The University of Texas at Austin
Center for Identity

A Large Publicly Available Corpus of Website Privacy Policies Based on DMOZ

Razieh Nokhbeh Zaeem
K. Suzanne Barber

UTCID Report #20-16

November 2020

A Large Publicly Available Corpus of Website Privacy Policies Based on DMOZ

Razieh Nokhbeh Zaeem
K. Suzanne Barber
nokhbeh@utexas.edu
sbarber@identity.utexas.edu
Center for Identity at the University of Texas at Austin

ABSTRACT

Studies have shown website privacy policies are too long and hard to comprehend for their target audience. These studies and a more recent body of research that utilizes machine learning and natural language processing to automatically summarize privacy policies greatly benefit, if not rely on, corpora of privacy policies collected from the web. While there have been smaller annotated corpora of web privacy policies made public, we are not aware of any large publicly available corpus. We use DMOZ, a massive open-content directory of the web, and its manually categorized 1.5 million websites, to collect hundreds of thousands of privacy policies associated with their categories, enabling research on privacy policies across different categories/market sectors. We review the statistics of this corpus and make it available for research. We also obtain valuable insights about privacy policies, e.g., which websites post them less often. Our corpus of web privacy policies is a valuable tool at the researchers' disposal to investigate privacy policies. For example, it facilitates comparison among different methods of privacy policy summarization by providing a benchmark, and can be used in unsupervised machine learning to summarize privacy policies.

CCS CONCEPTS

• **Social and professional topics** → **Privacy policies**; • **Security and privacy** → *Usability in security and privacy*.

KEYWORDS

datasets, privacy policies, DMOZ, corpus

ACM Reference Format:
Razieh Nokhbeh Zaeem and K. Suzanne Barber. 2021. A Large Publicly Available Corpus of Website Privacy Policies Based on DMOZ. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Online privacy policies are legal documents through which websites share how they collect, use, disclose, and manage users' information. While privacy policies are virtually ubiquitous on the web, studies have shown [8, 12, 13] that privacy policies are too long and hard to comprehend for their intended users. As a result, users barely ever take the time and effort to thoroughly read these policies. These and similar studies (e.g., [1, 6]) of online privacy policies at scale require large corpora of privacy policies that are proportionate to the size of the web.

Furthermore, the research to circumvent the poor readability of privacy policies has been on the rise over the past decade. Any tool or research technique that addresses the length and poor readability of these policies, such as those that apply machine learning, natural language processing, and crowd-sourcing to automatically summarize privacy policies (e.g., [2, 4, 9, 17, 24, 25]), would require (or benefit from) large corpora of web privacy policies.

Many studies have privately gathered corpora of privacy policies and some have publicly shared them with the research community. However, as we discuss in Section 2, there is a lack of very large corpora of web privacy policies, that are proportionate in size to the number of privacy policies on the web. Mobile apps and their privacy policies have received much more attention and very large corpora of mobile privacy policies exist. Nevertheless, the privacy policies of websites and apps have meaningful differences in content. For example, cookies are more often associated with websites than mobile apps. On the other hand, mobile devices can usually access a finer grade location history and a different set of personally identifiable information (e.g., fingerprints) and preserving the privacy of this information involves different privacy settings.

While the common approach to collecting privacy policies (from the web or for mobile apps) has been to crawl the web (or Google Play), we take advantage of the massive open-content directory of web links in the DMOZ project. DMOZ (also known as the Open Directory Project) was[1] a directory of web links. A community of volunteers created, manually categorized, and maintained a collection of over 1.5 million links according to a hierarchical ontology. Our use of DMOZ has an advantage over simply crawling the web: it also provides the market sector/category of privacy policies, enabling research and comparison across different categories.

In this work we make the following contributions:

(1) We make available the first very large corpus of web privacy policies with over 100K privacy policies.
(2) In lieu of simply crawling the web, we base our collection of privacy policies on a manually categorized hierarchy of over 1.5 million web links from DMOZ. As a result, we enable research on privacy policies across categories.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

[1] DMOZ was closed in 2017, and was inherited by another website—Curlie at https://curlie.org. We use the final collection of DMOZ links from 2017 because Curlie links are not available for bulk download.

(3) We make the details of our data collection publicly available in order to enhance reproducibility.

The rest of this paper is organized as follows. Section 2 reviews the related work on publicly available corpora of privacy policies and identifies the gap this paper seeks to fill. Section 3 elaborates on the process of building the privacy policy corpus and provides the link to download it. Finally, Section 4 concludes the paper with some final remarks about future work.

2 RELATED WORK

Applying machine learning, natural language processing, and crowdsourcing techniques to digest privacy policies has grown in popularity over the past decade. The existence and availability of privacy policy corpora is the foundation of most of the aforementioned techniques. In this section, we cover related work on publicly available privacy policy corpora.

Some researchers have dedicated their attention to manually annotating privacy policy corpora. The fact that these corpora are manually annotated by researchers or crowd-sourced workers is prohibiting them from including more than a couple hundred privacy policies. Two of the most widely used privacy policy corpora are OPP-115 [20] and APP-350 [25], containing 115 and 350 privacy policies respectively. There exist other manually labeled corpora containing less than 1K policies (e.g., 236 policies [4], 400 policies [21–24], 45 policies [18], and 64 policies [5]). While valuable for supervised machine learning, these corpora fall short when it comes to unsupervised machine learning, natural language processing, and testing/validation because of their limited size.

Among corpora gathered from the web but without annotation, some are larger but not available to the public. For instance, corpora of 9,295 policies [26] and 130K policies [9]. Some smaller corpora of web privacy policies are made publicly available, for example, a corpus of 1,010 policies [14].

Interestingly, there are multiple corpora of *mobile app* (particularly Google Play) privacy policies that are available. For example, Kumar et al. used 150K [11] policies from Google Play and Sunyaev et al. [16] considered the privacy policies of 183 health iOS and Android apps. Notable is the MAPS framework [25], which evaluated the privacy policies of over one million Android apps and released 441,626 app privacy policies with their app categories. Mobile app privacy policies have received a lot of attention, arguably, among other reasons, because of the research that analyzes a mobile app's code alongside its privacy policy [1–3, 7, 10]. Privacy policies of websites, nonetheless, are equally important. There are meaningful differences between the contents of web privacy policies and mobile app privacy policies. For instance, the use of cookies is more applicable to web privacy policies or, generally speaking, mobile apps can obtain finer grade location information when compared to websites and should address how they deal with this location information in their privacy policies. Despite these differences, there is a lack of sizable *web* privacy policy corpora.

Srinath and his colleagues [15] created a corpus of one million privacy policies. They crawled the web for links with the words "privacy" or "data protection" in the URL itself, similar to how we looked for URLs. Our work differs from their privacy policy corpus: (1) Their corpus, in its entirety, is not publicly available as of this writing (10/11/2020); and (2) Their work does not take into account, or in anyway group privacy policies across, categories.

3 PRIVACY POLICY CORPUS

In this section, we review our pipeline of collecting and cleaning up the corpus of privacy policies.

3.1 DMOZ

DMOZ (formerly known as the Open Directory Project) is a huge manually edited directory of the web, which contains over 1.5 million URLs in 15 categories. Even though DMOZ was closed in 2017, and was inherited by another website (Curlie[2]), its non-editable mirror remains available[3] and is still used for various purposes, including by the research community. We use the final collection of DMOZ links from 2017[4].

Our copy of the DMOZ dataset contains 1,562,978 URLs of websites across 15 categories. These categories, sorted alphabetically, are: (1) adult, (2) arts, (3) business, (4) computers, (5) games, (6) health, (7) home, (8) kids, (9) news, (10) recreation, (11) reference, (12) science, (13) shopping, (14) society, and (15) sports. The first row of Table 1 is the number of URLs in each category in DMOZ.

We accessed all the DMOZ URLs from September 25 to 27, 2020. As it is expected, not all of the DMOZ pages were working and we had to catch a variety of exceptions thrown by the DMOZ URLs. The second section of Table 1 shows the exceptions caught in each category. The table also shows, in its third section, the number of DMOZ pages that we successfully retrieved, which is (in each category) equal to the number of DMOZ URLs minus the total number of DMOZ exceptions (in that category). We were able to successfully retrieve 819,865 DMOZ URLs. We think the fact that this dataset has not been maintained in about three years (2017 to 2020) plays a role in the high number of broken DMOZ URLs. Unfortunately, Curlie links are not available for bulk download.

3.2 Finding Links to Privacy Policies of the DMOZ dataset

As the next step, we identified links to potential privacy policies on each of the pages in DMOZ. As others [15] have pointed out, common names for links to privacy policies include keywords like "Privacy Policy", "Privacy Notice", and "Data Protection", and these words are usually reflected in the URL itself. Looking for patterns of these words in the URLs has been common practice in the collection of privacy policies [15, 25]. In previous work [15, 25], however, the selection of these words has been largely ad-hoc. In order to collect a *comprehensive* set of keywords, we manually evaluated a previously available corpus of 400 [24] privacy policies with their URLs to find keywords [21] in URLs. The list of keywords we distill and use is as follows: privacy, legal, conditions, policy, policies, terms, help.

Parsing each page of DMOZ, we accessed any link on the page that had at least one of the above keywords in its target URL. Note that one URL in DMOZ may result in multiple privacy policy candidates, since we collect *all* the links on the DMOZ page with *at least* one of the above keywords. The fourth section of Table 1 further

[2]https://curlie.org
[3]http://dmoztools.net
[4]From https://www.kaggle.com/shawon10/url-classification-dataset-dmoz

Table 1: DMOZ URLs and candidate privacy policies.

	All DMOZ	Adult	Arts	Business	Computers	Games	Health	Home	Kids	News	Recreation	Reference	Science	Shopping	Society	Sports
DMOZ URLs	1,562,978	35,325	253,840	240,177	117,962	56,477	60,097	28,269	46,182	8,989	106,586	58,247	110,286	95,270	243,943	101,328
DMOZ URLError	305,900	13,844	40,664	67,743	20,309	8,308	9,791	6,299	8,140	1,291	19,024	11,883	20,485	18,222	37,264	22,633
DMOZ HTTPError	410,509	7,108	81,163	35,825	26,258	19,183	18,594	8,161	14,605	2,236	29,398	16,716	34,043	16,629	68,117	32,473
DMOZ timeout (3s)	23,085	246	3,255	4,381	1,726	349	1,116	409	619	208	1,640	673	1,339	1,465	3,852	1,807
DMOZ ConnectionResetErr.	2,315	15	370	371	119	32	104	23	207	5	163	97	220	89	409	91
DMOZ RemoteDisconnected	1,023	35	112	207	80	30	35	170	16	4	48	23	38	69	92	64
DMOZ InvalidURL	128	0	16	10	5	8	19	3	1	2	4	6	19	5	26	4
DMOZ IncompleteRead	40	0	6	10	3	0	1	0	0	2	2	2	1	3	8	2
DMOZ UnboundLocalError	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
DMOZ SSLWantReadError	31	0	6	6	2	0	5	1	1	0	3	1	1	1	4	0
DMOZ UnicodeEncodeError	29	0	3	2	0	0	1	0	2	0	1	5	2	2	9	2
DMOZ BadStatusLine	35	0	17	5	2	0	0	1	1	0	1	0	0	2	4	2
DMOZ TypeError	5	0	0	1	1	0	1	0	0	0	0	0	1	0	1	0
DMOZ UnicodeError	5	0	0	2	0	0	0	0	0	0	0	0	1	1	1	0
DMOZ HTTPException	2	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0
DMOZ ConnectionAbortedErr.	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
DMOZ ValueError	3	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0
DMOZ UnknownProtocol	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Retrieved DMOZ Page	819,865	14,077	128,228	131,611	69,455	28,567	30,430	13,202	22,590	5,241	56,302	28,841	54,136	58,780	134,155	44,250
Candidate Policy URLs	938,468	7,791	162,774	112,658	78,958	40,400	62,496	19,588	39,088	13,178	37,636	49,917	61,968	50,951	157,117	43,948
Avg. #Candidate Policies on a DMOZ Page	1.14	0.55	1.27	0.86	1.14	1.41	2.05	1.48	1.73	2.51	0.67	1.73	1.14	0.87	1.17	0.99
Duplicate Cand. Policy URLs	554,714	4,502	139,216	35,898	44,778	35,648	31,666	12,147	30,017	5,903	18,502	25,631	36,270	16,231	90,207	28,098
Percentage of Duplicate Candidate Policies	59%	58%	86%	32%	57%	88%	51%	62%	77%	45%	49%	51%	59%	32%	57%	64%
Unique Cand. Policy URLs	383,754	3,289	23,558	76,760	34,180	4,752	30,830	7,441	9,071	7,275	19,134	24,286	25,698	34,720	66,910	15,850

displays the number of candidate privacy policy URLs collected from the retrieved DMOZ pages.

An interesting observation pertains to the average number of candidate privacy policies on a given successfully retrieved DMOZ page across categories (shown in the fourth section of Table 1). This average is 1.14 for the entire DMOZ, i.e., there is almost one potential privacy policy link on every page. This average, however, varies drastically across categories of DMOZ, with a minimum of 0.55 for the pages in the Adult category to a maximum of 2.51 for the News category. Admittedly, we are collecting multiple candidate privacy policy links from one page in categories like News. These links might be to terms of service, privacy policies, or other non-policy pages that simply happen to include one of our keywords, such as "legal". We remove these non-policy pages in Section 3.6.

Nonetheless, the variation in the number of privacy policy links on a page across categories is insightful. We may conclude that the Adult category is particularly inferior in readily providing links to privacy policies. There is only a couple of academic studies focusing on Adult website privacy policies, but they have found similar results. For example, Vallina et al. [19] investigated 6,843 Adult websites and found that as low as 16% of them have privacy policies and Maris et al. considered 22,484 Adult websites to find that as low as 17% have privacy policies. It is likely that they are under-counting, as they do not look for "uncommon phrasing" in links to privacy policies, and we are over-counting, as (in this section) we are looking at *all candidate* policies. We do not recalculate the number of privacy policies per page after the removal of duplicate URLs, which unfairly decreases the number of privacy policies counted per page: sharing privacy policies is fine.

3.3 Removing Duplicate Privacy Policy URLs

We observe that there are duplicate URLs in the set of candidate privacy policies. In each category, we remove these duplicates, but do not remove a URL that appears in more than one category. Our rationale is that a DMOZ page and its privacy policy might be put under two categories, because they rightfully belong to both. However, the duplication of URLs inside a category is redundant.

The fifth section of Table 1 lists the number of duplicate URLs among the candidate privacy policy URLs, and the percentage of the candidate privacy policies that were duplicate. In fact, a high percentage (59% in the entire dataset, ranging from 32% in Shopping and Business to 88% in Games) of collected candidate privacy policy URLs were duplicate. We believe that such a high percentage of shared privacy policies is due to shared parent companies and widespread use of template policies.

After removing duplicates, the sixth section of Table 1 shows the number of unique candidate privacy policies. In the end, we have about 400K *candidate* privacy policy *URLs*.

3.4 Privacy Policy Text

The next step is to obtain the privacy policy text for the unique candidate URLs. We (1) access the candidate URL, and (2) obtain its main body after removing boilerplate. Table 2 starts with the last line of Table 1. Firstly, accessing the candidate URLs throws some extensions, as collected in the second section of Table 2. Once those exceptions are removed from unique candidate policy URLs, the third section of the table displays the number of successfully accessed candidate URLs. Secondly, we obtained the body of the candidate policy URL and removed boilerplate. We sometimes found

Table 2: Candidate and final privacy policies.

	All DMOZ	Adult	Arts	Business	Computers	Games	Health	Home	Kids	News	Recreation	Reference	Science	Shopping	Society	Sports
Unique Cand. Policy URLs	383,754	3,289	23,558	76,760	34,180	4,752	30,830	7,441	9,071	7,275	19,134	24,286	25,698	34,720	66,910	15,850
Cand. Policy URLError	15,535	43	1,233	7,663	586	73	260	92	91	76	225	1,145	952	1,245	1,635	216
Cand. Policy HTTPError	24,808	184	1,870	3,537	2,312	402	2,542	699	591	626	1,091	1,970	1,828	1,772	4,453	931
Cand. Policy timeout	6,150	32	380	1,296	421	47	638	105	71	83	342	416	360	473	1,176	310
Cand. Policy ConnectionResetError	295	2	36	25	13	30	28	10	4	2	10	27	39	9	53	7
Cand. Policy RemoteDisconnected	21	0	2	3	4	0	1	0	0	0	1	3	0	1	5	1
Cand. Policy InvalidURL	488	3	35	62	62	15	19	22	14	13	25	42	25	13	111	27
Cand. Policy IncompleteRead	20	0	3	5	1	0	4	0	0	0	0	0	0	2	5	0
Cand. Policy UnboundLocalError	6	0	0	2	0	0	2	0	0	0	0	1	1	0	0	0
Cand. Policy SSLWantReadError	22	1	1	2	2	0	4	1	0	1	1	2	2	0	4	1
Cand. Policy BadStatusLine	7	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0
Cand. Policy TypeError	13	0	0	5	2	0	0	0	0	0	0	2	0	0	2	2
Cand. Policy HTTPException	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Cand. Policy Unknown	150	0	17	7	6	0	7	0	2	37	1	4	12	4	53	0
Successfully Accessed Cand. Policy	336,238	3,024	19,974	64,153	30,771	4,185	27,325	6,512	8,298	6,437	17,437	20,674	22,479	31,201	59,413	14,355
Empty Body of Page	74,130	1,030	3,830	13,654	8,493	821	5,901	1,287	2,650	876	3,922	4,220	4,991	5,735	13,787	2,933
French (fr)	1,750	83	122	342	106	12	43	39	230	6	239	64	125	96	177	66
Dutch (nl)	946	43	66	184	122	18	28	6	180	4	59	48	47	40	79	22
Swedish (sv)	179	2	5	46	29	5	8	5	12	0	4	3	20	6	11	23
Portuguese (pt)	76	2	4	5	16	0	1	5	7	1	7	2	2	4	17	3
Norwegian (no)	39	3	2	13	1	0	0	2	2	0	0	2	1	2	8	3
Croatian (hr)	25	1	2	2	6	1	1	1	2	1	1	2	1	0	3	1
Vietnamese (vi)	19	2	0	6	1	0	0	0	0	0	0	0	6	1	3	0
Catalan (ca)	213	12	17	38	13	4	8	4	7	0	14	13	10	17	38	18
German (de)	732	88	55	128	89	11	36	1	53	3	45	18	57	55	81	12
Italian (it)	1,189	37	77	580	68	15	17	9	46	7	100	24	63	66	44	36
Danish (da)	302	5	9	59	43	8	9	4	16	0	20	3	16	46	41	23
Tagalog (tl)	44	1	2	4	9	1	1	9	2	1	3	3	1	1	5	1
Slovenian (sl)	8	1	1	1	1	1	0	1	0	0	1	0	0	0	1	0
Finnish (fi)	47	2	8	2	6	0	1	3	7	0	6	2	1	0	8	1
Romanian (ro)	30	3	3	2	2	1	0	6	1	0	0	0	1	1	9	1
Indonesian (id)	182	1	26	22	20	5	5	7	2	2	14	3	10	19	37	9
Spanish (es)	699	38	45	160	29	4	45	8	83	2	46	17	27	15	153	27
Afrikaans (af)	46	0	2	3	3	1	0	3	4	0	2	1	5	0	22	0
Welsh (cy)	110	0	15	6	16	5	2	3	11	0	15	6	11	0	14	6
Turkish (tr)	5	0	1	1	1	0	0	0	1	0	1	0	0	0	0	0
Polish (pl)	294	0	39	13	14	6	28	2	8	2	19	23	15	2	110	13
Swahili (sw)	21	0	1	4	2	0	1	1	1	1	2	0	2	1	4	1
Somali (so)	14	0	3	1	2	0	1	1	0	0	1	0	2	0	3	0
Slovak (sk)	2	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0
Estonian (et)	34	0	0	0	5	1	0	5	0	0	0	2	10	1	9	1
Hungarian (hu)	27	0	0	0	3	0	1	0	1	0	3	2	4	4	8	1
Albanian (sq)	12	0	0	0	5	4	0	0	0	0	0	0	0	0	2	1
Lithuanian (lt)	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Czech (cs)	3	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0
Number of English Cand. Policies	255,059	1,670	15,639	48,876	21,666	3,261	21,188	5,099	4,972	5,531	12,913	16,216	17,051	25,088	44,736	11,153
Percentage of English Candidates	97%	84%	97%	97%	97%	97%	99%	98%	88%	99%	96%	99%	98%	99%	98%	98%
Privacy Policies Selected from Candidates	**117,460**	**897**	**8,483**	**27,029**	**11,057**	**1,642**	**8,258**	**2,016**	**1,875**	**3,124**	**6,335**	**5,837**	**7,390**	**13,487**	**14,687**	**5,343**
Number of Policies with Unique Text	104,093	752	7,278	24,530	9,844	1,462	7,468	1,764	1,584	2,151	5,768	5,105	5,932	12,378	13,353	4,724
Percentage of Policies with Unique Text	89%	84%	86%	91%	89%	89%	90%	88%	84%	69%	91%	87%	80%	92%	91%	88%

that what was left was an empty body, as shown in the fourth section of Table 2.

3.5 Removing Non-English Pages

Next, we utilize the Python library *langdetect*[5], a language detection library ported from Google's language-detection, to identify the language of the text. We remove non-English text because we are not able to read and validate if the links are privacy policies after automatic classification (Section 3.6). The fifth section of Table 2 lists the number of non-English documents, followed by the number of English documents in the sixth section. The number of successfully accessed candidate policies is split between empty bodies, non-English documents, and English documents. We see that the documents we fetched are predominately in English (97%). Others have found similar, but not identical, distributions in the languages of privacy policies when crawling the web. For example, Srinath et al. [15] crawled the web for privacy policies and found non-English languages Italian, Dutch, German, Spanish, and French in their top ten common languages. They also found Asian languages such as Japanese, Russian, Chinese, and Korean that are notably missing from our dataset. Even though DMOZ frequently includes European languages, we still find the absence of these Asian languages

[5]https://pypi.org/project/langdetect

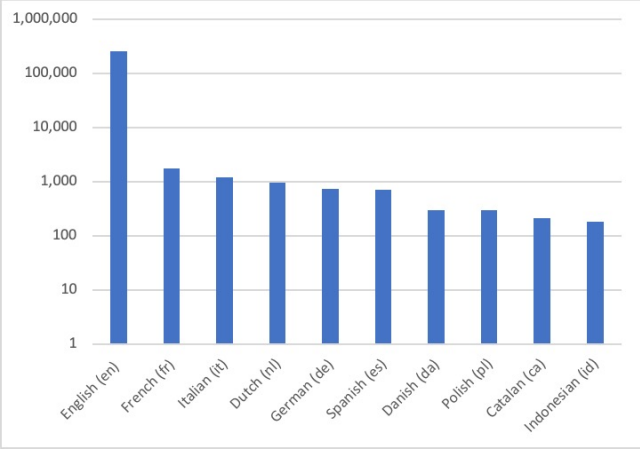

Figure 1: Top ten most frequently detected languages in candidate privacy policies (Y-axis in logarithmic scale).

curious as DMOZ does include websites in all of these languages. We do see, however, the Indonesian language present in our top ten. Figure 1 depicts the ten most frequently detected languages in our candidate privacy policies. (The Y axis is in logarithmic scale.)

3.6 Separating Actual Privacy Policies

The most challenging step in the creation of our corpus is to find out which of the candidates is indeed a privacy policy. We considered a variety of classification methods, from regular expression to machine learning. However, an effective classification method does not always have to be sophisticated—it just needs to deliver good results for the problem at hand. After the manual investigation of the candidates, we found that the expectation that a privacy policy discusses privacy can serve as a great classifier. We discovered that by keeping candidate policies that mention the word "privacy" more than twice, we are able to gather a corpus of privacy policies with very high precision and recall. The rationale behind looking for the word privacy to appear "more than twice" is to eliminate pages that link to privacy policies in headers and footers (two places) but are not privacy policies. The fact that we have already collected a set of documents with particular keywords in their URLs contributes to the success of this classification method.

The sevenths section of Table 2 shows the final number of policies after classification, with the number of policies released in our dataset in bold. We count the number of repeated *text* among these privacy policies but do not remove duplicates in our final dataset (section seven of the table). The majority of text extracted are unique, with 11% duplicate text in the entire dataset, stemming from using template policies. We report the number of repeated text but do not remove them as sharing privacy policy text is fine.

Our classification method, as basic as it is, yields high precision and recall in the final privacy policy corpus of 117,460 policies. We manually labeled a random 1% sample of the English candidate privacy policies in each category, adding up to 2,552 candidates. The first row of Table 3 shows the number of manually labeled candidates in each category. We thoroughly studied the content (not only the title) of these randomly selected policies to identify them as true privacy policy (including terms of service with information on privacy) or non-privacy policy. The second section of the table displays the details of classification. **True Positive:** The text was indeed a privacy policy, and was correctly identified as such by our classification. **True Negative:** The text was not a privacy policy, and was correctly identified as such by our classification. **False Positive:** The text was not a privacy policy but was incorrectly identified as one by our classification. **False Negative:** The text was a privacy policy but was incorrectly identified as not a privacy policy by our classification. True positive and true negative rates show the strength of our classification and should be considered together. One reason for the current true positive and true negative rates is that many of the initial links were not privacy policies, as we aimed to collect more links at the begging with a comprehensive set of keywords and narrow down as we go. In the end, our classification successfully detected/deleted non-privacy policies.

The last section of the table includes precision, recall, and F-1 scores. The weighted average precision of 95%, recall of 0.94%, and F-1 score of 0.95% indicate very good classification results. Across categories, the lowest F-1 scores belong to two categories: News and Society—the very two that are expected to report news and articles on privacy. These articles discuss privacy and might have related keywords in their URLs, which misleads our collection and classification. Yet, the F-1 scores for both are still over 90%.

Overall, the distribution of candidate policies across the categories (last section of Table 2) is slightly different than the distribution of URLs in DMOZ, which we assume to be somewhat representative of the entire web. The most populated categories in DMOZ are Arts, Society, and Business. The most privacy policies we found are in Business, Computers, Shopping, and Society. These results are interesting but rather intuitive: Arts websites discuss privacy policies less often than Computers and Shopping websites.

Our corpus is available for free download at https://github.com/UTCID/DMOZ-Privacy-Policy-Corpus-CODASPY21.

We saved each URL with its privacy policy text in a file named after its domain URL. If there were multiple URLs sharing the same domain, we numbered the files. Privacy practices are often scattered across pages (e.g., cookie consent pages, privacy policies, and terms of service). Hence, there might be multiple pages that cover privacy practices that should be considered together. Therefore, there sometimes are multiple files for the same domain URL in our dataset, sharing the same domain URL file name but numbered sequentially, to make it easier to identify which files belong to the same website. The full unique URL is included in the file.

4 CONCLUSIONS AND FUTURE WORK

We discussed the creation of a corpus of over 100K web privacy policies based on the open web directory DMOZ and made our corpus publicly available. Our corpus of web privacy policies is an order of magnitude bigger than similar available corpora. We manually labeled 1% of the candidate privacy policies–2,552 policies–and measured that 95% (the weighted average precision) of the corpus are indeed privacy policies. Along the way, we also made various observations as we took advantage of manually categorized DMOZ links in 15 categories. For example, the websites in the Adult category are less than half as likely to have a link to a potential privacy policy, when compared to the average DMOZ website. As another

Table 3: Validation of privacy policy classification.

	All DMOZ	Adult	Arts	Business	Computers	Games	Health	Home	Kids	News	Recreation	Reference	Science	Shopping	Society	Sports
Manually Labeled (Random 0.01 of English Cand. Policies)	2552	17	156	489	217	33	212	51	50	55	129	162	171	251	447	112
True Positive	1093	9	71	265	95	12	77	17	25	32	79	60	50	101	157	43
False Positive	55	0	1	9	6	0	3	0	0	3	1	3	2	3	24	0
True Negative	1340	8	75	204	114	19	127	34	24	16	44	96	113	141	259	66
False Negative	64	0	9	11	2	2	5	0	1	4	5	3	6	6	7	3
Precision	0.95	1.00	0.99	0.97	0.94	1.00	0.96	1.00	1.00	0.91	0.99	0.95	0.96	0.97	0.87	1.00
Recall	0.94	1.00	0.89	0.96	0.98	0.86	0.94	1.00	0.96	0.89	0.94	0.95	0.89	0.94	0.96	0.93
F-1	0.95	1.00	0.93	0.96	0.96	0.92	0.95	1.00	0.98	0.90	0.96	0.95	0.93	0.96	0.91	0.97

example, we saw that 59% of potential privacy policy URLs in the entire DMOZ dataset are repetitive. We think that such a high percentage of duplicate privacy policies is due to shared parent companies and widespread use of template policies. Our corpus is a valuable dataset for privacy policy research and study. It provides a benchmark to compare privacy policy summarization methods and enhances researchers' ability to take advantage of machine learning techniques that require bigger corpora. A future work avenue is to augment our corpus with more granular subcategories of DMOZ.

ACKNOWLEDGMENTS

We thank the Strategic Partners of the Center for Identity (http://identity.utexas.edu/strategic-partners) for their contributions.

REFERENCES

[1] 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*. 585–602.

[2] 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)*. 985–1002.

[3] Steven Arzt, Siegfried Rasthofer, Christian Fritz, Eric Bodden, Alexandre Bartel, Jacques Klein, Yves Le Traon, Damien Octeau, and Patrick McDaniel. 2014. Flowdroid: Precise context, flow, field, object-sensitive and lifecycle-aware taint analysis for android apps. *Acm Sigplan Notices* 49, 6 (2014), 259–269.

[4] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, et al. 2020. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In *Proceedings of The Web Conference 2020*. 1943–1954.

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

[6] 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. *Informatik Spektrum* 42, 5 (2019), 345–346.

[7] William Enck, Peter Gilbert, Seungyeop Han, Vasant Tendulkar, Byung-Gon Chun, Landon P Cox, Jaeyeon Jung, Patrick McDaniel, and Anmol N Sheth. 2014. TaintDroid: an information-flow tracking system for realtime privacy monitoring on smartphones. *ACM Transactions on Computer Systems (TOCS)* 32, 2 (2014), 1–29.

[8] Mark A Graber, Donna M D Alessandro, and Jill Johnson-West. 2002. Reading level of privacy policies on internet health web sites. *Journal of Family Practice* 51, 7 (2002), 642–642.

[9] 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*. 531–548.

[10] Daniel Kales, Christian Rechberger, Thomas Schneider, Matthias Senker, and Christian Weinert. 2019. Mobile private contact discovery at scale. In *28th USENIX Security Symposium (USENIX Security 19)*. 1447–1464.

[11] Vinayshekhar Bannihatti Kumar, Abhilasha Ravichander, Peter Story, and Norman Sadeh. 2019. Quantifying the effect of in-domain distributed word representations: A study of privacy policies. In *AAAI Spring Symposium on Privacy-Enhancing Artificial Intelligence and Language Technologies*.

[12] 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.

[13] Jonathan A Obar and Anne Oeldorf-Hirsch. 2020. The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services. *Information, Communication & Society* 23, 1 (2020), 128–147.

[14] Rohan Ramanath, Fei Liu, Norman Sadeh, and Noah A Smith. 2014. Unsupervised alignment of privacy policies using hidden markov models. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 605–610.

[15] Mukund Srinath, Shomir Wilson, and C Lee Giles. 2020. Privacy at Scale: Introducing the PrivaSeer Corpus of Web Privacy Policies. *arXiv preprint arXiv:2004.11131* (2020).

[16] Ali Sunyaev, Tobias Dehling, Patrick L Taylor, and Kenneth D Mandl. 2015. Availability and quality of mobile health app privacy policies. *Journal of the American Medical Informatics Association* 22, e1 (2015), e28–e33.

[17] 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*. 163–166.

[18] Welderufael B Tesfay, Peter Hofmann, Toru Nakamura, Shinsaku Kiyomoto, and Jetzabel Serna. 2018. PrivacyGuide: towards an implementation of the EU GDPR on internet privacy policy evaluation. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*. 15–21.

[19] Pelayo Vallina, Álvaro Feal, Julien Gamba, Narseo Vallina-Rodriguez, and Antonio Fernández Anta. 2019. Tales from the porn: A comprehensive privacy analysis of the web porn ecosystem. In *Proceedings of the Internet Measurement Conference*. 245–258.

[20] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N Cameron Russell, et al. 2016. The creation and analysis of a website privacy policy corpus. In *Annual Meeting of the Association for Computational Linguistics*. 1330–13340.

[21] Razieh Nokhbeh Zaeem, Safa Anya, Alex Issa, Jake Nimergood, Isabelle Rogers, Vinay Shah, Ayush Srivastava, and K. Suzanne Barber. 2020. PrivacyCheck v2: A Tool that Recaps Privacy Policies for You. In *29th ACM International Conference on Information and Knowledge Management (CIKM)*. ACM. To appear.

[22] Razieh Nokhbeh Zaeem and K Suzanne Barber. 2017. A study of web privacy policies across industries. *Journal of Information Privacy and Security* 13, 4 (2017), 169–185.

[23] 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 (TMIS)* 12, 1 (2020), 1–20.

[24] Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. 2018. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Transactions on Internet Technology (TOIT)* 18, 4 (2018), 53.

[25] 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 (2019), 66–86.

[26] Sebastian Zimmeck, Ziqi Wang, Lieyong Zou, Roger Iyengar, Bin Liu, Florian Schaub, Shomir Wilson, Norman M Sadeh, Steven M Bellovin, and Joel R Reidenberg. 2017. Automated Analysis of Privacy Requirements for Mobile Apps.. In *NDSS*.

WWW.IDENTITY.UTEXAS.EDU