Automatic Retrieval of Privacy Factors from IoMT Policies: ML and Custom NER Approach

Nyteisha Bookert and Mohd Anwar
North Carolina Agricultural and Technical State University
manwar@ncat.edu

***Abstract*—Patient-generated health data is growing at an unparalleled rate due to advancing technologies (e.g., the Internet of Medical Things, 5G, artificial intelligence) and increased consumer transactions. The influx of data has offered life-altering solutions. Consequently, the growth has created significant privacy challenges. A central theme to mitigating risks is promoting transparency and notifying stakeholders of data practices through privacy policies. However, natural language privacy policies have several limitations, such as being difficult to understand (by the user), lengthy, and having conflicting requirements. Yet they remain the de facto standard to inform users of privacy practices and how organizations follow privacy regulations. We developed an automated process to evaluate the appropriateness of combining machine learning and custom named entity recognition techniques to extract IoMT-relevant privacy factors in the privacy policies of IoMT devices. We employed machine learning and the natural language processing technique of named entity recognition to automatically analyze a corpus of policies and specifications to extract privacy-related information for the IoMT device. Based on the natural language analysis of policies, we provide fine-grained annotations that can help reduce the manual and tedious process of policy analysis and aid privacy engineers and policy makers in developing suitable privacy policies.**

I. Introduction

Data protection laws and regulations are increasing globally to protect users and their data from the concerns of emerging technologies. Most data protection laws require transparency into how an organization collects, shares, and handles data or data practices [2]. The data practices are commonly placed into privacy policies. Privacy policies should be clear and conspicuous for end-users to understand an organization's privacy posture and learn how to exercise their rights. Regulators leverage privacy policies to audit and verify an organization's compliance with data protection laws and regulations. For example, the Federal Trade Commission served several enforcement actions for companies failing to maintain accurate and consistent privacy policies. Google and YouTube agreed to pay $170 million to settle allegations of violating the Children's Online Privacy Protection Act Rule (COPPA) for failing to provide a COPPA-specific notice for their practices [16]. However, as emerging technologies continue to increase the number of systems, the volume and complexity of privacy policies increase.

Privacy policies serve an essential role as the primary option for users, especially lay users, to learn more about an organization's data practices. However, users often do not read privacy policies [24] because they are difficult to comprehend [20], [21], vague [8], and time-consuming to read [6], [29]. Several efforts have gone into creating tools to estimate the risk for users [43], [51], [53], evaluate compliance [4], [22], [54], and even seek potential alternatives [25], [34], [42]. Furthermore, regulators require an automated tool to conduct large-scale analyses to help identify potential non-compliant systems for more detailed examination. Accordingly, there is a need for scalable privacy policy analysis techniques and solutions to reduce the manual and tedious work to save time and reduce errors. Yet there are limited tools available to assist in analyzing privacy policies.

Privacy policy analysis includes policies for websites [11], [12], [43], [47], [50], [51], [53], mobile applications [5], [38], [45], [48], [55], and Internet of Things devices [26], [30], [31], [40], [49]. The privacy policies studies are usually general, capturing various fields [3], [17], [23], [39], but a few examine how privacy policies handle more sensitive data (e.g., financial and health data) [10], [14], [18], [19], [37], [41]. However, there is limited privacy policy analysis work [26], [49] on the intersection of healthcare and the Internet of Things (or IoMT). Therefore, our study builds a tool to create IoMT-specific fine-grained annotations to identify relevant information to aid users, including privacy policy analysis researchers and developers. We evaluate the appropriateness of combining machine learning and custom named entity recognition techniques to reveal IoMT-specific privacy-related topics in privacy policies.

The remainder of the paper is organized as follows. Section II briefly introduces the Internet of Medical Things. In Section III, we provide an overview of our methodology. We present our results in Section IV and discuss our findings in Section V. Then, the related work is discussed in Section VI, and future work is in Section VII. Finally, we draw our conclusions and closing remarks in Section VIII.

II. Internet of Medical Things Background

Medical information was once limited to oral stories provided by individuals and data collected in a medical facility. Emerging technologies, such as the Internet of Medical Things, are transforming healthcare. Internet of Medical Things (IoMT) describes connected devices and applications that collect, process, and analyze health, medical, fitness, and wellness-related data [27], [44]. Devices are found in traditional medical facilities, hospitals, clinics, and laboratories, and non-traditional settings, such as schools, homes, and vehicles

Symposium on Usable Security and Privacy (USEC) 2023
27 February 2023, San Diego, CA, USA
ISBN 1-891562-91-6
https://dx.doi.org/10.14722/usec.2023.235685
www.ndss-symposium.org, https://www.usablesecurity.net/USEC/

(ambulances). The devices and applications produce various types of data ranging from sleep logs to medical images to location to create a vast amount of data. The patient-generated health data is increasing at an astounding rate. The data can supplement clinical trials to help provide valuable insight into patients' day-to-day activities.

Patient-generated health data is one of the leading privacy risks in upcoming years [32]. Health-related data is the most breached data [1], [35]. Furthermore, it is unclear whether current legislation in the United States, where most IoMT key players are headquartered, adequately protects patient-generated health data and its usage [7], [32]. For example, IoMT devices collect health-related information, behavioral data, and variables that can become unique identifiers. Brain signals can be unique for each individual [15]. However, they are not covered under HIPAA. As the benefits of patient-generated health data would significantly improve healthcare, public health, and emergency preparedness operations it is imperative to understand how organizations handle this data.

III. Methodology

In this section, we discuss the method to design and implement a two-layer classification model with multi-label classification and a custom named entity (NER) model to analyze the privacy policies of IoMT devices. First, we adopt frameworks and a popular dataset to create the data for the models. Then, we build the classification tool. Next, we further refine the data to develop the NER model. Finally, we discuss the performance metrics to evaluate the two-layer classification model.

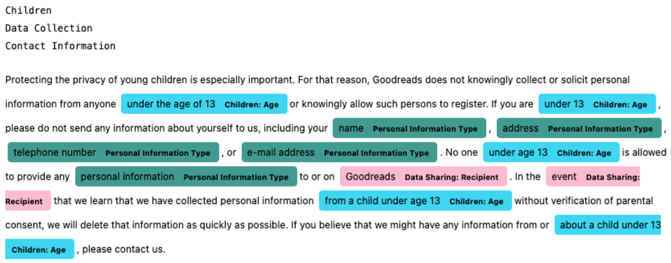
A. Creating the data

Several privacy policy corpora exist [3], [39], [46], [54]. We use the Online Privacy Policies, set of 115 Corpus (OPP-115) [46] because it contains labeled privacy policies for diverse privacy-related factors. Legal and privacy experts annotated the privacy policies for 115 websites from different sectors (e.g., Health, Business, Home) according to 10 privacy-related categories and more than 120 attributes. While OPP-115 does not contain IoMT-specific privacy policies, we opted to select this corpus instead of creating our own due to the annotation quality and comprehensive privacy-related factors.

However, since we are focusing on IoMT devices, we used the seven categories identified in the Privacy Policy Assessment Questionnaire (PPAQ) [9] that are specific to IoMT devices as our labels. We mapped the PPAQ labels to the OPP-115 categories to begin building our dataset, as shown in Table I. The OPP-115 attributes have required and optional values. Some PPAQ categories do not map directly to an OPP-115 category but a required value. For example, OPP-115 does not have a children category. Instead, the international and specific audience category has a required attribute, audience type, with a children option. In those cases, we use the attribute to collect the relevant data practices.

B. Multi-label Classification

We seek to determine whether the PPAQ topics are present in a privacy policy. A paragraph in a privacy policy may be related to one or more categories. For example, a paragraph may describe the *data collection* practices for *children* and provide guardians with *contact information*, as shown in Figure 1. Therefore, the first layer of our model is a multi-label classification problem. We preprocessed the data by removing HTML characters, punctuation, special characters, and non-alphabetic characters. We removed stop words and returned the stem for each remaining word. The corpus is split into 80% training and 20% testing. We vectorized the data using the term frequency-inverse document frequency and configured the vector with unigrams and bigrams. We consider several classification models to compare using a set of performance metrics. The following algorithms were used: Gradient Boosting, Bagging, Naïve Bayes, and Linear Support Vector Classification.



```
Children
Data Collection
Contact Information
```

Protecting the privacy of young children is especially important. For that reason, Goodreads does not knowingly collect or solicit personal information from anyone under the age of 13 **Children: Age** or knowingly allow such persons to register. If you are under 13 **Children: Age** , please do not send any information about yourself to us, including your name **Personal Information Type** , address **Personal Information Type** , telephone number **Personal Information Type** , or e-mail address **Personal Information Type** . No one under age 13 **Children: Age** is allowed to provide any personal information **Personal Information Type** to or on Goodreads **Data Sharing: Recipient** . In the event **Data Sharing: Recipient** that we learn that we have collected personal information from a child under age 13 **Children: Age** without verification of parental consent, we will delete that information as quickly as possible. If you believe that we might have any information from or about a child under 13 **Children: Age** , please contact us.

Fig. 1. Privacy policy paragraph discussing multiple privacy-related topics

C. Custom NER Model

We seek to provide more granular information from the privacy policy. Therefore, the second layer of our model is a custom named entity recognition (NER) model. NER is an information extraction task that locates and classifies key information in unstructured text. However, privacy policies are legally binding text and have a different structure than most documents. Therefore, we build a custom NER model with the following eleven entities: Address, Children: Age, Data Retention: Period, Data Sharing: Recipient, Date, Does/Does Not, Email, Personal Information Type, Phone Number, URL, and User Choice. We leverage the attributes of OPP-115 to create the annotated dataset for the custom NER model, shown in Table I.

The OPP-115 annotations range from one word to paragraphs requiring further processing. The data is processed to identify the entities in the text. The HTML characters were removed from each segment. We used spaCy rule-based matching, spaCy properties, and custom-based scripts to automatically refine the annotations. We do not claim this is a gold standard set of fine-grained annotations. Our goal is to show the usefulness of fine-grained annotations in privacy policy analysis.

From the annotated data, we constructed our model. We defined a blank model in English. The OPP-115 annotations contain the start and end indices for the attributes. However, there are overlaps in the labels. Therefore, we filter the spans to ensure that one label applies to a span. We randomly divided the corpus into 70% training and 30% testing. Next, we trained and evaluated our model using the built-in train command in spaCy.

TABLE I. MAPPING BETWEEN PPAQ CATEGORIES AND THE OPP-115 CATEGORIES. * DENOTES THE ATTRIBUTES FOR THE CUSTOM NER MODEL DATASET

PPAQ Category	OPP-115 Category	OPP-115 Attribute	OPP-115 Attribute Values
Data Collection	First Party Collection/Use	Does/Does Not*	All
		Identifiability	All
		Action First-Party	All
		Personal Information Type*	All
		Purpose	All
Data Sharing	Third Party Sharing/Collection	Does/Does Not*	All
		Identifiability	All
		Third Party Entity*	All
		Action Third Party	All
		Personal Information Type*	All
		Purpose	All
Data Retention	Data Retention	Retention Period*	All
		Retention Purpose	All
		Personal Information Type*	All
Data Security	Data Security	Security Measure	All
User Choice	User Access, Edit and Deletion	Access Type*	All
		Access Scope	All
		Choice Type*	All
		Choice Scope	All
		Personal Information Type*	All
		Purpose	All
Children	International and Specific Audiences	Audience Type*	Children
Contact Information	Other	Other Type*	Privacy contact information
Change Notification Process	Policy Change	Change Type	All
		Notification Type	All
		User Choice	All

D. Performance Evaluation Metrics

The model with the overall best metrics is selected to build the automated tool and is evaluated with precision, recall, F1-score, hamming loss, and subset accuracy. The precision measures how many positive predictions are correct, whereas the recall measures how many positive cases from the dataset are predicted correctly. The F1-score is the average of precision and recall. Hamming loss describes how many of the labels are classified incorrectly. The subset accuracy represents the predictions that are all classified as accurate.

IV. RESULTS

A multi-label classification system was used to label a paragraph in a privacy policy to determine whether the relevant privacy-related topics were present. The results of the machine learning models are shown in Table II with precision, recall, F1-score, hamming loss, and subset accuracy. We selected the algorithm with the highest F1-score and subset accuracy, Powerset SVC (Support Vector Machine). The performance metrics for each class are in Table III.

We built a custom NER model to create fine-grained annotations. The overall model has an F1-score of 45.48%. The entities range from an F1-score of 6% to 88%, as shown in Table IV.

V. DISCUSSION

We evaluated whether automation techniques can extract IoMT-relevant privacy factors from privacy policies. The results indicate that the quantity of the annotations varies by granularity and privacy factors. The coarse-grained multi-label classification and fine-grained custom NER model perform well and independently. However, NER techniques cannot capture some categories. Therefore, combining the two techniques for a hybrid approach seems to be the best idea for privacy policy analysis to help capture additional information and provide users with more detailed information. For example, the paragraph in Figure 1 returns different results depending on the model. The multi-label classification returns three labels related to Children, Data Collection, and Contact Information. The custom NER model returns the entities Children: Age, Personal Information Type, and Data Sharing recipient. The combined results provide a more comprehensive and accurate accounting of the paragraph topics.

The results might seem that NER techniques are not best suited for some privacy factors, such as data retention practices. But privacy policies cover several data practices, but most are related to data collection and sharing practices. There are more than 8,000 first party collection/use data practices compared to 370 data retention data practices in OPP-115 [46]. The imbalanced dataset likely leads to lower scores for data retention practices.

The results show inconsistency between the F1-scores of the classification and custom NER techniques for the Data Collection/Data Sharing and the fine-grained Personal Information Types/Data Sharing: Recipients annotations. With the imbalanced dataset, the results seem to contradict that NER techniques can provide additional details for privacy-related factors. However, data collection and sharing practices are often long lists of information within a sentence. The parser often captures most of the entities but due to their abundance some may be missed leading to a lower score. The automated parser helps enumerate several personal information types and recipients. The hybrid approach helps an individual identify the relevant paragraphs and allows them to note any missing or miscategorized information.

The entities well recognized (more than 50% F1-score) by an automated parser include Email, Children: Age, Date, Phone number, URL, Address, and User Choice and have well-defined values with similar compositions. The other entities generally contain vague language and differ considerably across privacy policies.

These results build on existing evidence that privacy poli-

TABLE II. MULTI-LABEL CLASSIFICATION RESULTS

Model	F1-score	Precision	Recall	Hamming Loss	Subset Accuracy
Bagging	71.01%	78.30%	64.97%	9.26%	47.42%
Boosting	70.02%	79.29%	62.69%	9.38%	46.06%
Multi-xnominal NB	67.24%	98.33%	51.08%	8.69%	43.94%
SVC SQ Hinge Loss (SVM)	75.00%	86.75%	66.05%	7.69%	53.79%
Binary Relevance	75.00%	86.75%	66.05%	7.69%	53.79%
SVM Transform	75.00%	86.75%	66.05%	7.69%	53.79%
Power Set SVC (SVM)	75.72%	82.67%	69.85%	7.82%	58.18%

TABLE III. POWER SET SVC CLASSIFICATION RESULTS

Category	Precision	Recall	F1-score	Support
data collection	81%	79%	80%	296
data sharing	86%	75%	80%	238
user choice	82%	58%	68%	151
data retention	44%	25%	32%	16
data security	81%	56%	67%	85
policy change	97%	77%	86%	44
children	100%	74%	85%	39
contact info	69%	58%	63%	53
micro avg	83%	70%	76%	922
macro avg	80%	63%	70%	922
weighted avg	83%	70%	75%	922
samples avg	85%	76%	78%	922

TABLE IV. CUSTOM NER MODEL RESULTS

Entity	F1-score	Precision	Recall
Performance	45.48%	50.60%	41.31%
Address	59.26%	80.00%	47.06%
Children: Age	85.71%	96.43%	77.14%
Data Retention: Period	6.67%	7.69%	5.88%
Data Sharing: Recipient	36.71%	49.34%	29.23%
Date	85.71%	88.24%	83.33%
Does/Does Not	31.35%	38.02%	26.67%
Email	88.24%	96.77%	81.08%
Personal Information Type	48.88%	48.05%	49.74%
Phone Number	80.00%	80.00%	80.00%
URL	67.39%	76.54%	60.19%
User Choice	58.36%	63.98%	53.65%

cies fail to capture users' actual concern. For example, data retention is a privacy factor that is especially relevant to users [33]. However, data retention practices have the worst performance metrics. The low-performance metrics make it difficult to provide users with relevant information for an organization's retention practices.

While previous research has focused on data collection, sharing, and retention practices, these results demonstrate that custom NER techniques can help provide fine-grained information for other relevant privacy factors, such as children-related statements.

While conducting our research, we encountered some limitations. OPP-115 does not contain privacy policies for IoMT devices. However, privacy policies have a general structure with similar language. Therefore, OPP-115 enabled us to create a tool that could automatically extract relevant information from the privacy policies of these specific devices. In the future, we intend to create a dataset of the privacy policies of IoMT devices.

OPP-115 contains relevant coarse-grained annotations serving as a starting place for the fine-grained annotations for the custom NER model. We refined the dataset using NLP techniques and custom scripts on the annotated portion of the paragraph. Consequently, some sentence contexts may be lost, causing the NLP techniques to miss or incorrectly label the fine-grained annotations. Furthermore, due to the lack of dates in the short date pattern (i.e., 8/15/22), our model cannot detect dates in this format. However, we intended to show the feasibility and usefulness of fine-grained annotations and not create a gold-standard dataset, and we are ok with some missed or mislabeled annotations.

Within the annotations, we sought to determine and list the personal information types that are collected and ignore any that are not collected. We leveraged the does/does not attribute to evaluate whether the privacy policy collected or shared information. However, this method was not reliable. The does/does not entity achieved a 31% F1-score. Future research needs to explore how to incorporate negation analysis into custom NER models. It is beyond the scope of this study to identify any contradictions within a privacy policy.

VI. RELATED WORK

Privacy policy analysis research is grouped into two categories, defining the criteria to assess privacy policies and examining the content or text of the privacy policy. Privacy policies are often legal documents whose requirements vary depending on the jurisdiction, industry, age, and domain. Therefore, an essential step in examining privacy policies is determining which privacy-related topics to investigate. Several articles considered legal frameworks and regulations as guidelines since organizations failing to comply could face fines and disciplinary action [17], [28], [30], [36], [40], [52]. In this study, we used the Privacy Policy Assessment Questionnaire [9] to capture IoMT-related concerns. While most research focuses on data collection, sharing, and retention because they align with users' concerns, there are several other categories related to IoMT to examine in IoMT privacy policies, such as data security.

Examining the content or text of the privacy policy is a growing research area. Content analysis approaches leverage manual, automatic, or semi-automatic techniques to extract factors from privacy policies. Fan et al. [22] leveraged binary classifiers to determine whether a sentence addressed a specific privacy factor. A sentence is processed by each classifier to capture multiple factors. Fan et al. classification was high-level, focusing on data collection and sharing. However, it is important to consider a more granular classification. Zimmeck et al. [55] classification includes the collection and sharing practices of specific data types, such as contact information, location, and device identifiers. Zimmeck et al. [54] expand further with even more granularity and modality. While those works focus on data collection and sharing techniques at different levels of granularity, we explore additional privacy-related topics, such as data retention and security practices, and combine both broad- and fine-grained analysis.

Natural language processing techniques can provide sentence and word-level context for automatic analysis. Liao et al. [26], Yu et al. [48], and Yu et al. [49] used NLP to extract and identify relevant phrases. Andow et al. [4] created data and entity dependency trees with NLP techniques to detect conflicting practices. Bui et al. [13] used BLSTM-CRF-based NER models to label data types. In this work, we combine machine learning and NLP techniques to build an automated two-layer model to detect privacy-related topics in privacy policies.

VII. Future Work

The future work of this study involves introducing more granularity to data collection and sharing related categories and entities. Specifically, the data collection-related entity, personal information type generally returns several data types in a paragraph. We intend to identify and group similar data types using topic modeling or multi-class classification to reveal more generic categories, such as health information, contact information, and device identifiers. We also intend to evaluate the approach with the privacy policies of IoMT devices.

VIII. Conclusion

As IoMT devices and applications and, by extension, privacy policies continue to rise, it is imperative to have tools to aid in privacy policy analysis. Our study explored whether machine learning and custom named-entity recognition techniques could extract IoMT-related privacy-relevant information from privacy policies. We found that combining the techniques provide relevant fine-grained information for several privacy topics, such as children and contact information. However, data collection and sharing practices revealed inconsistent results, with high coarse-grained F1-scores and low fine-grained F1-scores. This suggests that, while automation can reveal most personal information types and recipients, human intervention is still necessary for privacy policy analysis. In the future, we plan to investigate negation analysis to improve the accuracy and relevance of the results for potential use cases. Our tool can assist individuals, researchers, and developers, in identifying relevant privacy information from privacy policies for projects, such as automatically detecting mismatches between users' preferences and privacy policies.

Acknowledgment

This work was supported partially by the U. S. Department of Education under the Title III Historically Black Graduate Institutions (HBGI) grant. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U. S. Government. The U. S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] K. Abu Ali and S. Alyounis, "CyberSecurity in Healthcare Industry," in *2021 International Conference on Information Technology (ICIT)*. Amman, Jordan: IEEE, Jul. 2021, pp. 695–701. [Online]. Available: https://ieeexplore.ieee.org/document/9491669/

[2] A. Aljeraisy, M. Barati, O. Rana, and C. Perera, "Privacy laws and privacy by design schemes for the internet of things: A developer's perspective," *ACM Computing Surveys*, vol. 54, no. 5, May 2021, number of pages: 38 Place: New York, NY, USA Publisher: Association for Computing Machinery tex.articleno: 102 tex.issue_date: June 2021. [Online]. Available: https://doi-org.ncat.idm.oclc.org/10.1145/3450965

[3] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer, "Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset," in *Proceedings of the Web Conference 2021*. Ljubljana Slovenia: ACM, Apr. 2021, pp. 2165–2176. [Online]. Available: https://dl.acm.org/doi/10.1145/3442381.3450048

[4] B. Andow, S. Y. Mahmud, W. Wang, J. Whitaker, W. Enck, B. Reaves, K. Singh, and T. Xie, "PolicyLint: Investigating internal privacy policy contradictions on google play," in *28th USENIX security symposium (USENIX security 19)*. Santa Clara, CA: USENIX Association, Aug. 2019, pp. 585–602. [Online]. Available: https://www.usenix.org/conference/usenixsecurity19/presentation/andow

[5] B. Andow, S. Y. Mahmud, J. Whitaker, W. Enck, B. Reaves, K. Singh, and S. Egelman, "Actions speak louder than words: Entity-sensitive privacy policy and data flow analysis with PoliCheck," in *29th USENIX security symposium (USENIX security 20)*. USENIX Association, Aug. 2020, pp. 985–1002. [Online]. Available: https://www.usenix.org/conference/usenixsecurity20/presentation/andow

[6] A. I. Antón, E. Bertino, N. Li, and T. Yu, "A roadmap for comprehensive online privacy policy management," *Communications of the ACM*, vol. 50, no. 7, pp. 109–116, Jul. 2007, number of pages: 8 Publisher: Association for Computing Machinery tex.address: New York, NY, USA tex.issue_date: July 2007. [Online]. Available: https://doi-org.ncat.idm.oclc.org/10.1145/1272516.1272522

[7] S. S. Banerjee, T. Hemphill, and P. Longstreet, "Wearable devices and healthcare: Data sharing and privacy," *The Information Society*, vol. 34, no. 1, pp. 49–57, Jan. 2018. [Online]. Available: https://www.tandfonline.com/doi/full/10.1080/01972243.2017.1391912

[8] J. Bhatia, T. D. Breaux, J. R. Reidenberg, and T. B. Norton, "A theory of vagueness and privacy risk perception," in *2016 IEEE 24th international requirements engineering conference (RE)*, 2016, pp. 26–35.

[9] N. Bookert, W. Bondurant, and M. Anwar, "Data practices of internet of medical things: A look from privacy policy perspectives," *Smart Health*, vol. 26, p. 100342, Dec. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352648322000769

[10] J. Bowers, B. Reaves, I. N. Sherman, P. Traynor, and K. Butler, "Regulators, mount up! Analysis of privacy policies for mobile money services," in *Thirteenth symposium on usable privacy and security (SOUPS 2017)*. Santa Clara, CA: USENIX Association, Jul. 2017, pp. 97–114. [Online]. Available: https://www.usenix.org/conference/soups2017/technical-sessions/presentation/bowers

[11] T. D. Breaux and A. Rao, "Formal analysis of privacy requirements specifications for multi-tier applications," in *2013 21st IEEE international requirements engineering conference (RE)*, 2013, pp. 14–23.

[12] T. D. Breaux, H. Hibshi, and A. Rao, "Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements," *Requirements Engineering*, vol. 19, no. 3, pp. 281–307, Sep. 2014. [Online]. Available: http://link.springer.com/10.1007/s00766-013-0190-7

[13] D. Bui, K. G. Shin, J.-M. Choi, and J. Shin, "Automated extraction and presentation of data practices in privacy policies," *Proceedings on Privacy Enhancing Technologies*, vol. 2021, no. 2, pp. 88 – 110, Apr. 2021, publisher: Sciendo tex.address: Berlin. [Online]. Available: https://petsymposium.org/2021/files/papers/issue2/popets-2021-0019.pdf

[14] J. Burkell and A. Fortier, "Privacy policy disclosures of behavioural tracking on consumer health Websites: Privacy Policy Disclosures of Behavioural Tracking on Consumer Health Websites," *Proceedings of the American Society for Information Science and*

Technology, vol. 50, no. 1, pp. 1–9, 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/meet.14505001087

[15] P. Campisi and D. L. Rocca, "Brain waves for automatic biometric-based user recognition," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 5, pp. 782–800, 2014.

[16] F. T. Commission, "Google and YouTube Will Pay Record $170 Million for Alleged Violations of Children's Privacy Law," Sep. 2019. [Online]. Available: https://www.ftc.gov/news-events/press-releases/2019/09/google-youtube-will-pay-record-170-million-alleged-violations

[17] E. Costante, Y. Sun, M. Petković, and J. 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 - WPES '12*. Raleigh, North Carolina, USA: ACM Press, 2012, p. 91. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2381966.2381979

[18] L. F. Cranor, K. Idouchi, P. G. Leon, M. Sleeper, and B. Ur, "Are they actually any different? Comparing thousands of financial institutions ' privacy practices," *The Twelfth Workshop on the Economics of Information Security (WEIS 2013)*, Jun. 2013.

[19] L. F. Cranor, P. G. Leon, and B. Ur, "A large-scale evaluation of U.S. financial institutions' standardized privacy notices," *ACM Transactions on the Web*, vol. 10, no. 3, Aug. 2016, number of pages: 33 Publisher: Association for Computing Machinery tex.address: New York, NY, USA tex.articleno: 17 tex.issue_date: August 2016. [Online]. Available: https://doi-org.ncat.idm.oclc.org/10.1145/2911988

[20] T. Ermakova, B. Fabian, and E. Babina, "Readability of privacy policies of healthcare websites." *Wirtschaftsinformatik*, vol. 15, 2015.

[21] B. Fabian, T. Ermakova, and T. Lentz, "Large-scale readability analysis of privacy policies," in *Proceedings of the international conference on web intelligence*, ser. WI '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 18–25, number of pages: 8 Place: Leipzig, Germany. [Online]. Available: https://doi-org.ncat.idm.oclc.org/10.1145/3106426.3106427

[22] M. Fan, L. Yu, S. Chen, H. Zhou, X. Luo, S. Li, Y. Liu, J. Liu, and T. Liu, "An empirical evaluation of GDPR compliance violations in android mHealth apps," in *2020 IEEE 31st international symposium on software reliability engineering (ISSRE)*, 2020, pp. 253–264.

[23] N. Guntamukkala, R. Dara, and G. Grewal, "A machine-learning based approach for measuring the completeness of online privacy policies," in *2015 IEEE 14th international conference on machine learning and applications (ICMLA)*, 2015, pp. 289–294.

[24] C. Jensen and C. Potts, "Privacy policies as decision-making tools: An evaluation of online privacy notices," in *Proceedings of the SIGCHI conference on human factors in computing systems*, ser. CHI '04. New York, NY, USA: Association for Computing Machinery, 2004, pp. 471–478, number of pages: 8 Place: Vienna, Austria. [Online]. Available: https://doi-org.ncat.idm.oclc.org/10.1145/985692.985752

[25] P. G. Kelley, J. Bresee, L. F. Cranor, and R. W. Reeder, "A "Nutrition Label" for privacy," in *Proceedings of the 5th symposium on usable privacy and security*, ser. SOUPS '09. New York, NY, USA: Association for Computing Machinery, 2009, number of pages: 12 Place: Mountain View, California, USA tex.articleno: 4. [Online]. Available: https://doi.org/10.1145/1572532.1572538

[26] S. Liao, C. Wilson, L. Cheng, H. Hu, and H. Deng, "Measuring the effectiveness of privacy policies for voice assistant applications," in *Annual computer security applications conference*, ser. ACSAC '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 856–869, number of pages: 14 Place: Austin, USA. [Online]. Available: https://doi-org.ncat.idm.oclc.org/10.1145/3427228.3427250

[27] H. Lin, S. Garg, J. Hu, X. Wang, M. J. Piran, and M. S. Hossain, "Privacy-enhanced Data Fusion for COVID-19 Applications in Intelligent Internet of Medical Things," *IEEE Internet of Things Journal*, pp. 1–1, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9235575/

[28] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz, "The Privacy Policy Landscape After the GDPR," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 1, pp. 47–64, Jan. 2020. [Online]. Available: https://www.sciendo.com/article/10.2478/popets-2020-0004

[29] A. M. McDonald and L. Cranor, "The cost of reading privacy policies," 2009.

[30] N. Paul, W. B. Tesfay, D.-K. Kipker, M. Stelter, and S. Pape, "Assessing Privacy Policies of Internet of Things Services," in *ICT Systems Security and Privacy Protection*, L. J. Janczewski and M. Kutyłowski, Eds. Cham: Springer International Publishing, 2018, vol. 529, pp. 156–169, series Title: IFIP Advances in Information and Communication Technology. [Online]. Available: http://link.springer.com/10.1007/978-3-319-99828-2_12

[31] A. J. Perez, S. Zeadally, and J. Cochran, "A review and an empirical analysis of privacy policy and notices for consumer Internet of things," *Security and Privacy*, vol. 1, no. 3, May 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/spy2.15

[32] J. Polonetsky and E. Renieris, "Privacy 2020: 10 Privacy Risks and 10 Privacy Enhancing Technologies to Watch in the Next Decade," Future of Privacy Forum, Tech. Rep., Jan. 2020. [Online]. Available: https://fpf.org/wp-content/uploads/2020/01/FPF_Privacy2020_WhitePaper.pdf

[33] J. R. Reidenberg, N. C. Russell, A. Callen, S. Qasir, and T. Norton, "Privacy Harms and the Effectiveness of the Notice and Choice Framework," *SSRN Electronic Journal*, 2014. [Online]. Available: https://www.ssrn.com/abstract=2418247

[34] D. Reinhardt, J. Borchard, and J. Hurtienne, "Visual Interactive Privacy Policy: The Better Choice?" in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama Japan: ACM, May 2021, pp. 1–12. [Online]. Available: https://dl.acm.org/doi/10.1145/3411764.3445465

[35] A. H. Seh, M. Zarour, M. Alenezi, A. K. Sarkar, A. Agrawal, R. Kumar, and R. Ahmad Khan, "Healthcare Data Breaches: Insights and Implications," *Healthcare*, vol. 8, no. 2, p. 133, May 2020. [Online]. Available: https://www.mdpi.com/2227-9032/8/2/133

[36] X. Sheng and L. F. Cranor, "An evaluation of the effect of us financial privacy legislation through the analysis of privacy policies," *ISJLP*, vol. 2, p. 943, 2005, publisher: HeinOnline.

[37] L. Shipp and J. Blasco, "How private is your period?: A systematic analysis of menstrual app privacy policies," *Proceedings on Privacy Enhancing Technologies*, vol. 2020, no. 4, pp. 491–510, Oct. 2020. [Online]. Available: https://sciendo.com/article/10.2478/popets-2020-0083

[38] R. Slavin, X. Wang, M. B. Hosseini, J. Hester, R. Krishnan, J. Bhatia, T. D. Breaux, and J. Niu, "Toward a framework for detecting privacy policy violations in android application code," in *Proceedings of the 38th International Conference on Software Engineering*. Austin Texas: ACM, May 2016, pp. 25–36. [Online]. Available: https://dl.acm.org/doi/10.1145/2884781.2884855

[39] M. Srinath, S. Wilson, and C. L. Giles, "Privacy at scale: Introducing the PrivaSeer corpus of web privacy policies," in *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)*. Online: Association for Computational Linguistics, Aug. 2021, pp. 6829–6839. [Online]. Available: https://aclanthology.org/2021.acl-long.532

[40] A. Subahi and G. Theodorakopoulos, "Ensuring Compliance of IoT Devices with Their Privacy Policy Agreement," in *2018 IEEE 6th International Conference on Future Internet of Things and Cloud (FiCloud)*, Aug. 2018, pp. 100–107.

[41] A. Sunyaev, T. Dehling, P. L. Taylor, and K. D. Mandl, "Availability and quality of mobile health app privacy policies," *Journal of the American Medical Informatics Association*, vol. 22, no. e1, pp. e28–e33, Apr. 2015. [Online]. Available: https://academic.oup.com/jamia/article/22/e1/e28/700676

[42] M. Tabassum, A. Alqhatani, M. Aldossari, and H. Richter Lipford, "Increasing User Attention with a Comic-based Policy," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. Montreal QC Canada: ACM, Apr. 2018, pp. 1–6. [Online]. Available: https://dl.acm.org/doi/10.1145/3173574.3173774

[43] 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*. Tempe AZ USA: ACM, Mar. 2018, pp. 15–21. [Online]. Available: https://dl.acm.org/doi/10.1145/3180445.3180447

[44] X. Wang, L. Wang, Y. Li, and K. Gai, "Privacy-Aware Efficient

Fine-Grained Data Access Control in Internet of Medical Things Based Fog Computing," *IEEE Access*, vol. 6, pp. 47 657–47 665, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8412491/

[45] X. Wang, X. Qin, M. B. Hosseini, R. Slavin, T. D. Breaux, and J. Niu, "GUILeak: tracing privacy policy claims on user input data for Android applications," in *Proceedings of the 40th International Conference on Software Engineering*. Gothenburg Sweden: ACM, May 2018, pp. 37–47. [Online]. Available: https://dl.acm.org/doi/10.1145/3180155.3180196

[46] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. Giovanni Leon, M. Schaarup Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, T. B. Norton, E. Hovy, J. Reidenberg, and N. Sadeh, "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: Association for Computational Linguistics, 2016, pp. 1330–1340. [Online]. Available: http://aclweb.org/anthology/P16-1126

[47] S. Winkler and S. Zeadally, "Privacy Policy Analysis of Popular Web Platforms," *IEEE Technology and Society Magazine*, vol. 35, no. 2, pp. 75–85, Jun. 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7484849/

[48] L. Yu, X. Luo, X. Liu, and T. Zhang, "Can we trust the privacy policies of android apps?" in *2016 46th annual IEEE/IFIP international conference on dependable systems and networks (DSN)*, 2016, pp. 538–549.

[49] X. Yu, Y. Yang, W. Wang, and Y. Zhang, "Whether the sensitive information statement of the IoT privacy policy is consistent with the actual behavior," in *2021 51st annual IEEE/IFIP international conference on dependable systems and networks workshops (DSN-W)*, 2021, pp. 85–92.

[50] R. N. Zaeem, S. Anya, A. Issa, J. Nimergood, I. Rogers, V. Shah, A. Srivastava, and K. S. Barber, "PrivacyCheck v2: A Tool that Recaps Privacy Policies for You," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. Virtual Event Ireland: ACM, Oct. 2020, pp. 3441–3444. [Online]. Available: https://dl.acm.org/doi/10.1145/3340531.3417469

[51] R. N. Zaeem, R. L. German, and K. S. Barber, "PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining," *ACM Transactions on Internet Technology*, vol. 18, no. 4, pp. 1–18, Nov. 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3127519

[52] B. C. Zapata, A. Hernandez Ninirola, J. L. Fernandez-Aleman, and A. Toval, "Assessing the privacy policies in mobile personal health records," in *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. Chicago, IL: IEEE, Aug. 2014, pp. 4956–4959. [Online]. Available: http://ieeexplore.ieee.org/document/6944736/

[53] S. Zimmeck and S. M. Bellovin, "Privee: An architecture for automatically analyzing web privacy policies," in *23rd USENIX security symposium (USENIX security 14)*. San Diego, CA: USENIX Association, Aug. 2014, pp. 1–16. [Online]. Available: https://www.usenix.org/conference/usenixsecurity14/technical-sessions/presentation/zimmeck

[54] S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. Reidenberg, N. Cameron Russell, and N. Sadeh, "MAPS: Scaling Privacy Compliance Analysis to a Million Apps," *Proceedings on Privacy Enhancing Technologies*, vol. 2019, no. 3, pp. 66–86, Jul. 2019. [Online]. Available: https://content.sciendo.com/doi/10.2478/popets-2019-0037

[55] S. Zimmeck, Z. Wang, L. Zou, R. Iyengar, B. Liu, F. Schaub, S. Wilson, N. Sadeh, S. M. Bellovin, and J. Reidenberg, "Automated Analysis of Privacy Requirements for Mobile Apps," in *Proceedings 2017 Network and Distributed System Security Symposium*. San Diego, CA: Internet Society, 2017. [Online]. Available: https://www.ndss-symposium.org/ndss2017/ndss-2017-programme/automated-analysis-privacy-requirements-mobile-apps/