Privacy Policies Across the Ages: Content of Privacy Policies 1996–2021

ISABEL WAGNER*, University of Basel, Switzerland and De Montfort University, UK

It is well-known that most users do not read privacy policies, but almost always tick the box to agree with them. While the length and readability of privacy policies have been well studied, and many approaches for policy analysis based on natural language processing have been proposed, existing studies are limited in their depth and scope, often focusing on a small number of data practices at single point in time. In this paper, we fill this gap by analyzing the 25-year history of privacy policies using machine learning and natural language processing and presenting a comprehensive analysis of policy contents. Specifically, we collect a large-scale longitudinal corpus of privacy policies from 1996 to 2021 and analyze their content in terms of the data practices they describe, the rights they grant to users, and the rights they reserve for their organizations. We pay particular attention to changes in response to recent privacy regulations such as the GDPR and CCPA. We observe some positive changes, such as reductions in data collection post-GDPR, but also a range of concerning data practices, such as widespread implicit data collection for which users have no meaningful choices or access rights. Our work is an important step towards making privacy policies machine-readable on the user-side, which would help users match their privacy preferences against the policies offered by web services.

CCS Concepts: • **Social and professional topics** → **Privacy policies**; • **Security and privacy** → **Usability in security and privacy**; • **General and reference** → *Empirical studies.*

Additional Key Words and Phrases: privacy policy, longitudinal study, natural language processing, machine learning, neural networks, BERT

ACM Reference Format:
Isabel Wagner. 2023. Privacy Policies Across the Ages: Content of Privacy Policies 1996–2021. *ACM Trans. Priv. Sec.* 1, 1, Article 1 (January 2023), 34 pages. https://doi.org/10.1145/3590152

1 INTRODUCTION

A website's privacy policy is a legal document that explains what data the site collects from its users, how and for what purpose it processes the data, and with what other parties it shares the data. In addition, privacy policies can explain the users' rights regarding opting in or out of data collection, data correction, and data deletion. Privacy policies are notorious for being lengthy documents that are hard to understand [30, 31, 34]. They are rarely read by users, but website owners assert that by visiting their site users agree to their privacy policy. This gap in understanding between users and website owners deserves closer study: which data practices do users unwittingly agree to?

In the 25-year history of website privacy policies, the privacy and data protection rules have changed several times and in different jurisdictions, and privacy policies were updated to accommodate new requirements. For example, when the General Data Protection Regulation (GDPR)

Author's address: Isabel Wagner, isabel.wagner@unibas.ch, University of Basel, Spiegelgasse 1, Basel, Switzerland, 4051 and De Montfort University, UK.

2471-2566/2023/1-ART1 $15.00
https://doi.org/10.1145/3590152

came into force in 2018, privacy policies became longer and many websites introduced new privacy policies [16, 26].

However, despite much progress in natural language processing, existing analyses of privacy policies have been limited with respect to the scale and depth with which they analyze the specific data practices in privacy policies. In this paper, we fill this gap by analyzing *longitudinal changes in the content* of privacy policies over the last 25 years in terms of specific data practices and their attributes. Studying the content of privacy policies as well as longitudinal changes is important to understand whether new privacy regulations lead to increased user privacy or not; to understand whether privacy regulations lead to substantive changes in data practices or whether they merely result in corporate box-ticking exercises; to understand the state of privacy on the web and the extent to which websites respect user privacy; to guide the design of new privacy protections; and to guide the design of next-generation privacy regulations.

We leverage recent advances in machine learning and natural language processing [23] to automate the analysis of privacy policies at scale. In this paper, we answer the following research questions:

- How has the content of privacy policies evolved in terms of covered privacy practices and user rights?
- What was the effect of the GDPR and California Consumer Privacy Act (CCPA) on the privacy policy landscape?

To answer these research questions, we collect a corpus of more than 50,000 unique privacy policy texts spanning 25 years, from 1996 to 2021[1]. We train 22 machine learning classifiers using Bidirectional Encoder Representations from Transformers (BERT) to label data practices and their attributes, and present a detailed analysis of the policies in our corpus. To the best of our knowledge, we are the first large-scale study to classify the *attributes* of privacy policy statements and the first large-scale *longitudinal study of the contents* of privacy policies.

Overall, our results indicate that recent privacy regulations have not substantially improved the privacy of users online, but rather led to longer privacy policies that describe more categories of data practices, but in often vague and non-specific terms. In more detail, our findings include:

- *Completeness*: Privacy policies are becoming more complete in the sense that they address more categories of privacy practices, and as a result, are becoming longer. This confirms findings in prior work [4, 31] (Section 4.1).
- *Modality*: Assertions that personal data is not collected are much less common than assertions of collection and have been decreasing over time. For several personal information types, these decreases in non-collection mirror increases in collection, most notably for user online activities and cookies (Section 4.2).
- *Collection of personal information types*: We find a 5–10% reduction in the collection of some personal information types, including contact information, cookies, and user online activities, after the introduction of the GDPR and CCPA. Although this is a positive development, collection of user online activities remains at a high level: their implicit collection is asserted in 52% of policies in 2021 (Section 4.3).
- *Third-party sharing*: Another positive development is that sharing of identifiable personal information with third parties is decreasing – although it is important to keep in mind that tracking and profiling do not require identifiable personal information, and harm can be caused even when individuals can be only singled-out, but not identified (Section 4.4).

[1]Our policy corpus, including readability data and policy content labels, is available as an open access dataset [58].

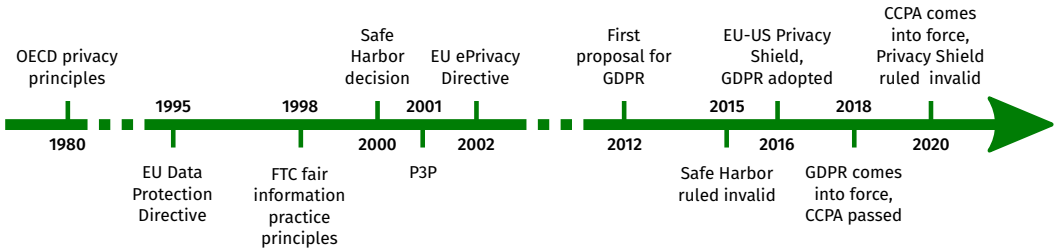


Fig. 1. Timeline of privacy-relevant events, regulations, and principles since 1980.

- *Notification of policy change*: Among policies that mention policy change in 2021, 78% notify users of changes either on the website or within the privacy policy itself, which means that most users are unlikely to become aware of changes to privacy policies (Section 4.5).
- *User choice*: Opt-in choices are offered mostly for explicitly collected data, such as contact information. Choices are often left unspecified for implicitly collected data, which encompasses user online activities, cookies, and device identifiers – commonly used for tracking and profiling of users (Section 4.6).
- *Vagueness*: Many attributes of data practices are left unspecified. For example, third parties are unnamed, identifiability of data is not stated, or collected personal information types are not enumerated. In addition, 72% of policy sentences in 2021 contained at least one obfuscating word. This confirms prior findings about the vagueness of policies [6] (Section 4.10).

The remainder of this paper is structured as follows. Section 2 discusses background information and related work. Section 3 explains our methodology for collecting and analyzing privacy policies, followed by our results on the contents of privacy policies in Section 4. Limitations of our approach are discussed in Section 5, and Section 6 concludes.

2 BACKGROUND AND RELATED WORK

In this section, we briefly describe background on privacy regulations as well as relevant work on privacy policy corpora, studies of the length and readability of privacy policies, and studies that work towards machine-readable privacy policies.

2.1 Privacy regulations

In the last two decades, the landscape of privacy regulations has seen significant changes (see Figure 1), including the introduction of Europe's GDPR in 2018 [18] and California's CCPA in 2020 [13]. The OECD privacy principles, introduced in 1980, encompass collection limitation, data quality, purpose specification, use limitation, security safeguards, openness, individual participation, and accountability [41]. These principles were intended to be implemented into national law by member countries. For example, European privacy legislation–most recently the GDPR–closely follows these principles. In contrast, privacy rules in the United States are more closely aligned to the Federal Trade Commission's fair information practice principles (FIPPs), which are based on notice/awareness and choice/consent. The FIPPs are seen as less comprehensive than the European privacy regime [48].

The International Safe Harbor Privacy Principles were developed between 1998 and 2000 as a way to reconcile the stricter European privacy regime with the FIPPs. The seven principles contain rules about notice, choice, onward transfer, security, data integrity, access, and enforcement. The Safe Harbor decision by the European Commission in 2000 ruled that the Safe Harbor principles

complied with the EU Data Protection Directive, and as a result companies who complied with the principles could register their (self-) certification and transfer data from EU to US.

However, the Safe Harbor decision was invalidated by the European Court of Justice (ECJ) in October 2015. In its place, the EU-US Privacy Shield was agreed in February 2016, introducing stronger obligations on US companies, stronger monitoring, and stronger enforcement. However, in 2020 the Privacy Shield was again declared invalid by the ECJ. As a result, EU-US data transfers are now governed for example by Article 49 of the GDPR which allows data transfers subject to explicit informed consent, or by contracts between EU data subjects and US data controllers. The CCPA, passed in 2018 and taking effect in 2020, was the first US state law that introduced comprehensive privacy rules in the European sense [54].

2.2 Length and readability of privacy policies

Privacy policies have been studied for more than a decade. For example, in 2008, the annual economic opportunity cost for reading privacy policies was estimated at $781 billion for US internet users [34]. In 2009, privacy policies were found to have Flesch Reading Ease (FRE) scores between 32 and 46 [35], which means that policies are *difficult* to understand, requiring high school or some college education[2]. In a 2017 study of 50,000 privacy policies, Fabian et al. [19] also find that policies are *difficult* to comprehend and require on average 13.6 years of education to understand, according to the Flesch-Kincaid grade level (FKG). In 2018, privacy policies from the top 1 million websites had an average FRE score of 39.8 [30]. In addition, higher-ranked policies were found to be longer (2,000+ words) than lower-ranked policies (1,400 words on average). An analysis of a longitudinal corpus of policies from 130,000 websites [4] shows a steady increase in policy length with the word count doubling between 2009 and 2019. The median FKG shows decreasing readability, indicated by an increase of roughly 0.5 years in the amount of required education between 2009 and 2019. These studies all focus on analyzing length and readability, but do not analyze the content of policies.

After the GDPR came into effect, more than 84% of European websites had a privacy policy and 62% displayed a cookie consent notice, an increase of 4.9% and 16%, respectively [16]. Policies became significantly longer, increasing from 2,145 words on average in 2016 to 3,603 words on average in May 2018 [16]. In addition, the use of GDPR-related terms, such as *complaint*, *data portability*, or *erasure*, increased by 6–12% between January and May 2018. Even though some studies find improvements in readability of privacy policies after introduction of the GDPR, Kretschmer et al. [26] conclude that privacy policies are still not understandable by the general public.

In contrast to these works, in this paper we dive deeper by analyzing the specific data practices described in privacy policies and their evolution over a 25-year period.

2.3 Machine-readable privacy policies

The Platform for Privacy Preferences (P3P) was introduced in 2001 as a W3C recommendation. P3P aimed at making privacy policies machine-readable, opening complex privacy policies to automated analysis and matching against user preferences. P3P relied on two components: the server-side component provides a policy following P3P's XML schema, and the user-side component retrieves and analyzes the policy, matches it against user preferences, and displays the result to the user. However, P3P was never mandated and as a result was never widely adopted [45]. In addition, many P3P policies were erroneous and seldom corrected or updated [45], and there were no mechanisms to ensure that the natural-language and P3P policy versions were equivalent, or that a website's actual practices conformed to the stated policy [33].

[2]Lower FRE scores mean that texts are harder to understand. For example, articles in the Harvard Law Review, which certainly require university education, have scores of about 30.

Standardized presentation of privacy policies has been proposed to make privacy policies easier to understand. For example, a tabular presentation can significantly increase comprehension and usability [25]. However, in 2022 natural-language policies are still the norm and neither standardized presentations nor machine-readable policies have been adopted.

2.4 Natural language processing for privacy policy analysis

With the application of natural language processing (NLP) and machine learning to label the contents of privacy policies, it might be possible to realize the intended benefits of P3P purely on the client-side. The approaches proposed for automated analysis of privacy policies in the past decade fall into two groups: symbolic NLP and statistical NLP [17]. Symbolic NLP approaches range from morphological/lexical analysis (e.g., matching of key terms [4]) to analysis of sentence-level syntax and semantics [62] and ontology reasoning [5, 6]. Statistical NLP approaches encompass unsupervised learning (e.g., topic modeling [49]), supervised learning [3], and artificial neural networks [12].

2.4.1 Statistical NLP. Early machine learning classifiers for privacy policies were based on support vector machines and hidden Markov models [59], presented together with the OPP-115 corpus of labeled privacy policies. Classifiers based on convolutional neural networks [23] and BERT models [36, 53] improved the classification performance. These classifiers were used to build a user interface that maps privacy icons to policy statements [23] and to analyze GDPR-related changes in the policy landscape using queries that assess the specificity and compliance of privacy policies [31].

Another strand of work predicts privacy categories together with risk levels for each category. For example, Tesfay et al. [55] use a Naive Bayes classifier to predict 11 privacy categories derived from the GDPR and three risk levels for each category. However, this classifier is not applied to privacy policies outside of the training set. In a similar approach, Zaeem and Barber [63] predict ten privacy categories and three risk levels (protected, at risk, compromised), and evaluate how risk levels have changed between 2016 and 2019 (pre-/post-GDPR). They find moderate improvements in risk levels throughout, but most significantly in the protection of children's privacy (improved risk level in 29% of policies). However, their policy corpus is quite small (550 policies), and the reported F1 scores (between 0.48 and 0.76) are low compared to our work. A follow-up study by the same authors increased the number of categories to 20, but did not conduct a large-scale analysis of privacy policies [37]. In addition, the categories analyzed in these studies only partially overlap with a unified list of 15 privacy categories recently proposed [7].

2.4.2 Symbolic NLP. Symbolic (NLP) has been used to identify contradictory statements in privacy policies of mobile apps [5], where 14% of policies contained misleading statements, including redefinitions of common understandings of terms and conflicts between terms used in different privacy regulations. NLP was also used to study the flow-to-policy consistency of mobile apps and their privacy policies, showing that the behavior of 40% of apps was not consistent with the app's privacy policy [6]. Topic modeling applied to a corpus of 1 million policies found 9 cohesive topics, with the most frequently addressed topic being "1st Party Information Type & Purpose" [53]. Compared with the hierarchy of labels in the OPP-115 corpus, these topics are less fine-grained and do not support a detailed analysis of policy contents.

2.4.3 Results from the application of NLP. In many cases, papers proposing new NLP approaches focus on evaluating the performance of the approach, but not on presenting insight from the application of the approach to a large corpus of privacy policies. The papers summarized in Table 1 are the exceptions: they all present results from the application of NLP to privacy policies on a

Table 1. Prior work that present results from large-scale application of NLP for privacy policy analysis. Symbolic NLP approaches (SY) include morphological/lexical (M), syntax/semantics (S), and ontology reasoning (O). Statistical NLP approaches (ST) include supervised learning (S) and artificial neural networks (N). Filled circles indicate in-depth analysis of the category or attribute; half-filled circles indicate limited or partial analysis.

				Categories									Attributes				
Reference	# of websites / apps	Longitudinal?	NLP approach	Entities	User access, edit & deletion	Data retention	Data security	Audience types	Do not track	Policy change	User choice & control	Privacy contact information	Personal information type	Modality	Purpose	Collection mode	Identifiability
Amos et al. [4]	108,499	1997–2019	SY-M	◐													
Andow et al. [5]	11,430	no	SY-O	●									●	●			
Andow et al. [6]	13,796	no	SY-O	●									●	●			
Fan et al. [20]	796	no	ST-S		◐		◐					●	●				
Kumar et al. [28]	6,885	no	ST-S							◐							
Linden et al. [31]	6,278	2016–2019	ST-N	◐	◐	◐	◐	◐		◐	◐	●			◐		
Slavin et al. [52]	477	no	SY-O										●				
Verderame et al. [57]	4,567	no	ST-S										●				
Xie et al. [61]	5,024	no	ST-S			●		●					●				
Yu et al. [62]	1,197	no	SY-S										●	●			
Zaeem and Barber [63]	550	2016, 2019	ST-S	◐	◐			◐		●			◐		◐		
Zimmeck et al. [65]	9,050	no	ST-S		◐					◐			◐				
Zimmeck et al. [64]	1,035,853	no	ST-S	●									●	●			
this paper	4,997	1997–2021	ST-N	●	●		●	●	●	●	●	●	●	●	●	●	●

large scale. Using the fine-grained categories and attributes from the OPP-115 corpus [59] as a guide, the table shows which aspects of privacy policies have been analyzed in prior work. It is clear that, while prior work has undoubtedly pushed the state of the art in NLP techniques, a comprehensive analysis of both categories and attributes is still missing, especially when taking a long-term longitudinal view. This detailed analysis of data practices at the attribute level is the main contribution of this paper.

2.5 Privacy policy corpora

To support the analysis of privacy policies, several research groups have collected and published corpora of privacy policies. The OPP-115 corpus consists of 115 website privacy policies collected in 2015 [59]. The corpus includes labels that indicate which high-level category each policy segment belongs to, and which detailed attribute-value pairs apply to each segment. This corpus has been widely used to train classifiers for policy segments [12, 23, 32, 36, 40, 53], however, prior work has largely focused on evaluating classifier performance as well as using the classifiers to support user-facing tools.

More recently, three large (unlabeled) corpora were published in 2021: a corpus of 1 million policies collected in 2019 [53], a corpus of 100,000 policies collected in 2020 [38], and a longitudinal corpus of policies from 130,000 websites, spanning the years from 1997 to 2019 [4].

Similar work exists for privacy policies of mobile apps: the APP-350 corpus includes labels for entities, privacy practices, and modality [64], and unlabeled corpora with more than 400,000 policies each exist for Android [64] and iOS [2].

In this paper, we focus on web privacy policies and use the OPP-115 corpus to train machine learning classifiers. For our large-scale analysis, we did not use the two single-snapshot corpora because we our analysis required a longitudinal corpus. We labeled and analyzed the policies from the Princeton-Leuven corpus [4], however, we use our own corpus to present the results because our corpus includes policies for 2020 and 2021 which are important to study because new privacy regulation (CCPA) came into force in early 2020. We did not find substantial differences in the results for 1997–2019 between the Princeton-Leuven corpus and ours (see Appendix A for details), which indicates that our corpus contains a sufficiently large sample of privacy policies to ensure generalizability of the insights.

3 METHODOLOGY

Our methodology for collecting and analyzing a longitudinal corpus of privacy policies consists of 4 steps: (1) crawling websites to find links to their privacy policies, (2) retrieving the policy texts, relying on the Wayback Machine to retrieve historical policy texts going back to 1996, (3) training and evaluating 22 machine learning classifiers for data practices and their attributes, and (4) evaluating the data practices described in policies using the trained machine learning classifiers.

We do not use the corpus published by Amos et al. [4], because their policy collection ended in 2019 and thus does not reflect the policy updates made after the CCPA came into effect. However, we find that our analysis results are very similar for both corpora (see Appendix A).

For all crawls, we use computers on our university campus between January 2020 and December 2021. In cases where access to privacy policies was filtered by our university firewall (e.g., for pornographic websites)[3], we used supplementary crawls from a residential internet connection. We use the Wayback Machine to retrieve policies between 1996 and February 2020, and monthly live crawls between March 2020 and December 2021.

3.1 Selecting websites and dates to crawl

To select websites for our longitudinal analysis of privacy policies, we combine two approaches: sampling from a recent version of the Tranco list [42], and sampling from historical versions of the Alexa toplist, for a total of 4,997 sites. We focus primarily on higher-ranked websites because rankings of lower-ranked websites are not available prior to 2010. Our sample of sites is large compared to prior work that has analyzed the content of privacy policies, e.g., topics covered in policies (550 policies [63]) or data deletion and opt-out choices (150 policies [22]).

Specifically, we use a stratified sample from the Tranco list (1 October 2019[4]), consisting of the top 1,000 websites plus 1,000 sites drawn uniformly at random from the top 1,000 to 10,000. In addition, we select the top 1,000 sites from the Tranco list from 31 March 2021[5]. Second, we add sites from the historical Alexa toplist for each year [29]: for 2010–2021, we use the top 1,000 sites of the Alexa top one million; between 2003 and 2009, we scrape the top 500 sites from the Alexa website as archived by the Wayback Machine; and for 2002, we use the top 100 sites from the archived Alexa website.

For each site, we retrieve the list of available snapshots for the landing page using the Wayback Machine's CDX API. To pick up when a site's privacy policy moves to a new URL, we select one

[3]This filtering is easy to detect because the firewall serves a block page which mentions the university name.
[4]https://tranco-list.eu/list/JL9Y
[5]https://tranco-list.eu/list/ZLZG

snapshot per year between 1996 and 2008, quarterly snapshots between 2009 and 2017, and monthly snapshots after that. In addition, we retrieve the category for each site from Alexa.

3.2 Finding privacy policy links

We use Firefox, automated with Selenium, to load the landing pages and parse the HTML with BeautifulSoup 4 [47]. To locate links to privacy policies, we search through link titles and link URLs in reverse order. Because there is no standard naming scheme for privacy policy links, we search for each of the terms *privacy polic*, *privacy*, *terms of service*, *web policies*, *cookie polic*, *data polic* and *legal*. Across all snapshots, we find 27,329 privacy policy links.

3.3 Retrieving the policy text

To retrieve the privacy policy text, we identify available snapshots for each policy link using the Wayback Machine's CDX API and fetch one snapshot per month, as far back as available. For each snapshot, we load the link with Firefox/Selenium, scroll to the bottom, and save the resulting HTML. We follow HTTP redirects within the Wayback Machine. If a page has not loaded completely after two minutes, we trigger a timeout and extract the policy text from the partially loaded page. This often succeeds when the page was waiting for embedded resources. To extract the policy text from HTML pages, we use both Firefox's reader mode and the readability-lxml library to strip navigational elements, page headers, and page footers. While both reader mode and readability-lxml consistently remove non-policy elements, they sometimes also remove parts of policy text. To mitigate this, we compare the length of both extracted texts and keep the longer of the two.

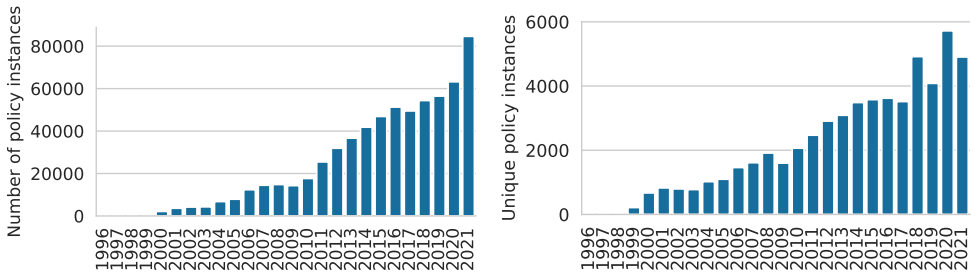
Some sites only display a short summary of the policy instead of the full policy when clicking on the landing page's privacy policy link. To catch these cases, we search for links within policies whose titles contain *privacy* or *policy* as well as *full*, *entire*, or *complete*, plus titles that contain *privacy statement*, *privacy polic*, *privacy notice*, or *privacy*, and add these links to our list.

In total, we fetched 1,068,683 documents as potential privacy policies, with 120,265 unique documents (an average of 39.1 policy instances and 4.4 unique policy texts per link).

3.4 Data cleaning

Because the training data for our machine learning classifiers is in English, we remove all non-English policies from our database. In particular, we use the PYCLD2 package [1] for language detection and remove all policies where English was not the language detected with highest confidence. In addition, we remove short policies with fewer than 100 words because they usually contain brief summaries or error messages, not policy text.

In addition, we implemented the classifier from Linden et al. [31] to identify which of the documents in our corpus are privacy policies. We trained the classifier with the same corpus of 1,000 privacy policies as Linden et al. [31], but used our own set of non-policy documents because theirs was not available. We trained three versions of the classifier with different sets of non-policy documents: (1) a selection of 1,000 landing pages from our crawls, (2) landing pages longer than 5,000 characters, and (3) a selection of 1,000 subsites crawled from the landing pages of the Tranco top 500 sites. To ensure that the non-policy corpus does indeed not contain privacy policies, we filter the non-policy corpus so that it does not include keywords expected in privacy policies using the same list of keywords as above. Evaluated on the test set (10% of samples), the three classifiers have F1 scores of 0.97, 1.0, and 0.98, respectively. We remove policy texts that have a low probability to be a privacy policy according to all three classifiers, using empirically determined thresholds of 0.9 for classifier (1), 0.6 for classifier (2), and 0.1 for classifier (3). The combination of the three classifiers labels all samples in the test set correctly.



(a) Non-unique (b) Unique

Fig. 2. Number of policy instances collected for each year. Unique policy texts are counted in the year they first appeared.

After filtering, our corpus contains 56,416 unique privacy policy texts. Although this is less than the 1m policies in recent work [4], we note that our analysis of detailed data practices is computationally expensive, whereas prior work has only focused on simple metrics like policy length and readability. Figure 2 shows how many policy snapshots and unique policies were collected for each year. The peaks for unique policy texts in 2018 and 2020 (Figure 2b) show that the introductions of the GDPR and CCPA, respectively, caused many organizations to update their privacy policies.

3.5 Classifying content of privacy policies

To evaluate which data practices are described in privacy policies, we follow Harkous et al. [23]. Specifically, we implement a hierarchy of classifiers so that the top-level classifier labels the topic, or category, of each segment of a privacy policy, and the lower-level classifiers label the attributes described in each segment.

For example, consider the segment: "As you navigate through and interact with our Website, we may use automatic data collection technologies to collect certain usage information about your equipment, browsing actions, and patterns, including: details of your visits to our Website, including traffic data, location data, logs, and other communication data and the resources that you access and use on the Website." This segment is labeled as *First Party Collection/Use*, and its 9 attribute-value pairs are Does/Does Not=*Does*, Collection Mode=*Implicit*, Action First-Party=*Collect on website*, Identifiability=*Unspecified*, Personal Information Type=*Location* and *User online activities*, Purpose=*Unspecified*, User Type=*Unspecified*, Choice Type=*Unspecified*, Choice Scope=*Unspecified*.

To train these classifiers, we rely on the OPP-115 corpus [59], which is a labeled collection of 115 privacy policy texts. Each privacy policy segment was labeled with one or more of ten top-level categories, and then further labeled with attribute-value pairs that represent its data practices in detail. Each policy in this corpus was annotated independently by three legal experts. The inter-rater consistency as measured by Fleiss' Kappa ranged between 0.91 and 0.49, depending on the top-level category.

3.5.1 Preprocessing. To prepare the OPP-115 corpus for training, we ensure consistent spelling of all attribute labels, in particular consistent use of upper-/lower-case (e.g., "User Profile" vs. "User profile"). We use the full set of annotations (i.e., the *annotations* folder), but apply majority vote consolidation [36], i.e., we only include labels if at least two annotators agree on the label. This is applied for top-level category labels as well as for attribute labels.

Table 2. Hyperparameter settings for training top- and attribute-level classifiers

attention_probs_dropout_prob	0.1
gradient_checkpointing	false
hidden_act	gelu
hidden_dropout_prob	0.1
hidden_size	768
initializer_range	0.02
intermediate_size	3072
layer_norm_eps	1e-12
max_position_embeddings	512
num_attention_heads	12
num_hidden_layers	12
pad_token_id	0
type_vocab_size	2
vocab_size	30522

In addition, we restrict attribute labels to those reported in Harkous et al. [23]. The omitted labels have very small support and would therefore be difficult to train correctly. Accordingly, restricting attribute labels improves the performance of our classifiers. The labels for all classifiers are one-hot encoded using a multi-label binarizer, that is, all policy segments can be assigned more than one label. For example, it is possible that a policy segment covers more than one top-level category, and that it describes several lower-level attributes.

3.5.2 Classifier training. We train one top-level classifier and 21 attribute-level classifiers using the fast-bert library [56], which is based on HuggingFace transformers [60]. In a pretraining step, we first fine-tune the *bert-base-uncased* language model using all unique policy texts in our corpus (4 epochs, batch size=8). Fine-tuning is a computationally expensive step (33 hours per epoch on our hardware), but improves classifier performance.

For the top-level classifier, we use the train-test-validation split reported in Mousavi Nejad et al. [36], which is a random split with a 3:1:1 ratio using the majority-vote version of the OPP-115 dataset. We train the classifier for 100 epochs with a batch size of 8 using the *train* portion of the dataset. Hyperparameter settings are shown in Table 2. We use the *validation* portion of the dataset to evaluate the loss after each epoch.

Table 3 shows the performance of our top-level classifier based on the *test* portion of the dataset in terms of F1 score, compared with prior work (detailed performance results are in Appendix B). The performance on average is in line with the state of the art [36, 53]. Differences are most likely due to the fine-tuning step, where we used a different corpus than existing works, and possibly due to differences in the batch size and number of epochs. Our results are slightly worse than the results from [23] which may be due to differences in their train/test split and data augmentation process. Overall, the classifier performance is at about the same level as the inter-rater consistency reported for the OPP-115 dataset [59].

For the attribute-level classifiers, the train-test-validation split from [36] results in imbalanced splits where some attribute labels are missing from some splits. Therefore, we create a separate stratified 3:1:1 split for each attribute. Because we have multi-label data, we apply an algorithm for multi-label stratification instead of the default stratifiers in scikit-learn [10, 50]. We tune the number of training epochs by comparing training loss and validation loss and selecting the final epoch as the one just before the two losses start to diverge.

Table 4 shows the macro F1 scores of all 21 attribute-level classifiers, compared with prior work. We note that BERT has not been applied to attribute-level classifiers before. On average, our BERT

Table 3. F1 score for our top-level classifier vs. prior work. The best F1 scores for each category are in **bold**.

	CNN, maj. [36]	BERT, maj. [36]	CNN, union [23]	PrivBERT [53]	BERT, maj. (here)
First Party Collection/Use	82	**91**	79	92	90
Third Party Sharing/Collection	82	90	79	**91**	87
User Access, Edit & Deletion	70	73	80	84	**85**
Data Retention	40	56	71	**77**	56
Data Security	75	80	85	**86**	85
International/Specific Audiences	82	83	**95**	86	84
Do Not Track	100	**100**	95	**100**	**100**
Policy Change	88	90	88	**91**	89
User Choice & Control	72	81	74	**83**	82
Introductory/Generic	73	79	70	77	**81**
Practice Not Covered	13	35	**70**	52	47
Privacy Contact Information	84	78	**87**	81	78
Micro average	78	85	–	**87**	85
Macro average	71	79	81	**83**	80

Table 4. Macro F1 scores for attribute-level classifiers vs. prior work. Some results, shown as n/a, were not reported in Harkous et al. [23].

	CNN [23]	BERT (here)
Access Scope	n/a	**67**
Access Type	62	**90**
Action First-Party	65	**87**
Action Third Party	n/a	**74**
Audience Type	**97**	**97**
Change Type	76	**90**
Choice Scope	59	**63**
Choice Type	73	**78**
Collection Mode	n/a	**85**
Do Not Track Policy	**100**	**100**
Does/Does Not	86	**93**
Identifiability	77	**91**
Notification Type	71	**94**
Personal Information Type	81	**83**
Purpose	83	**84**
Retention Period	73	**89**
Retention Purpose	n/a	**84**
Security Measure	74	**82**
Third Party Entity	73	**80**
User Choice	n/a	**81**
User Type	n/a	**92**

classifiers clearly outperform the CNN-based prior work by about 10%. Detailed performance results for all attribute classifiers are in Appendix C.

3.5.3 Policy segmentation. To apply these classifiers to policy texts from our corpus, we have to split each policy into semantically coherent segments. We find that the list aggregation technique

proposed by Harkous et al. [23], which relies on HTML tags, does not work consistently on our policy corpus, in part due to our reliance on reader mode and readability-lxml. Instead, we use GraphSeg [21] (relatedness threshold 0.25, minimal segment size 1). Instead of the default word embeddings in GraphSeg, we use custom word embeddings that are specific to the privacy policy domain. Specifically, we use an unsupervised fastText model (model type: skipgram, dimensions: 300, minimum word count: 5) based on our corpus of unique policy texts.

3.5.4 Policy content labeling. Before labeling the policy segments, we discard non-English segments. This step improves labeling results for policies that include several languages in one document. Then, we apply the top-level classifier to label the top-level category for each segment. Finally, for each attribute that is relevant for the labeled top-level category, we apply the corresponding attribute classifier. In addition to the predicted labels, we record the numeric prediction confidence (i.e., the output of the final sigmoid function).

3.5.5 Comparison to manual annotations. In addition to evaluating the performance of our classifiers, we evaluate the agreement between the three subject-matter experts who annotated the OPP-115 corpus and our classifiers. To evaluate agreement, we use a set of policies from the OPP-115 corpus that was not used during classifier training and compute Krippendorff's alpha (preferable to Fleiss' kappa because it can handle multiple labels for each document, documents being rated by a variable number of raters, and different raters rating each document [[27]]). We find that the average agreement between three subject-matter experts is 0.75 for top-level categories and 0.53 for attribute values. Agreement between the majority vote among human annotators and the classifiers is 0.83 for top-level categories and 0.68 for attribute value, indicating good and acceptable agreement, respectively [27]).

4 CONTENT OF PRIVACY POLICIES

In this section, we report the results of our analysis, for top-level categories of data practices (Section 4.1) and for specific attributes of data practices (Sections 4.2–4.9).

To analyze the semantic content of privacy policies, we are interested in how many policies each year address each privacy practice, and in what way, e.g., whether they assert collection or sharing of a specific information type. To this end, we label each privacy policy segment with the top-level category classifier and with each attribute classifier that is relevant to its category labels. We then eliminate segments with duplicate labels, i.e., we remove segments that belong to the same unique policy if they have the same labels, regardless of the segment text or the policy's timestamp. To analyze top-level categories, we retain only the first mention of each category per policy, at its earliest instance.

In addition, we exclude categories and attributes for which the classifiers did not perform well. Specifically, we retain only labels that have a *precision* of at least 75%. This ensures that the predicted labels are most likely correct (low false positives), while accepting the possibility that the classifiers miss some labels (low recall corresponding to high false negatives). As a result, the reported results may underestimate the true prevalence of privacy practices in privacy policies. Specifically, for top-level categories, we exclude the *data retention* and *practice not covered* categories. For attribute classifiers, we exclude a total of nine labels (out of 98 labels total) across the 21 classifiers (marked with an asterisk in Appendix C).

In the subsequent figures, we present the fraction of policies labeled with specific attributes or combinations of attributes as bar plots, where each bar represents policies from one year. The height of each bar indicates the fraction of policies for which the classifier's confidence was above 0.5 (where combinations of attributes are plotted, both confidences are above 0.5). We compute 95% prediction intervals for all results, shown as grey error bars. Our calculation of prediction intervals

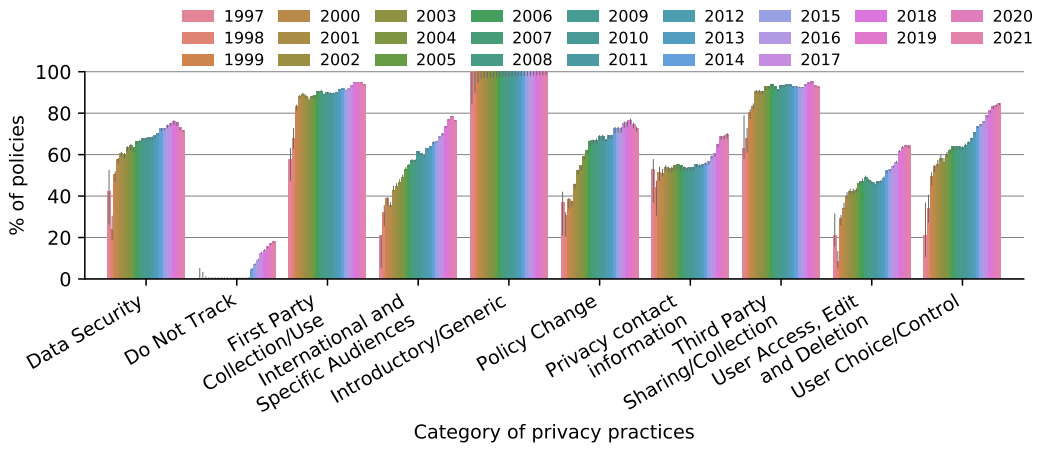


Fig. 3. Fraction of privacy policies each year that address each category of data practices.

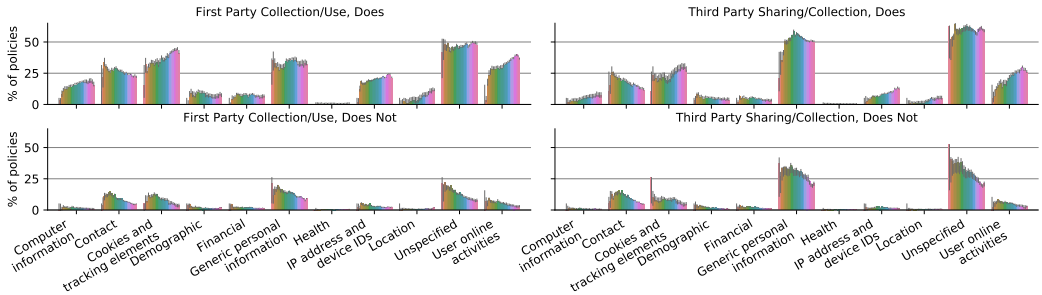


Fig. 4. Percentage of policies that mention specific personal information types, split by whether the collection is for first party use vs. third party sharing, and by whether the policy *does* or *does not* assert the data collection. The figure omits personal information types with percentages close to zero (personal identifier, social media data, survey data).

relies on the insight that the number of policies with a given label is a random variable with a Poisson Binomial probability distribution [24]. The number of policies corresponds to the mean of this distribution, and our prediction intervals are computed based on the 0.975 and 0.025 quantiles. Importantly, this approach takes into account all estimated probabilities, not just the ones above the classification threshold of 0.5 [15].

4.1 Top-level categories of data practices

Figure 3 shows the fraction of policies that address each category of data practices over the past 25 years. Almost all policies contain *introductory/generic* statements which we disregard in the remainder of the analysis. Overall, the trend is that privacy policies address more data practices each year, i.e., they have become more comprehensive and thereby longer. This is supported by studies that have analyzed readability and length of privacy policies [4, 30].

The most commonly addressed categories are *first-party collection and use* (94% of privacy policies in 2021), followed by *third-party sharing and collection* (93% in 2021).

Statements about *user access, edit and deletion* rights and *user choice/control* show a large increase of 30–40% in the early 2000s. These categories correspond closely to two Safe Harbor principles, *choice* and *access*. The increase is likely caused by websites aiming to facilitate data transfers from the EU to the US by complying with the Safe Harbor principles after the European Commission's Safe Harbor decision in 2000.

Two categories show a notable ~10% increase after 2018: *privacy contact information* and *user access, edit, and deletion*. This increase is most likely caused by the GDPR which requires that users are informed about these topics. However, from a regulatory viewpoint many privacy policies are still lacking: in 2021, only 67% of policies explain users' access, edit, and deletion rights, and only 71% give contact details for privacy-related queries.

Starting in 2020, however, there is a slight decline for some data practices, including statements about *data security* and *policy change*. One possible reason for this is that new regulations (GDPR and CCPA) introduced specific wording for these categories which privacy policies subsequently adopted, but which was not common at the time the training data was collected. However, we believe the more likely reason is that, even though both GDPR and CCPA require data security measures, there does not seem to be a requirement to inform users about the specific security measures taken. As a result, some websites may simply have removed corresponding statements from their policies. For example, the privacy policy of *sagepub.com* contained a paragraph about security measures until April 2018 ("SAGE uses industry-standard encryption technologies when transferring and receiving consumer data exchanged with our Web Site."), but not thereafter. Similarly, organizations may have reasoned that the regulations do not require them to state explicitly how users will be informed of policy changes.

Finally, we note that the do not track header is mentioned in 20% of policies, which all assert that they do not respect the header.

4.2 Personal Information Types

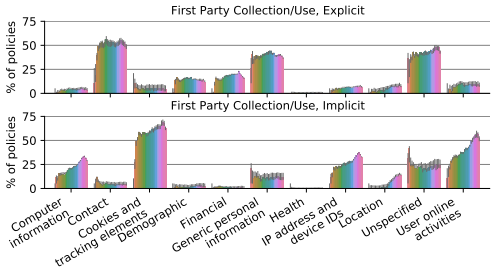
Figure 4 shows which personal information types are mentioned in privacy policies, split by first- or third-party data collection and by whether the policy *does* or *does not* assert the data collection. We observe that in most cases, policies assert that data is indeed collected, especially for first-party collection (although collection rates are not much lower for third-party sharing). In addition, the percentage of policies that do not collect data shows a decreasing trend for most personal information types. For third-party sharing, assertions of non-collection are more frequent, most notably for sharing of *generic personal information*, where 22% of 2021 policies state that data is not collected. However, many more policies (50%) assert that they do share *generic personal information* with third parties. We also note the increasing rate of location data collection, especially for first-party use. This is concerning due to the sensitive nature of location data.

In the following sections, we filter the results to only include policy statements that *assert* data collection.

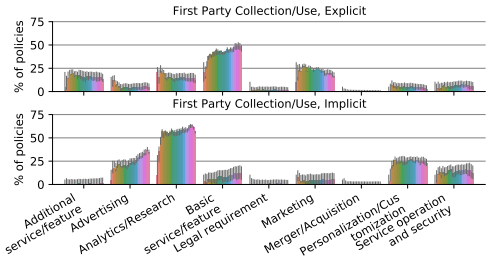
4.3 First-party data collection/use

The most salient attributes for first-party data collection/use are the collection mode, purpose, personal information types, choice and controls offered to users, and identifiability of collected data.

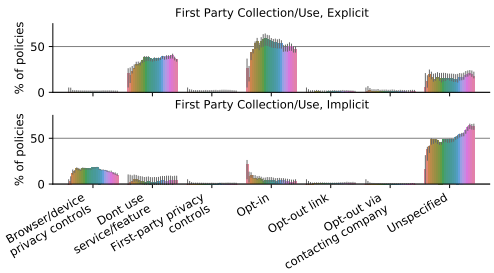
4.3.1 Collection mode vs. personal information type, purpose, choice/control, and identifiability. The *collection mode* describes whether information is explicitly provided by the user (e.g., data entered in a form), or collected implicitly, e.g., in the background, possibly without the user's knowledge. Figure 5a shows personal information types by collection mode. *Generic personal information* and



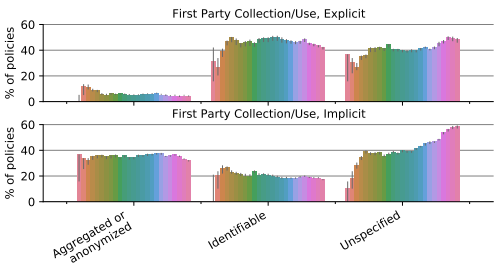
(a) Personal information type



(b) Purpose



(c) Type of choice and controls offered to users



(d) Identifiability of the collected data.

Fig. 5. Collection mode for first-party data collection.

contact data is most often collected explicitly, whereas *cookies and tracking* data, data about *user online activities*, and device identifiers (*computer information*, *IP address and device IDs*) are mostly collected implicitly.

The trend, particularly for implicit data collection, is clearly towards more data collection. However, starting between 2018 and 2020, we can see a 5–10% decline for many personal information types, for example explicitly collected contact information and implicitly collected cookies, device identifiers, and user online activities. This points to a positive effect of new data protection regulations. The reduction in online tracking in particular has also been confirmed by measurement studies [14]. Nevertheless, we note that online tracking and profiling (e.g., via analysis of user online activities) is still very common: in 2021, 52% policies assert that they implicitly record user online activities.

Location data is collected implicitly at almost double the rate than explicitly (14% vs. 7.5% in 2021), which is concerning due to the steep rise in location data collection and the sensitivity of location data. The rise in the implicit collection of location data could be related to the fact that most browsers implemented the W3C Geolocation API around 2010/2011, and to the fact that increasingly "interesting" location data is available with the increasing use of mobile devices as opposed to stationary PCs.

Figure 5b shows the purposes stated for data collection by collection mode. Explicitly collected data is most often used for *basic service features* (e.g., logins), whereas implicitly collected data is used most for *analytics* and *advertising*. As before, the trend is towards more data collection, with two exceptions. First, explicit data collection for marketing shows a decreasing trend. This may be because revenue generation through marketing of the website's own products has been supplanted by revenue generation through advertising, which has become more and more important as a

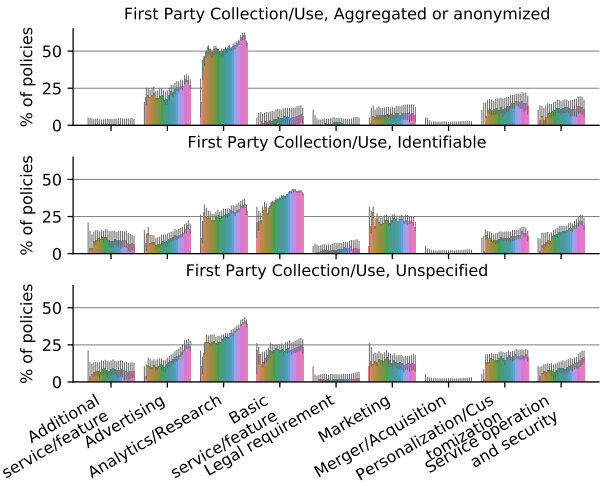


Fig. 6. Purpose of data collection, by identifiability of the collected data.

business model on the Internet. Second, similar to the personal information types above, some purposes show a decline after 2018–2020, e.g., implicit data collection for analytics, pointing to positive effects of data protection regulations.

Figure 5c shows the types of choice and control mechanisms offered to users for each collection mode. For explicitly collected data, the rate at which opt-ins are offered has decreased over the last decade. In 2021, opt-ins are offered at almost the same rate as the "choice" to stop using the service or feature. For implicitly collected data, most policies leave user choices unspecified, with a rate that was increasing until 2018. For example, the privacy policy of *waldenu.edu* referred users to their browser's privacy settings to control Google Analytics cookies up until January 2018 ("You may refuse the use of cookies by selecting the appropriate settings on your browser"), after which the mention of this choice disappeared. Opt-ins are offered less frequently than asking users to rely on their browser's privacy controls or to stop using the service.

If we link the different choices users get for implicit vs explicit data collection with the different personal information types and purposes for each collection mode, we can see that users rarely get to opt in before their online activities are recorded, made linkable over time via cookies and device identifiers, and used for advertising and analytics purposes. This is concerning, not least because the frequency with which opt-ins are offered has not increased after the introduction of GDPR and CCPA.

Figure 5d shows that implicitly collected data is more often aggregated or anonymized than explicitly collected data. This is positive because it shows an effort to protect data that may have been collected without the user's knowledge. However, because anonymizing data is notoriously difficult, most websites are likely to use simple aggregation. In addition, many policy segments leave the identifiability of collected data unspecified, with a high rate especially for implicit collection. This indicates that policies are often vague.

4.3.2 Identifiability vs. purpose. Figure 6 shows the purpose of data collection by identifiability of the data. Aggregated or anonymized data is most commonly used for *analytics*, followed by *advertising*. Identifiable data is used to provide *basic service features*, but also for analytics, advertising, marketing, and service operation. It is concerning that the number of data collection

purposes where identifiability is left unspecified is similar to the other two groups. Over the years, an increasing number of policies asserts data collection for *analytics*, *advertising*, *service provision*, and *service operation*. The first two purposes reflect web business models driven by advertising revenue, but the last two may indicate an increasing use of *legitimate interest* as a lawful basis for data processing instead of user consent, as for example allowed by the GDPR [26].

4.4 Third-party data sharing/collection

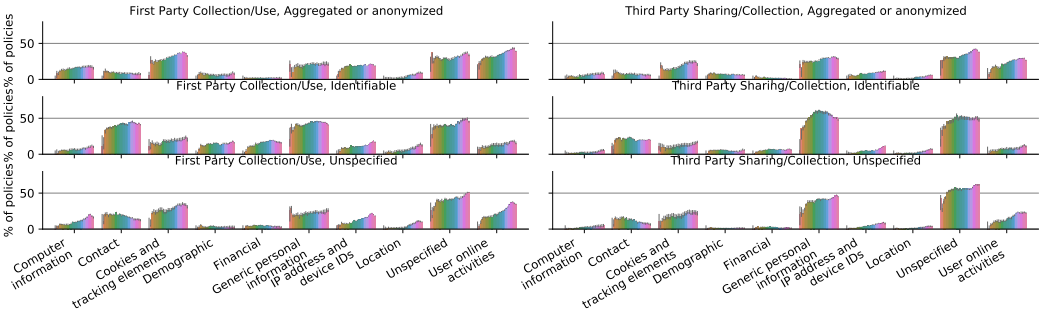


Fig. 7. Personal information types collected by first- vs. third-parties, by identifiability of the collected data.

4.4.1 Identifiability vs. personal information type. Figure 7 compares the personal information types collected by first- and third-parties, split by identifiability of the collected data. This Figure expands the top row of Figure 4, separating data collection by whether the collected data identifies an individual, whether it is aggregated or anonymized, or whether the policy leaves identifiability unspecified.

In most cases, policies collect more different personal information types for first-party use than for third-party sharing, with the exception of *generic personal information*, which is shared with third parties more often than it is used by first parties, regardless of identifiability. Even though the percentage of policies sharing identifiable PI with third parties has decreased since 2009 (from 61% to 48%), this is partially compensated by an increase in policies that do not specify identifiability (from 41% to 45%). We can draw two conclusions from this data: First, sharing of identifiable personal information with third parties is decreasing. This is a positive development over the last decade. However, it is important to keep in mind that the creation of user profiles does not require identifiable data: it is possible to single-out users without being able to identify them [9], and this can also cause harm, e.g., through discriminatory targeting. Second, policies are becoming more vague by leaving more attributes of their data practices unspecified. Figures 5 and 6 already showed several examples for this trend. This is a concerning development because it indicates that, while policies are becoming longer and more comprehensive in terms of the categories of data practices they address, they actually contain less specific detail about the attributes of their data practices. We analyze this finding further by studying the use of obfuscating words in privacy policies in Section 4.10.

4.4.2 Third-party entity. Figure 8 shows which types of third-party entities are mentioned in privacy policies. Over the last decade, we can see an increase in *named* third parties, which is a positive development. However, named third parties include categorized third parties that are not identified by name, such as *advertisers*. Even though this is likely GDPR compliant (articles 13 and 15, for example, only require categories of recipients to be specified), more specific information may be more desirable from a user's point of view. In addition, the majority of third parties is unnamed.

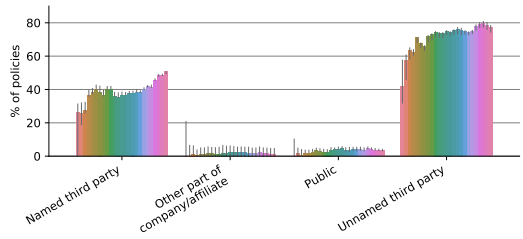


Fig. 8. Third-party entities described in privacy policies.

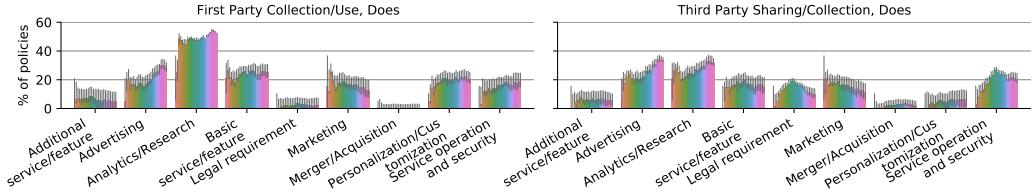


Fig. 9. Purposes for data collection for first- vs. third-parties.

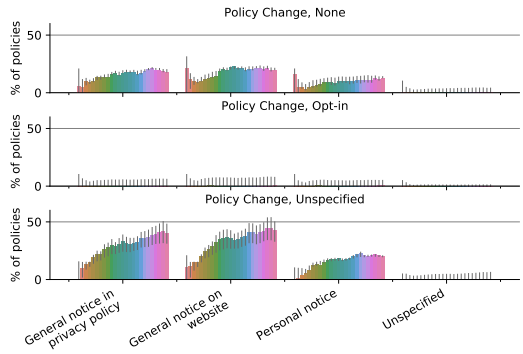


Fig. 10. How users are notified of privacy policy changes, by choices they are given.

4.4.3 Purpose. Figure 9 compares the purposes of data collection for first- and third-parties. The trends for first-party collection are similar to Figure 5b. We can see that slightly fewer purposes are given for third-party sharing, however, some purposes are more prevalent for third-party than first-party collection, including *advertising* and *legal requirements*. For both first and third parties, data collection for *advertising* and *analytics* has been decreasing slightly since 2018/19 (post-GDPR), but in both cases collection rates are still much higher than ten years ago.

4.5 Policy change

Figure 10 shows how users are notified of changes to privacy policies and what choices they are given when this happens. In 2021, 73% of policies include a statement about policy change. Of these, 38% state that changes will be announced by a notice in the privacy policy, 40% will post a notice on the website, and 22% will send a personal notice (the remaining policies leave the notification

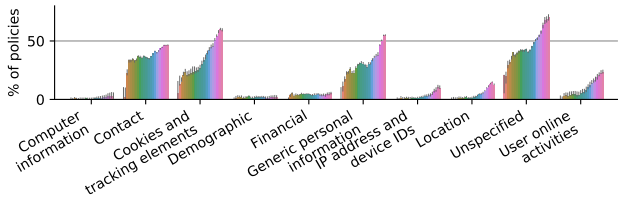


Fig. 11. Fraction of policies that offer choice or control mechanisms for specific personal information types. The figure omits four personal information types with fractions close to zero (health, personal identifier, social media data, survey data).

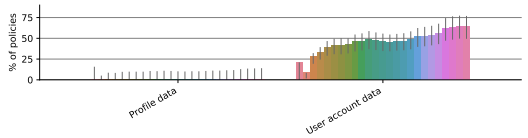


Fig. 12. Scope of data for which users are offered access, edit, and deletion rights.

type unspecified). Most users – the 78% notified on the website or in the policy – are unlikely to become aware of changes in privacy policies.

In addition, users are offered almost no meaningful choice when policies change: very few policies offer a new opt-in (middle row of Figure 10), whereas most policies leave the user's choice unspecified (bottom row).

4.6 User choice/control

Figure 11 shows the personal information types for which users are offered choice/control mechanisms. Over the past decade, an increasing percentage of policies offers choice or control mechanisms for almost all personal information types, which is a positive development. There is a particularly notable increase for *generic personal information* after 2018, most likely caused by the introduction of the GDPR. For example, the privacy policy of *airtable.com* introduced new text in June 2018 explaining users' choices: "You have many choices to access information we collect about you and about how we use or disclose that information. This section details many of those choices [...]."

However, comparing the percentage of policies that offer choice/control with those that collect specific personal information types (Figure 5a), we note that computer information, cookies, and user online activities, for example, are collected at much higher rates than choice/control is offered. In addition, the choices regarding cookies, although offered by more than half of policies, are insufficient to protect users from tracking: first, because choice or control mechanisms are rarely offered for *computer information*, *device identifiers*, and *personal identifiers*, which allow tracking of users via fingerprinting [44]; and second, because the use of dark patterns in cookie banners is widespread, which can lead to users making unintended choices [39].

4.7 User access, edit, deletion

Figure 12 shows the scope of data for which privacy policies offer access, edit, and deletion rights to users. User access is mostly offered for account data, i.e., data explicitly specified by users, but very rarely for profile data which is collected implicitly. This is concerning because profile data, such as interests inferred from analyzing user online activities, are widely used for targeting. Users

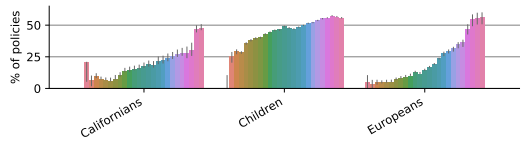


Fig. 13. Audience types described in privacy policies.

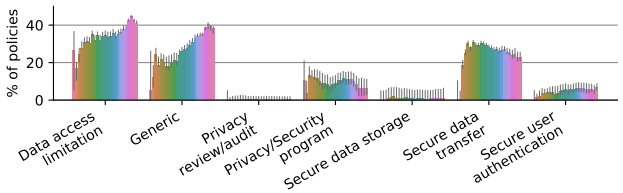


Fig. 14. Security measures described in privacy policies.

are very rarely offered to see or correct this data, and therefore have little opportunity to rectify any discriminatory targeting they may be subject to.

4.8 International/specific audiences

Figure 13 shows the audience types that are singled-out in privacy policies. Children are most frequently mentioned, due to longstanding legislation in many countries that require differential treatment of minors.

Following the introduction of the GDPR, we observe a 21% increase in mention of Europeans (from 35% in 2016 to 56% in 2021), and a similar 20% increase in mention of Californians (from 28% in 2018 to 48% in 2021) after the introduction of the CCPA. This indicates an increasing tailoring of privacy policies to specific audiences. As a result, other audiences may not benefit from the increased protections afforded to Europeans and Californians with the respective regulations. This increasing tailoring to specific audiences has already been observed in measurement studies where a user's location determines which tracking methods are used and what content is served [14], indicating an increasing fragmentation of the web.

4.9 Data security

Figure 14 shows the security measures mentioned in privacy policies. Most policies mention data access limitations and generic security measures. Statements about secure data transfer have been decreasing for more than a decade. This is possibly caused by the increased use and indeed normalization of TLS: the more common TLS is, the less websites see a need to state its use.

After 2018, we observe a decrease in mentions of privacy/security programs, data access limitations, and generic security measures. As we have hypothesized earlier in this Section, the reason for this may be that GDPR and CCPA do not require users to be informed about security measures – they only require security measures to be in place.

4.10 Obfuscating words

The use of obfuscating words is a measure for language accessibility [51]. Obfuscating words, such as *acceptable*, *significant*, *mainly*, or *predominantly* make text harder to understand because they reduce the clarity of statements. We use the list of obfuscating adjectives and adverbs published

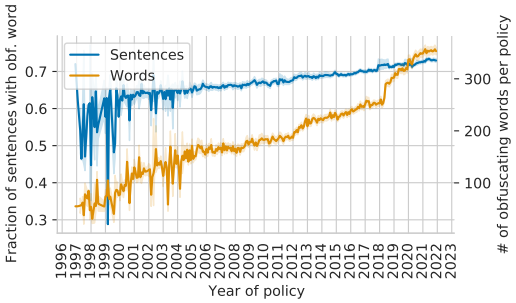


Fig. 15. Number of obfuscating words in policies and fraction of sentences with obfuscating words.

by Shipp and Blasco [51] to analyze to what extent privacy policies use obfuscating words. To do this, we count instances of obfuscating words in the policy text after preprocessing both our policy texts and the list of obfuscating words with the *gensim* preprocessor [46], which strips punctuation and numbers, removes whitespace and stop words, converts text to lowercase, and applies the PorterStemmer algorithm [43].

Figure 15 shows the average number of obfuscating words in privacy policies (orange) as well as the fraction of sentences that contain obfuscating words (blue). We can see that the absolute number of obfuscating words increased steadily before 2018, but then increased rapidly from a median of 229 in January 2018 to 304 in June 2020 ($p < .001$, Cohen's $d = 0.372$). In contrast, the fraction of sentences that contain obfuscating words continued its steady increase, indicating that the increase in the absolute number can be explained by increased policy length post-GDPR. In 2021, 72% of sentences in privacy policies contained at least one obfuscating word. This use of obfuscating words is another indicator that privacy policies are becoming less specific.

5 LIMITATIONS

We have presented a large-scale, longitudinal study of the contents of privacy policies. A major but common limitation of our study is our focus on English-language policies. To lift this limitation in a future study, we would need a labeled corpus of privacy policies for each target language.

Additional limitations stem from our process of retrieving privacy policies and from the methods for evaluating their content.

5.1 Retrieval of privacy policies

We focus on a longitudinal evaluation of privacy policies and rely on the Internet Archive's Wayback Machine because it is commonly seen as the most complete and reliable source of archived Internet sources [11, 29]. However, if a site is not archived by the Wayback Machine, we do not have its historical privacy policies available for analysis. Sites can be excluded from the Wayback Machine for different reasons, including a restriction in the robots.txt file.

Our crawler is sensitive to variations in how websites link to their privacy policies. Even though we attempt to find privacy policy content under various names, the process can fail if websites are creative in how they name their links and link titles. For example, if the link title is *here* (as in "read our privacy policy *here*"), and the URL includes the word "policy" but not "privacy", the crawler fails to find the policy. This limitation could be lifted by adding privacy policy links manually, as was done by Degeling et al. [16]. In addition, the crawler fails for some cases of regional differentiation, where the website asks the user to select a language before showing the privacy policy. If the

names and titles of policy links correspond to the chosen language ("English") instead of indicating presence of a privacy policy, the crawler fails to find the policy.

Finally, some websites choose to make their privacy policies available as a PDF download only. Our crawler detects these links and downloads the policy, but is not able to analyze the binary file.

5.2 Content evaluation

Our use of machine learning to segment and label the contents of privacy policies introduces limitations related to training data, segmentation, and classifier accuracy.

The OPP-115 corpus was published in 2016 and is based on privacy policies from that time. It is not clear whether classifiers trained based on this data are applicable to a longitudinal policy corpus from 1996 to 2021. For example, new regulations may have introduced new terms or new ways of phrasing data practices which are not present in the training data. In addition, the corpus only consists of 115 policies and as a result some data practices, e.g., in the *data retention* or *do not track* categories, occur infrequently. This makes it difficult to train accurate classifiers for some attributes.

The policy segments created by GraphSeg are sometimes longer than segments we would have created manually. This may negatively influence labeling accuracy.

6 CONCLUSION

In this paper, we have presented a longitudinal corpus of over 50,000 privacy policies from 1996 to 2021 and a detailed analysis of the data practices described in privacy policies. We find some improvements in the policy landscape after the introduction of the GDPR and CCPA, for example a 5–10% reduction in the collection of some personal information types, including contact information, cookies, and user online activities. However, we also identify several concerning trends, including the increasing use of location data, increasing use of implicitly collected data, lack of meaningful choice, lack of effective notification of privacy policy changes, increasing data sharing with unnamed third parties, and lack of specific information about security and privacy measures.

It is especially concerning that these data practices are obscured in lengthy policies that require university education to understand. Websites have shown that they can adopt standards for machine-readable formats quickly. For example, the ads.txt standard has reached 60% adoption rate within two years [8]. It is therefore not unreasonable to expect that privacy policies could be treated similarly. However, as the lack of adoption of P3P and the lack of respect for the DNT header show, it does not appear to be in the industry's interest to respect user privacy. The quantitative evidence presented in this paper shows that privacy policies are a mechanism that fails users and serves website owners.

As a result, we believe that three different approaches may together form a way forward: first, technical measures on the user-side that automatically classify privacy practices, match them against user preferences, and block unwanted data collection–in essence realizing P3P on the client-side; second, regulatory measures that mandate specific formats and locations for privacy policies and respect for specific privacy standards such as DNT; and third, measurement approaches that verify compliance of policies with actual data flows, e.g., building on work on flow-to-policy consistency [6].

REFERENCES

[1] Rami Al-Rfou. 2019. Pycld2: Python Bindings around Google Chromium's Embedded Compact Language Detection Library (CLD2). https://github.com/aboSamoor/pycld2

[2] Hamad Alamri, Carsten Maple, Saad Mohamad, and Gregory Epiphaniou. 2022. Do the Right Thing: A Privacy Policy Adherence Analysis of over Two Million Apps in Apple iOS App Store. *Sensors* 22, 22 (Jan. 2022), 8964.

https://doi.org/10.3390/s22228964

[3] Orlando Amaral, Sallam Abualhaija, Damiano Torre, Mehrdad Sabetzadeh, and Lionel Briand. 2022. AI-enabled Automation for Completeness Checking of Privacy Policies. *IEEE Transactions on Software Engineering* 48, 11 (Nov. 2022), 4647–4674. https://doi.org/10.1109/TSE.2021.3124332

[4] Ryan Amos, Gunes Acar, Elena Lucherini, Mihir Kshirsagar, Arvind Narayanan, and Jonathan Mayer. 2021. Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset. In *Proceedings of The Web Conference 2021 (WWW '21)*. ACM, Ljubljana, Slovenia, 22. https://doi.org/10.1145/3442381.3450048 arXiv:2008.09159

[5] Benjamin Andow, Samin Yaseer Mahmud, Wenyu Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. PolicyLint: Investigating Internal Privacy Policy Contradictions on Google Play. In *28th USENIX Security Symposium (USENIX Security 19)*. USENIX Association, Santa Clara, CA, USA, 585–602. https://www.usenix.org/conference/usenixsecurity19/presentation/andow

[6] Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions Speak Louder than Words: Entity-Sensitive Privacy Policy and Data Flow Analysis with PoliCheck. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX, online, 985–1002. https://www.usenix.org/conference/usenixsecurity20/presentation/andow

[7] Susanne Barth, Dan Ionita, and Pieter Hartel. 2022. Understanding Online Privacy - A Systematic Review of Privacy Visualizations and Privacy by Design Guidelines. *Comput. Surveys* 55, 3 (Feb. 2022), 63:1–63:37. https://doi.org/10.1145/3502288

[8] Muhammad Ahmad Bashir, Sajjad Arshad, Engin Kirda, William Robertson, and Christo Wilson. 2019. A Longitudinal Analysis of the Ads.Txt Standard. In *Proceedings of the Internet Measurement Conference (IMC '19)*. Association for Computing Machinery, Amsterdam, Netherlands, 294–307. https://doi.org/10.1145/3355369.3355603

[9] Frederik J. Zuiderveen Borgesius. 2016. Singling out People without Knowing Their Names – Behavioural Targeting, Pseudonymous Data, and the New Data Protection Regulation. *Computer Law & Security Review* 32, 2 (April 2016), 256–271. https://doi.org/10.1016/j.clsr.2015.12.013

[10] Trent J. Bradberry. 2022. Iterative-Stratification. https://github.com/trent-b/iterative-stratification

[11] Justin F. Brunelle, Mat Kelly, Hany SalahEldeen, Michele C. Weigle, and Michael L. Nelson. 2015. Not All Mementos Are Created Equal: Measuring the Impact of Missing Resources. *International Journal on Digital Libraries* 16, 3 (Sept. 2015), 283–301. https://doi.org/10.1007/s00799-015-0150-6

[12] Duc Bui, Kang G. Shin, Jong-Min Choi, and Junbum Shin. 2021. Automated Extraction and Presentation of Data Practices in Privacy Policies. *Proceedings on Privacy Enhancing Technologies* 2021, 2 (April 2021), 88–110. https://doi.org/10.2478/popets-2021-0019

[13] California State Legislature. 2018. The California Consumer Privacy Act of 2018. https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180AB375

[14] Adrian Dabrowski, Georg Merzdovnik, Johanna Ullrich, Gerald Sendera, and Edgar Weippl. 2019. Measuring Cookies and Web Privacy in a Post-GDPR World. In *Passive and Active Measurement (Lecture Notes in Computer Science)*, David Choffnes and Marinho Barcellos (Eds.). Springer International Publishing, Cham, 258–270. https://doi.org/10.1007/978-3-030-15986-3_17

[15] Caleb M. DeChant and Hamid Moradkhani. 2015. On the Assessment of Reliability in Probabilistic Hydrometeorological Event Forecasting. *Water Resources Research* 51, 6 (2015), 3867–3883. https://doi.org/10.1002/2014WR016617

[16] Martin Degeling, Christine Utz, Christopher Lentzsch, Henry Hosseini, Florian Schaub, and Thorsten Holz. 2019. We Value Your Privacy ... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. In *Network and Distributed Systems Security (NDSS) Symposium*. Internet Society, San Diego, CA, USA, 1–15. https://doi.org/10.14722/ndss.2019.23378

[17] Jose M. Del Alamo, Danny S. Guaman, Boni García, and Ana Diez. 2022. A Systematic Mapping Study on Automated Analysis of Privacy Policies. *Computing* 104, 9 (Sept. 2022), 2053–2076. https://doi.org/10.1007/s00607-022-01076-3

[18] European Parliament and Council of the European Union. 2016. General Data Protection Regulation. https://eur-lex.europa.eu/eli/reg/2016/679/oj

[19] Benjamin Fabian, Tatiana Ermakova, and Tino Lentz. 2017. Large-Scale Readability Analysis of Privacy Policies. In *Proceedings of the International Conference on Web Intelligence (WI '17)*. Association for Computing Machinery, Leipzig, Germany, 18–25. https://doi.org/10.1145/3106426.3106427

[20] Ming Fan, Le Yu, Sen Chen, Hao Zhou, Xiapu Luo, Shuyue Li, Yang Liu, Jun Liu, and Ting Liu. 2020. An Empirical Evaluation of GDPR Compliance Violations in Android mHealth Apps. In *2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE)*. IEEE, Coimbra, Portugal, 253–264. https://doi.org/10.1109/ISSRE5003.2020.00032

[21] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. 2016. Unsupervised Text Segmentation Using Semantic Relatedness Graphs. In **SEM 2016: The Fifth Joint Conference on Lexical and Computational Semantics : Proceedings of the Conference ; August 11-12 2016, Berlin, Germany*, Claire Gardent (Ed.). Association for Computational Linguistics,

Stroudsburg, Pa., 125–130. https://madoc.bib.uni-mannheim.de/41341

[22] Hana Habib, Yixin Zou, Aditi Jannu, Neha Sridhar, Chelse Swoopes, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2019. An Empirical Analysis of Data Deletion and Opt-Out Choices on 150 Websites. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, Santa Clara, CA, USA, 387–406. https://www.usenix.org/conference/soups2019/presentation/habib

[23] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. 2018. Polisis: Automated Analysis and Presentation of Privacy Policies Using Deep Learning. In *27th USENIX Security Symposium (USENIX Security 18)*. USENIX, Baltimore, MD, USA, 531–548. https://www.usenix.org/conference/usenixsecurity18/presentation/harkous

[24] Yili Hong. 2013. On Computing the Distribution Function for the Poisson Binomial Distribution. *Computational Statistics & Data Analysis* 59 (March 2013), 41–51. https://doi.org/10.1016/j.csda.2012.10.006

[25] Patrick Gage Kelley, Lucian Cesca, Joanna Bresee, and Lorrie Faith Cranor. 2010. Standardizing Privacy Notices: An Online Study of the Nutrition Label Approach. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 1573–1582. https://doi.org/10.1145/1753326.1753561

[26] Michael Kretschmer, Jan Pennekamp, and Klaus Wehrle. 2021. Cookie Banners and Privacy Policies: Measuring the Impact of the GDPR on the Web. *ACM Transactions on the Web* 15, 4 (July 2021), 20:1–20:42. https://doi.org/10.1145/3466722

[27] Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Los Angeles.

[28] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, Florian Schaub, and Norman Sadeh. 2020. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In *Proceedings of The Web Conference 2020 (WWW '20)*. Association for Computing Machinery, Taipei, Taiwan, 1943–1954. https://doi.org/10.1145/3366423.3380262

[29] Ada Lerner, Anna Kornfeld Simpson, Tadayoshi Kohno, and Franziska Roesner. 2016. Internet Jones and the Raiders of the Lost Trackers: An Archaeological Study of Web Tracking from 1996 to 2016. In *25th USENIX Security Symposium (USENIX Security 16)*. USENIX Association, Austin, TX, USA, 997–1013. https://www.usenix.org/conference/usenixsecurity16/technical-sessions/presentation/lerner

[30] Timothy Libert. 2018. An Automated Approach to Auditing Disclosure of Third-Party Data Collection in Website Privacy Policies. In *Proceedings of the 2018 World Wide Web Conference (WWW '18)*. International World Wide Web Conferences Steering Committee, Lyon, France, 207–216. https://doi.org/10.1145/3178876.3186087

[31] Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2020. The Privacy Policy Landscape After the GDPR. *Proceedings on Privacy Enhancing Technologies* 2020, 1 (Jan. 2020), 47–64. https://doi.org/10.2478/popets-2020-0004

[32] Frederick Liu, Shomir Wilson, Florian Schaub, and Norman Sadeh. 2016. Analyzing Vocabulary Intersections of Expert Annotations and Topic Models for Data Practices in Privacy Policies. In *2016 AAAI Fall Symposium Series*. Association for the Advancement of Artificial Intelligence, Arlington, VA, USA, 264–269. https://www.aaai.org/ocs/index.php/FSS/FSS16/paper/view/14099

[33] Miti Mazmudar and Ian Goldberg. 2020. Mitigator: Privacy Policy Compliance Using Trusted Hardware. *Proceedings on Privacy Enhancing Technologies* 2020, 3 (2020), 204–221. https://doi.org/10.2478/popets-2020-0049

[34] Aleecia M. McDonald and Lorrie Faith Cranor. 2008. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society* 4 (2008), 543. https://heinonline.org/HOL/Page?handle=hein.journals/isjlpsoc4&id=563&div=&collection=

[35] Aleecia M. McDonald, Robert W. Reeder, Patrick Gage Kelley, and Lorrie Faith Cranor. 2009. A Comparative Study of Online Privacy Policies and Formats. In *Privacy Enhancing Technologies*, Ian Goldberg and Mikhail J. Atallah (Eds.). Number 5672 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, Seattle, WA, USA, 37–55. http://link.springer.com/chapter/10.1007/978-3-642-03168-7_3

[36] Najmeh Mousavi Nejad, Pablo Jabat, Rostislav Nedelchev, Simon Scerri, and Damien Graux. 2020. Establishing a Strong Baseline for Privacy Policy Classification. In *ICT Systems Security and Privacy Protection (IFIP Advances in Information and Communication Technology)*, Marko Hölbl, Kai Rannenberg, and Tatjana Welzer (Eds.). Springer International Publishing, Maribor, Slovenia, 370–383. https://doi.org/10.1007/978-3-030-58201-2_25

[37] Razieh Nokhbeh Zaeem, Ahmad Ahbab, Josh Bestor, Hussam H. Djadi, Sunny Kharel, Victor Lai, Nick Wang, and K. Suzanne Barber. 2022. 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 (WSDM '22)*. Association for Computing Machinery, New York, NY, USA, 1593–1596. https://doi.org/10.1145/3488560.3502184

[38] Razieh Nokhbeh Zaeem and K. Suzanne Barber. 2021. 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 (CODASPY '21)*. Association for Computing Machinery, New York, NY, USA, 143–148. https://doi.org/10.1145/3422337.3447827

[39] Midas Nouwens, Ilaria Liccardi, Michael Veale, David Karger, and Lalana Kagal. 2020. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating Their Influence. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376321

[40] Alessandro Oltramari, Dhivya Piraviperumal, Florian Schaub, Shomir Wilson, Sushain Cherivirala, Thomas B. Norton, N. Cameron Russell, Peter Story, Joel Reidenberg, and Norman Sadeh. 2018. PrivOnto: A Semantic Framework for the Analysis of Privacy Policies. *Semantic Web* 9, 2 (Jan. 2018), 185–203. https://doi.org/10.3233/SW-170283

[41] Organisation for Economic Co-operation and Development (OECD). 2013. *The OECD Privacy Framework*. Technical Report. OECD. http://www.oecd.org/sti/ieconomy/oecd_privacy_framework.pdf

[42] Victor Le Pochat, Tom van Goethem, and Wouter Joosen. 2019. Rigging Research Results by Manipulating Top Websites Rankings.. In *26th Annual Network and Distributed System Security Symposium*. Internet Society, San Diego, CA, USA, 1–15. https://doi.org/10.14722/ndss.2019.23386

[43] M.F. Porter. 1980. An Algorithm for Suffix Stripping. *Program* 14, 3 (Jan. 1980), 130–137. https://doi.org/10.1108/eb046814

[44] Gaston Pugliese, Christian Riess, Freya Gassmann, and Zinaida Benenson. 2020. Long-Term Observation on Browser Fingerprinting: Users' Trackability and Perspective. *Proceedings on Privacy Enhancing Technologies* 2020, 2 (April 2020), 558–577. https://doi.org/10.2478/popets-2020-0041

[45] IK Reay, P. Beatty, S. Dick, and J. Miller. 2007. A Survey and Analysis of the P3P Protocol's Agents, Adoption, Maintenance, and Future. *IEEE Transactions on Dependable and Secure Computing* 4, 2 (April 2007), 151–164. https://doi.org/10.1109/TDSC.2007.1004

[46] Radim Řehůřek. 2021. Gensim: Topic Modelling for Humans. https://radimrehurek.com/gensim/

[47] Leonard Richardson. 2020. Beautiful Soup: We Called Him Tortoise Because He Taught Us. https://www.crummy.com/software/BeautifulSoup/

[48] Ira S. Rubinstein and Nathaniel Good. 2013. Privacy by Design: A Counterfactual Analysis of Google and Facebook Privacy Incidents. *Berkeley Technology Law Journal* 28 (2013), 1333. https://heinonline.org/HOL/Page?handle=hein.journals/berktech28&id=1367&div=&collection=

[49] David Sarne, Jonathan Schler, Alon Singer, Ayelet Sela, and Ittai Bar Siman Tov. 2019. Unsupervised Topic Extraction from Privacy Policies. In *Companion Proceedings of The 2019 World Wide Web Conference (WWW '19)*. Association for Computing Machinery, San Francisco, CA, USA, 563–568. https://doi.org/10.1145/3308560.3317585

[50] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the Stratification of Multi-label Data. In *Machine Learning and Knowledge Discovery in Databases (Lecture Notes in Computer Science)*, Dimitrios Gunopulos, Thomas Hofmann, Donato Malerba, and Michalis Vazirgiannis (Eds.). Springer, Berlin, Heidelberg, 145–158. https://doi.org/10.1007/978-3-642-23808-6_10

[51] Laura Shipp and Jorge Blasco. 2020. How Private Is Your Period?: A Systematic Analysis of Menstrual App Privacy Policies. *Proceedings on Privacy Enhancing Technologies* 2020, 4 (2020), 491–510. https://doi.org/10.2478/popets-2020-0083

[52] Rocky Slavin, Xiaoyin Wang, Mitra Bokaei Hosseini, James Hester, Ram Krishnan, Jaspreet Bhatia, Travis D. Breaux, and Jianwei Niu. 2016. Toward a Framework for Detecting Privacy Policy Violations in Android Application Code. In *Proceedings of the 38th International Conference on Software Engineering (ICSE '16)*. Association for Computing Machinery, New York, NY, USA, 25–36. https://doi.org/10.1145/2884781.2884855

[53] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2021. 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)*. Association for Computational Linguistics, Online, 6829–6839. https://doi.org/10.18653/v1/2021.acl-long.532

[54] William Stallings. 2020. Handling of Personal Information and Deidentified, Aggregated, and Pseudonymized Information Under the California Consumer Privacy Act. *IEEE Security Privacy* 18, 1 (Jan. 2020), 61–64. https://doi.org/10.1109/MSEC.2019.2953324

[55] 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 *Companion Proceedings of the The Web Conference 2018 (WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 163–166. https://doi.org/10.1145/3184558.3186969

[56] Kaushal Trivedi. 2022. Fast-Bert. Utterworks. https://github.com/utterworks/fast-bert

[57] Luca Verderame, Davide Caputo, Andrea Romdhana, and Alessio Merlo. 2020. On the (Un)Reliability of Privacy Policies in Android Apps. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, Glasgow, UK, 1–9. https://doi.org/10.1109/IJCNN48605.2020.9206660

[58] Isabel Wagner. 2022. Longitudinal Corpus of Privacy Policies [Data Set]. https://doi.org/10.5281/zenodo.7426577. https://doi.org/10.5281/zenodo.7426577

[59] 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, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. 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)*. Association for Computational Linguistics, Berlin, Germany, 1330–1340. https://doi.org/10.18653/v1/P16-1126

[60] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. https://doi.org/10.18653/v1/2020.emnlp-demos.6

[61] Fuman Xie, Yanjun Zhang, Chuan Yan, Suwan Li, Lei Bu, Kai Chen, and Zi Huang. 2022. Scrutinizing Privacy Policy Compliance of Virtual Personal Assistant Apps. In *37th IEEE/ACM International Conference on Automated Software Engineering (ASE 2022)*. ACM, Oakland Center, MI, USA, 12.

[62] Le Yu, Xiapu Luo, Xule Liu, and Tao Zhang. 2016. Can We Trust the Privacy Policies of Android Apps?. In *2016 46th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*. IEEE, Toulouse, France, 538–549. https://doi.org/10.1109/DSN.2016.55

[63] Razieh Nokhbeh Zaeem and K. Suzanne Barber. 2020. The Effect of the GDPR on Privacy Policies: Recent Progress and Future Promise. *ACM Transactions on Management Information Systems* 12, 1 (Dec. 2020), 2:1–2:20. https://doi.org/10.1145/3389685

[64] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Cameron Russell, and Norman Sadeh. 2019. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. *Proceedings on Privacy Enhancing Technologies* 2019, 3 (July 2019), 66–86. https://doi.org/10.2478/popets-2019-0037

[65] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman Sadeh, Steven M. Bellovin, and Joel Reidenberg. 2017. Automated Analysis of Privacy Requirements for Mobile Apps. In *Proceedings 2017 Network and Distributed System Security Symposium*. Internet Society, San Diego, CA, 1–15. https://doi.org/10.14722/ndss.2017.23034

A COMPARISON WITH PRINCETON-LEUVEN CORPUS

We have labeled the data practices of all policies in the Princeton-Leuven corpus [4] to evaluate how similar the data practices in their corpus are compared to our corpus. Figure 16 shows the percentage of policies in the Princeton-Leuven corpus addressing each of the top-level categories, i.e., the equivalent of Figure 3. We observe that the two figures are very similar up to 2019, which is the last year available in the Princeton-Leuven corpus. As another example, Figure 17 – the equivalent to Figure 5 – shows the collection mode for first-party data collection, and we also observe that the figures are very similar.

Checking the numeric differences in percentages of policies across all reported findings, we find that the average difference in the percentage of policies for each top-level category is 2.9%, and the average difference for analyses on the attribute level is 0.43%.

We present results from our corpus in the paper because it includes policies up to 2021, noting that 2020 and 2021 are interesting to study because new privacy regulation (the CCPA) came into force in early 2020. Given the similarity of Figures 3 and 16, and Figures 5 and 17, as well as the small numeric differences, we expect that the results presented in this paper generalize well to larger corpora of privacy policies.

B RESULTS FOR TOP-LEVEL CLASSIFIER

Tables 5 and 6 show the performance of the top-level classifier on the validation and test sets, respectively.

C RESULTS FOR ATTRIBUTE-LEVEL CLASSIFIERS

Tables 7 to 27 show classification results for attribute-level classifiers. Labels marked with an asterisk have been excluded from the analysis due to their low precision.

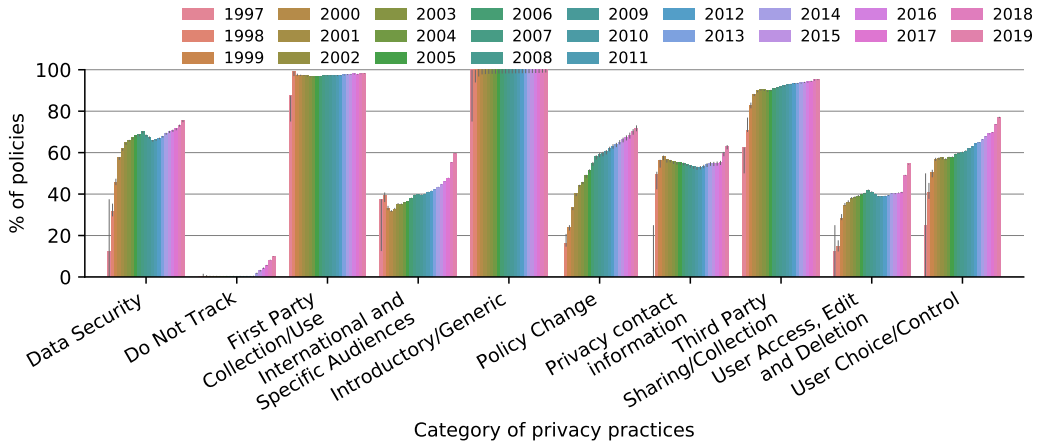
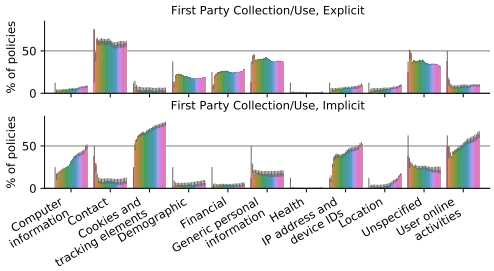
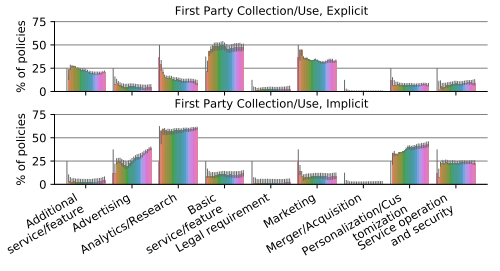


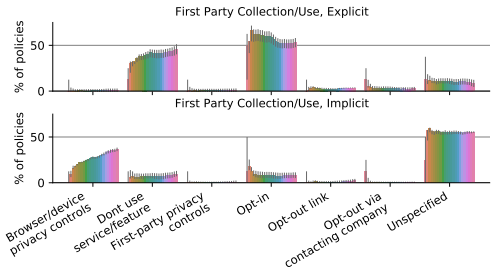
Fig. 16. Percentage of policies addressing each category of data practices in 1997–2019, based on the Princeton-Leuven corpus [4].



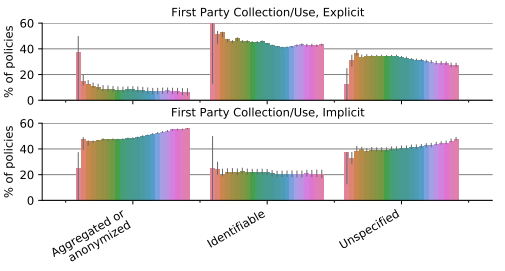
(a) Personal information type



(b) Purpose



(c) Type of choice and controls offered to users



(d) Identifiability of the collected data.

Fig. 17. Collection mode for first-party data collection, based on the Princeton-Leuven corpus.

Table 5. Top-level classifier, performance on validation dataset

	precision	recall	f1-score	support
Data Retention	0.833	0.357	0.500	14
Data Security	0.897	0.839	0.867	31
Do Not Track	1.000	0.667	0.800	6
First Party Collection/Use	0.866	0.886	0.876	175
International and Specific Audiences	0.929	0.975	0.951	40
Introductory/Generic	0.870	0.671	0.758	70
Policy Change	0.833	0.800	0.816	25
Practice not covered	0.471	0.381	0.421	21
Privacy contact information	0.793	0.719	0.754	32
Third Party Sharing/Collection	0.864	0.886	0.875	158
User Access, Edit and Deletion	0.800	0.833	0.816	24
User Choice/Control	0.800	0.750	0.774	48
micro avg	0.849	0.812	0.830	644
macro avg	0.830	0.730	0.767	644
weighted avg	0.847	0.812	0.825	644
samples avg	0.847	0.840	0.831	644

Table 6. Top-level classifier, performance on test dataset

	precision	recall	f1-score	support
Data Retention	0.636	0.500	0.560	14
Data Security	0.939	0.775	0.849	40
Do Not Track	1.000	1.000	1.000	3
First Party Collection/Use	0.909	0.883	0.896	248
International and Specific Audiences	0.852	0.821	0.836	56
Introductory/Generic	0.879	0.744	0.806	78
Policy Change	0.833	0.952	0.889	21
Practice not covered	0.692	0.360	0.474	25
Privacy contact information	0.879	0.707	0.784	41
Third Party Sharing/Collection	0.904	0.833	0.867	203
User Access, Edit and Deletion	0.870	0.833	0.851	24
User Choice/Control	0.827	0.816	0.821	76
micro avg	0.882	0.812	0.845	829
macro avg	0.852	0.769	0.803	829
weighted avg	0.879	0.812	0.842	829
samples avg	0.879	0.851	0.851	829

Table 7. F1 for attribute-level classifier: Access Scope (85 epochs)

	precision	recall	f1-score	support
Profile data	1.000	0.167	0.286	6
*Unspecified	0.600	1.000	0.750	3
User account data	0.950	1.000	0.974	19
micro avg	0.885	0.821	0.852	28
macro avg	0.850	0.722	0.670	28

Table 8. F1 for attribute-level classifier: Access Type (165 epochs)

	precision	recall	f1-score	support
Edit information	0.957	1.000	0.978	22
Unspecified	1.000	1.000	1.000	1
View	1.000	0.556	0.714	9
micro avg	0.966	0.875	0.918	32
macro avg	0.986	0.852	0.897	32

Table 9. F1 for attribute-level classifier: Action First-Party (40 epochs, data augmentation: segments labeled with different classes were combined into a new segment labeled with the union of the classes)

	precision	recall	f1-score	support
Collect in mobile app	0.975	0.928	0.951	83
Collect on mobile website	1.000	0.429	0.600	7
Collect on website	0.970	0.989	0.979	360
Unspecified	0.975	0.909	0.941	298
micro avg	0.972	0.945	0.959	748
macro avg	0.980	0.814	0.868	748

Table 10. F1 for attribute-level classifier: Action Third Party (62 epochs)

	precision	recall	f1-score	support
*Collect on first party website/app	0.688	0.500	0.579	22
Receive/Shared with	0.971	0.937	0.953	142
See	1.000	0.857	0.923	7
Track on first party website/app	0.923	0.923	0.923	26
Unspecified	1.000	0.200	0.333	5
micro avg	0.941	0.866	0.902	202
macro avg	0.916	0.683	0.742	202

Table 11. F1 for attribute-level classifier: Audience Type (117 epochs)

	precision	recall	f1-score	support
Californians	0.941	0.941	0.941	17
Children	0.968	0.968	0.968	31
Europeans	1.000	1.000	1.000	3
micro avg	0.961	0.961	0.961	51
macro avg	0.970	0.970	0.970	51

Table 12. F1 for attribute-level classifier: Change Type (91 epochs)

	precision	recall	f1-score	support
Privacy relevant change	1.000	0.714	0.833	7
Unspecified	0.917	1.000	0.957	11
micro avg	0.941	0.889	0.914	18
macro avg	0.958	0.857	0.895	18

Table 13. F1 for attribute-level classifier: Choice Scope (104 epochs)

	precision	recall	f1-score	support
Both	1.000	0.200	0.333	10
Collection	0.882	0.857	0.870	70
First party collection	0.900	0.692	0.783	13
First party use	0.848	0.812	0.830	48
Third party sharing/collection	0.864	0.679	0.760	28
*Third party use	0.000	0.000	0.000	9
*Unspecified	0.692	0.882	0.776	51
Use	1.000	0.535	0.697	43
micro avg	0.835	0.724	0.776	272
macro avg	0.773	0.582	0.631	272

Table 14. F1 for attribute-level classifier: Choice Type (90 epochs)

	precision	recall	f1-score	support
Browser/device privacy controls	0.900	0.923	0.911	39
Dont use service/feature	0.811	0.750	0.779	40
First-party privacy controls	0.857	0.400	0.545	15
Opt-in	0.909	0.811	0.857	74
Opt-out link	0.970	0.800	0.877	40
Opt-out via contacting company	0.923	0.828	0.873	29
*Third-party privacy controls	0.733	0.458	0.564	24
Unspecified	0.831	0.844	0.837	64
micro avg	0.875	0.778	0.824	325
macro avg	0.867	0.727	0.780	325

Table 15. F1 for attribute-level classifier: Collection Mode (50 epochs)

	precision	recall	f1-score	support
Explicit	0.938	0.882	0.909	68
Implicit	0.920	0.920	0.920	100
*Unspecified	0.696	0.762	0.727	21
micro avg	0.898	0.889	0.894	189
macro avg	0.851	0.855	0.852	189

Table 16. F1 for attribute-level classifier: Do Not Track policy (400 epochs)

	precision	recall	f1-score	support
Honored	1.000	1.000	1.000	1
Not honored	1.000	1.000	1.000	4
micro avg	1.000	1.000	1.000	5
macro avg	1.000	1.000	1.000	5

Table 17. F1 for attribute-level classifier: Does/Does Not (24 epochs)

	precision	recall	f1-score	support
Does	0.984	0.978	0.981	323
Does Not	0.944	0.829	0.883	41
micro avg	0.980	0.962	0.971	364
macro avg	0.964	0.904	0.932	364

Table 18. F1 for attribute-level classifier: Identifiability (100 epochs)

	precision	recall	f1-score	support
Aggregated or anonymized	0.912	0.963	0.937	54
Identifiable	0.976	0.910	0.942	134
Unspecified	0.767	0.958	0.852	48
micro avg	0.909	0.932	0.921	236
macro avg	0.885	0.944	0.910	236

Table 19. F1 for attribute-level classifier: Notification Type (150 epochs, data augmentation: segments labeled with different classes were combined into a new segment labeled with the union of the classes)

	precision	recall	f1-score	support
General notice in privacy policy	0.931	1.000	0.964	27
General notice on website	0.963	1.000	0.981	26
Personal notice	1.000	0.900	0.947	20
Unspecified	0.875	0.875	0.875	8
micro avg	0.951	0.963	0.957	81
macro avg	0.942	0.944	0.942	81

Table 20. F1 for attribute-level classifier: Personal Information Type (50 epochs, data augmentation: segments labeled with different classes were combined into a new segment labeled with the union of the classes)

	precision	recall	f1-score	support
Computer information	0.954	0.926	0.940	135
Contact	0.978	0.952	0.965	330
Cookies and tracking elements	0.985	0.997	0.991	339
Demographic	0.963	0.895	0.928	86
Financial	0.991	0.973	0.982	112
Generic personal information	0.953	0.950	0.951	577
Health	1.000	0.852	0.920	27
IP address and device IDs	1.000	0.960	0.980	176
Location	0.991	0.924	0.957	119
Personal identifier	1.000	0.548	0.708	31
Social media data	1.000	0.074	0.138	27
Survey data	1.000	0.200	0.333	15
Unspecified	0.882	0.848	0.865	395
User online activities	0.959	0.924	0.941	277
micro avg	0.958	0.917	0.937	2646
macro avg	0.975	0.787	0.828	2646

Table 21. F1 for attribute-level classifier: Purpose (65 epochs)

	precision	recall	f1-score	support
Additional service/feature	0.881	0.552	0.679	67
Advertising	0.941	0.909	0.925	88
Analytics/Research	0.887	0.910	0.899	78
Basic service/feature	0.909	0.738	0.814	122
Legal requirement	0.969	0.838	0.899	37
Marketing	0.924	0.839	0.880	87
Merger/Acquisition	1.000	0.895	0.944	19
Personalization/ Customization	0.933	0.764	0.840	55
Service operation and security	0.879	0.797	0.836	64
*Unspecified	0.589	0.825	0.688	40
micro avg	0.885	0.799	0.840	657
macro avg	0.891	0.807	0.840	657

Table 22. F1 for attribute-level classifier: Retention Period (161 epochs)

	precision	recall	f1-score	support
Indefinitely	1.000	1.000	1.000	3
*Limited	0.667	0.857	0.750	7
Unspecified	0.833	1.000	0.909	5
micro avg	0.778	0.933	0.848	15
macro avg	0.833	0.952	0.886	15

Table 23. F1 for attribute-level classifier: Retention Purpose (600 epochs)

	precision	recall	f1-score	support
Legal requirement	1.000	0.800	0.889	5
Perform service	0.800	0.800	0.800	5
Service operation and security	1.000	0.500	0.667	2
Unspecified	1.000	1.000	1.000	1
micro avg	0.909	0.769	0.833	13
macro avg	0.950	0.775	0.839	13

Table 24. F1 for attribute-level classifier: Security Measure (300 epochs, data augmentation: segments labeled with different classes were combined into a new segment labeled with the union of the classes)

	precision	recall	f1-score	support
Data access limitation	1.000	0.879	0.935	33
Generic	1.000	1.000	1.000	64
Privacy review/audit	1.000	0.333	0.500	3
Privacy/Security program	1.000	1.000	1.000	3
Secure data storage	1.000	0.429	0.600	7
Secure data transfer	1.000	1.000	1.000	26
Secure user authentication	1.000	0.500	0.667	4
micro avg	1.000	0.914	0.955	140
macro avg	1.000	0.734	0.815	140

Table 25. F1 for attribute-level classifier: Third Party Entity (90 epochs)

	precision	recall	f1-score	support
Named third party	0.868	0.787	0.825	75
Other part of company/affiliate	0.923	0.800	0.857	15
Public	0.857	0.667	0.750	9
Unnamed third party	0.885	0.959	0.921	121
Unspecified	0.625	0.625	0.625	8
micro avg	0.872	0.868	0.870	228
macro avg	0.832	0.767	0.796	228

Table 26. F1 for attribute-level classifier: User Choice (199 epochs)

	precision	recall	f1-score	support
None	0.875	0.875	0.875	8
Opt-in	1.000	0.500	0.667	2
Unspecified	0.800	1.000	0.889	4
micro avg	0.857	0.857	0.857	14
macro avg	0.892	0.792	0.810	14

Table 27. F1 for attribute-level classifier: User Type (19 epochs)

	precision	recall	f1-score	support
Unspecified	0.939	0.939	0.939	132
User with account	0.941	0.877	0.908	73
micro avg	0.940	0.917	0.928	205
macro avg	0.940	0.908	0.924	205