacy policies [48][49]. Those policies are well-suited for compliance analysis, because they are studied along with the app code and the traffic generated by the app [49][50].

References

1. Wagner, I. Privacy Policies Across the Ages: Content and Readability of Privacy Policies 1996–2021. Technical Report. arXiv 2022, arXiv:2201.08739.
2. Article 29 Working Party: Guidelines on Transparency under Regulation 2016/679. Available online: https://ec.europa.eu/newsroom/article29/items/622227/en (accessed on 15 November 2023).
3. Amos, R.; Acar, G.; Lucherini, E.; Kshirsagar, M.; Narayanan, A.; Mayer, J. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In Proceedings of the Web Conference 2021, Ljubljana, Slovenia, 19–23 April 2021; pp. 2165–2176.
4. Usable Privacy Policy Project. Available online: https://usableprivacy.org/ (accessed on 18 June 2023).
5. Wilson, S.; Schaub, F.; Dara, A.A.; Liu, F.; Cherivirala, S.; Giovanni Leon, P.; Schaarup Andersen, M.; Zimmeck, S.; Sathyendra, K.M.; Russell, N.C.; 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), Berlin, Germany, 7–12 August 2016; Association for Computational Linguistics: Berlin, Germany, 2016; pp. 1330–1340.
6. Bannihatti Kumar, V.; Iyengar, R.; Nisal, N.; Feng, Y.; Habib, H.; Story, P.; Cherivirala, S.; Hagan, M.; Cranor, L.; Wilson, S.; et al. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In Proceedings of the Web Conference 2020, Virtural, 20–24 April 2020; ACM: Taipei, Taiwan, 2020; pp. 1943–1954.
7. Ahmad, W.U.; Chi, J.; Le, T.; Norton, T.; Tian, Y.; Chang, K.W. Intent Classification and Slot Filling for Privacy Policies. arXiv 2021, arXiv:2101.00123.
8. Nokhbeh Zaeem, R.; Barber, K.S. A Large Publicly Available Corpus of Website Privacy Policies Based on DMOZ. In Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy, Virtual, 26–28 April 2021; ACM: New York, NY, USA, 2021; pp. 143–148.
9. Srinath, M.; Sundareswara, S.N.; Giles, C.L.; Wilson, S. PrivaSeer: A Privacy Policy Search Engine. In Proceedings of the Web Engineering; Brambilla, M., Chbeir, R., Frasincar, F., Manolescu, I., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2021; pp. 286–301.
10. Kaur, J.; Dara, R.A.; Obimbo, C.; Song, F.; Menard, K. A comprehensive keyword analysis of online privacy policies. Inf. Secur. J. Glob. Perspect. 2018, 27, 260–275.
11. Audich, D.; Dara, R.; Nonnecke, B. Privacy Policy Annotation for Semi-Automated Analysis: A Cost-Effective Approach. In Trust Management XII. IFIPTM 2018. IFIP Advances in Information and Communication Technology; Springer: Cham, Switzerland; Toronto, ON, Canada, 2018; pp. 29–44.
12. Kumar, V.B.; Ravichander, A.; Story, P.; Sadeh, N. Quantifying the Effect of In-Domain Distributed Word Representations: A Study of Privacy Policies. In AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence

and Language Technologies. 2019. Available online: https://usableprivacy.org/static/files/kumar_pal_2019.pdf (accessed on 18 June 2023).

13. Liu, F.; Wilson, S.; Story, P.; Zimmeck, S.; Sadeh, N. Towards Automatic Classification of Privacy Policy Text. Technical Report, CMU-ISR-17-118R, Institute for Software Research and Language Technologies Institute, School of Computer Science, Carnegie Mellon University, 2018. Available online: http://reports-archive.adm.cs.cmu.edu/anon/isr2017/CMU-ISR-17-118R.pdf (accessed on 18 June 2023).
14. Mousavi, N.; Jabat, P.; Nedelchev, R.; Scerri, S.; Graux, D. Establishing a Strong Baseline for Privacy Policy Classification. In Proceedings of the IFIP International Conference on ICT Systems Security and Privacy Protection, Maribor, Slovenia, 21–23 September 2020.
15. Mustapha, M.; Krasnashchok, K.; Al Bassit, A.; Skhiri, S. Privacy Policy Classification with XLNet (Short Paper). In Data Privacy Management, Cryptocurrencies and Blockchain Technology; Garcia-Alfaro, J., Navarro-Arribas, G., Herrera-Joancomarti, J., Eds.; Springer International Publishing: Cham, Switzerland, 2020; Volume 12484, pp. 250–257.
16. Bui, D.; Shin, K.G.; Choi, J.M.; Shin, J. Automated Extraction and Presentation of Data Practices in Privacy Policies. Proc. Priv. Enhancing Technol. 2021, 2021, 88–110.
17. Alabduljabbar, A.; Abusnaina, A.; Meteriz-Yildiran, U.; Mohaisen, D. Automated Privacy Policy Annotation with Information Highlighting Made Practical Using Deep Representations. In Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual, 15–19 November 2021; CCS '21. Association for Computing Machinery: New York, NY, USA, 2021; pp. 2378–2380.
18. Alabduljabbar, A.; Abusnaina, A.; Meteriz-Yildiran, U.; Mohaisen, D. TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights. In Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society, Virtual, 15 November 2021; ACM: New York, NY, USA, 2021; pp. 103–118.
19. Sathyendra, K.M.; Schaub, F.; Wilson, S.; Sadeh, N.M. Automatic Extraction of Opt-Out Choices from Privacy Policies. In AAAI Fall Symposia, 2016, Association for the Advancement of Artificial Intelligence. 2016. Available online: https://api.semanticscholar.org/CorpusID:32896562 (accessed on 18 June 2023).
20. Sathyendra, K.M.; Wilson, S.; Schaub, F.; Zimmeck, S.; Sadeh, N. Identifying the Provision of Choices in Privacy Policy Text. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 7–11 September 2017; Association for Computational Linguistics: Copenhagen, Denmark, 2017; pp. 2774–2779.
21. Keymanesh, M.; Elsner, M.; Parthasarathy, S. Toward Domain-Guided Controllable Summarization of Privacy Policies. In Proceedings of the 2020 Natural Legal Language Processing (NLLP) Workshop, Virtual Event/San Diego, CA, USA, 24 August 2020; ACM: New York, NY, USA, 2020; pp. 18–24.
22. Ravichander, A.; Black, A.W.; Wilson, S.; Norton, T.; Sadeh, N. Question Answering for Privacy Policies: Combining Computational and Legal Perspectives. arXiv 2019, arXiv:1911.00841.
23. Ahmad, W.U.; Chi, J.; Tian, Y.; Chang, K.W. PolicyQA: A Reading Comprehension Dataset for Privacy Policies. arXiv 2020, arXiv:2010.02557.
24. Keymanesh, M.; Elsner, M.; Parthasarathy, S. Privacy Policy Question Answering Assistant: A Query-Guided Extractive Summarization Approach. arXiv 2021, arXiv:2109.14638.
25. Shankar, A.; Waldis, A.; Bless, C.; Andueza Rodriguez, M.; Mazzola, L. PrivacyGLUE: A Benchmark Dataset for General Language Understanding in Privacy Policies. Appl. Sci. 2023, 13, 3701.
26. Tesfay, W.B.; Hofmann, P.; Nakamura, T.; Kiyomoto, S.; Serna, J. 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, Tempe, AZ, USA, 19–21 March 2018; IWSPA '18. Association for Computing Machinery: New York, NY, USA, 2018; pp. 15–21.
27. Harkous, H.; Fawaz, K.; Lebret, R.; Schaub, F.; Shin, K.G.; Aberer, K. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In Proceedings of the 27th USENIX Security Symposium, Baltimore, MD, USA, 15–17 August 2018; USENIX Association: Berkeley, CA, USA, 2018; pp. 531–548.
28. PriBOT. Available online: https://pribot.org/ (accessed on 24 June 2023).
29. Zaeem, R.N.; German, R.L.; Barber, K.S. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. ACM Trans. Internet Technol. 2018, 18, 53:1–53:18.
30. Nokhbeh Zaeem, R.; Anya, S.; Issa, A.; Nimergood, J.; Rogers, I.; Shah, V.; Srivastava, A.; Barber, K.S. PrivacyCheck v2: A Tool that Recaps Privacy Policies for You. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, Virtual, 19–23 October 2020; Association for Computing Machinery: New York, NY, USA, 2020. CIKM '20. pp. 3441–3444.

31. Nokhbeh Zaeem, R.; Ahbab, A.; Bestor, J.; Djadi, H.H.; Kharel, S.; Lai, V.; Wang, N.; Barber, K.S. PrivacyCheck v3: Empowering Users with Higher-Level Understanding of Privacy Policies. In Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining, Virtual, 21–25 February 2022; WSDM ’22. Association for Computing Machinery: New York, NY, USA, 2022; pp. 1593–1596.
32. Privacy Lab|Center for Identity. Available online: https://identity.utexas.edu/privacy-lab (accessed on 24 June 2023).
33. Opt-Out Easy. Available online: https://optouteasy.isr.cmu.edu/ (accessed on 24 June 2023).
34. Contissa, G.; Docter, K.; Lagioia, F.; Lippi, M.; Micklitz, H.W.; Pałka, P.; Sartor, G.; Torroni, P. Claudette Meets GDPR: Automating the Evaluation of Privacy Policies Using Artificial Intelligence; SSRN Scholarly; Social Science Research Network: Rochester, NY, USA, 2018.
35. Liepina, R.; Contissa, G.; Drazewski, K.; Lagioia, F.; Lippi, M.; Micklitz, H.; Palka, P.; Sartor, G.; Torroni, P. GDPR Privacy Policies in CLAUDETTE: Challenges of Omission, Context and Multilingualism. In Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2019), Montreal, QC, Canada, 21 June 2019.
36. Mousavi, N.; Scerri, S.; Lehmann, J. KnIGHT: Mapping Privacy Policies to GDPR. In Knowledge Engineering and Knowledge Management; Faron Zucker, C., Ghidini, C., Napoli, A., Toussaint, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; Volume 11313, pp. 258–272.
37. Cejas, O.A.; Abualhaija, S.; Torre, D.; Sabetzadeh, M.; Briand, L. AI-enabled Automation for Completeness Checking of Privacy Policies. IEEE Trans. Softw. Eng. 2021, 48, 4647–4674.
38. Qamar, A.; Javed, T.; Beg, M.O. Detecting Compliance of Privacy Policies with Data Protection Laws. arXiv 2021, arXiv:2102.12362.
39. Sánchez, D.; Viejo, A.; Batet, M. Automatic Assessment of Privacy Policies under the GDPR. Appl. Sci. 2021, 11, 1762.
40. Degeling, M.; Utz, C.; Lentzsch, C.; Hosseini, H.; Schaub, F.; Holz, T. We Value Your Privacy … Now Take Some Cookies: Measuring the GDPR’s Impact on Web Privacy. In Proceedings of the 2019 Network and Distributed System Security Symposium, San Diego, CA, USA, 24–27 February 2019.
41. Linden, T.; Khandelwal, R.; Harkous, H.; Fawaz, K. The Privacy Policy Landscape After the GDPR. arXiv 2019, arXiv:1809.08396.
42. Zaeem, R.N.; Barber, K.S. The Effect of the GDPR on Privacy Policies: Recent Progress and Future Promise. ACM Trans. Manag. Inf. Syst. 2020, 12, 2:1–2:20.
43. Libert, T. An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies. In Proceedings of the 2018 World Wide Web Conference on World Wide Web-WWW ’18, Lyon, France, 23–27 April 2018; pp. 207–216.
44. Fabian, B.; Ermakova, T.; Lentz, T. Large-scale readability analysis of privacy policies. In Proceedings of the International Conference on Web Intelligence (WI ’17), Leipzig, Germany, 23–26 August 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 18–25.
45. Ermakova, T.; Fabian, B.; Babina, E. Readability of Privacy Policies of Healthcare Websites. In Proceedings of the 12th International Conference on Wirtschaftsinformatik, Osnabrück, Germany, 4–6 March 2015.
46. Kotal, A.; Joshi, A.; Pande Joshi, K. The Effect of Text Ambiguity on creating Policy Knowledge Graphs. In Proceedings of the 2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), New York City, NY, USA, 30 September–3 October 2021; IEEE: New York City, NY, USA, 2021; pp. 1491–1500.
47. Lebanoff, L.; Liu, F. Automatic Detection of Vague Words and Sentences in Privacy Policies. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 31 October–4 November 2018; pp. 3508–3517.
48. Zimmeck, S.; Story, P.; Smullen, D.; Ravichander, A.; Wang, Z.; Reidenberg, J.; Cameron Russell, N.; Sadeh, N. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. Proc. Priv. Enhancing Technol. 2019, 2019, 66–86.
49. Story, P.; Zimmeck, S.; Ravichander, A.; Smullen, D.; Wang, Z.; Reidenberg, J.; Russell, N.; Sadeh, N. Natural Language Processing for Mobile App Privacy Compliance. In Proceedings of the PAL: Privacy-Enhancing Artificial Intelligence and Language Technologies AAAI Spring Symposium, Palo Alto, CA, USA, 25–27 March 2019.
50. Hashmi, S.S.; Waheed, N.; Tangari, G.; Ikram, M.; Smith, S. Longitudinal Compliance Analysis of Android Applications with Privacy Policies. arXiv 2021, arXiv:2106.10035.