Privacy Not Found: A Study of the Availability of Privacy Policies on the Web

Soundarya Nurani Sundareswara
Pennsylvania State University
University Park, PA, USA
sxn5310@psu.edu

Shomir Wilson
Pennsylvania State University
University Park, PA, USA
shomir@psu.edu

Mukund Srinath
Pennsylvania State University
University Park, PA, USA
mus824@psu.edu

C. Lee Giles
Pennsylvania State University
University Park, PA, USA
clg20@psu.edu

Abstract

Legal jurisdictions around the world typically require websites to post privacy policies if they collect information about their users, but some websites that purport to post privacy policies actually do not provide them. We investigate how widely available privacy policies are on a large sample of websites. To find a privacy policy, a user typically visits the landing page of a website (i.e., by entering a URL that chiefly contains the domain as the address) and follows a link from that page to the policy. We automate this exploration on a set of 7 million companies sampled from Free Company Dataset, and we examine the contents of the documents we collected. We identify potential causes for unavailability of privacy policies such as dead links, documents with empty content, egregiously short documents, documents unavailable in certain languages as compared to landing page and documents that consist solely of placeholder text. We estimate the frequencies of these failures and discuss their ramifications.

1 Introduction

Privacy policies are legal documents that service providers use to inform users about practices of data collection, storage and use. Many legal jurisdictions require privacy policies for websites that collect users' information. However, users are often faced with a major challenge of reading privacy policies as they tend to be lengthy and confusing, and they are rarely read and seldom help in decision making. According to a study, an individual would spend about 154 hours per year to even skim through privacy policies of websites they visit throughout the year [4]. Even prior to readability and length, some privacy policies are unavailable: a link on the website purportedly leads to the privacy policy, but the link is broken or leads to a document that does not fulfill the request. This presents a basic obstacle to notice and choice, with legal implications for the website operator.

In this study, we focus on availability of privacy policies on the world wide web. By collecting privacy policy documents from a large set of company domains, we observe different inconsistencies associated with availability of these documents.

As such, we make the following contributions:

- Provide a detailed analysis on various anomalies found related to availability of privacy policies in the data set at each stage of a document collection and classification pipeline.
- Estimate the frequencies of such inconsistencies and overall unavailability of privacy policies.

2 Related Work

Prior research has addressed the problem of presenting privacy policies in a manner that is suitable for reading. For example, Wilson et al., 2016 [9] introduced a corpus of 115 privacy policies that were annotated manually by skilled workers with information about different data collection and use practices. Harkous et al., 2018 [2] built an automated framework for privacy policy analysis and provided users with structured and free form querying. These efforts have significantly improved the comprehensibility of these documents.

One of the prior works on assessing privacy policies involved identifying mismatches between user expectations and privacy practices stated in privacy policy documents [5]. The paper describes the potential of highlighting unexpected practices in websites thereby helping users to make better privacy decision. On the other hand, Sunyaev et al., 2014 assessed the availability, scope and transparency of privacy policies of mobile health apps on iOS and Android [7]. They showed that these privacy policies have poor transparency and availability rates indicating that app developers seem to be constantly competing and have failed to provide app privacy practices. Though our study has a similar objective, we differ in our work by examining privacy policies derived from a wide range of company websites available on the web to determine how often privacy policies have malformed URLs and unrelated or empty content.

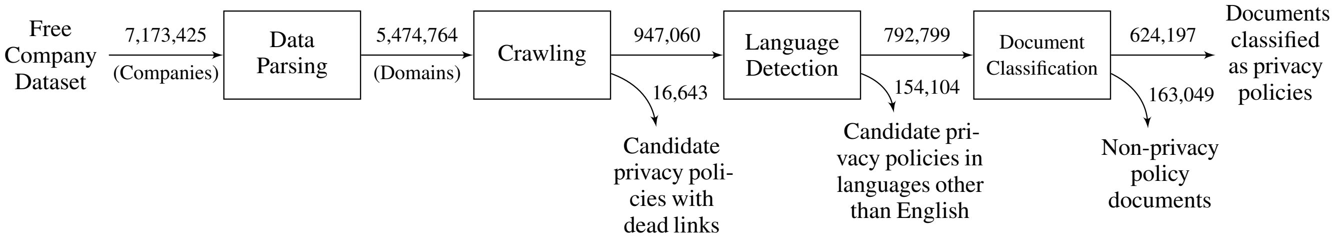
Free Company Dataset
7,173,425
(Companies)
Data Parsing
5,474,764
(Domains)
Crawling
947,060
16,643
Candidate privacy policies with dead links
Language Detection
792,799
154,104
Candidate privacy policies in languages other than English
Document Classification
624,197
163,049
Non-privacy policy documents
Documents classified as privacy policies

Figure 1: Processing pipeline for document collection and classification.

3 Document Collection and Classification

We used a list of company domains from the Free Company Dataset[1] provided by People Data Labs. The dataset is a collection of over 7 million global companies, including fields such as name, domain, year founded, industry, size range, locality, country, LinkedIn URL, current employee estimate, and total employee estimate. The main field of interest for our study is the domain, as it provides the company's website URL. Figure 1 shows our processing pipeline for extracting privacy policies from this dataset. This pipeline was largely inspired from the work of Mukund et al., 2020 to gather privacy policies from the web [6].

We extracted a list of company domain URLs from the dataset. Since the dataset reports duplicates and 35% null values in the domain field, we eliminated such records. In the end, 5,474,764 non-null unique company domain URLs were used for further processing. We used Scrapy[2] to crawl the domain URLs to obtain candidate privacy policies by searching for HTML *href* attributes containing the words "privacy" or "data" and "protection" after case folding the characters in *href*. This technique is used as company home pages usually contain a hyperlink with these keywords that point to a privacy policy document. A total of 947,060 candidate privacy policy URLs were successfully crawled. The remainder either did not have a privacy policy hyperlink on the landing page or had errors from crawling domain URLs and candidate privacy policy URLs. As errors from candidate privacy policy URLs indicate unavailability, we captured these for further analysis.

We detected the language of 947,060 successfully crawled candidate privacy policy documents using LangID [3], a language identification tool that is available as a Python library. It accepts a segment of text and returns the identified language of the text. It is capable of detecting 97 languages across different domains. Figure 2 shows the top ten languages with the highest distribution of documents. 83.7% (792,799) of the documents were in English. Out of the remaining, Dutch was most commonly used. For our analysis, we examined for abnormalities in the non-English candidate privacy policies using a random sample of 500 documents. Additionally, since Latin is not a widely used language in practical legal contexts, we gave the Latin documents further scrutiny, described in the next section.

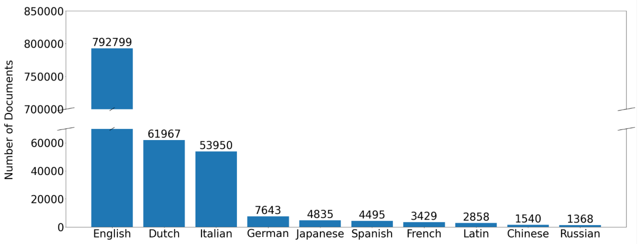

Figure 2: Top ten languages with highest distribution of documents.

We classified the candidate privacy policies in English to separate out documents that were not actually privacy policies. 1,000 randomly selected candidate privacy policies were manually labelled, and we used a supervised machine learning approach where a random forest classifier was trained on the manually labelled documents with features extracted from the privacy policy URLs and words in the document. Features from URLs were obtained by determining TF-IDF for each term in the URL path. Features from the words in the document were extracted after tokenizing the document using a regex tokenizer and removing stop words. The output was the probability of each document being a privacy policy. We thus considered documents with probability less than 0.5 as non-privacy policy and examined them further. 163,049 documents fell under this label.

4 Analysis

In this section, we estimate accuracy of the technique used for crawling to obtain candidate privacy policy documents. We then categorize the unavailability of privacy policies based on the results obtained at different stages of document collection and classification. Further, we provide our observations by examining the content of these documents.

[1] https://docs.peopledatalabs.com/docs/free-company-dataset

[2] https://scrapy.org/

4.1 Crawling Technique Evaluation

During crawling, we extracted candidate privacy policies by searching for particular keywords such as “privacy” or “data” and “protection” in the HTML *href* attributes in the domain web pages. To estimate the accuracy of this technique, we manually examined a random sample of 500 companies excluding those without a domain URL, from the Free Company Dataset. Table 1 lists our observations from this sample. 157 website URLs redirected to an error page. 216 did not disclose a hyperlink to privacy policy document on the landing page whereas 127 company web pages provided the hyperlink. Among these, 86 contained the chosen keywords (“privacy” or “data” and “protection”) as is and 16 with chosen keywords translated in other languages, in the *href* attribute of the hyperlink. In addition, there were 25 company web pages that displayed the privacy policy document with other HTML elements such as a dialog box or with a URL having other keywords. Overall, excluding the company website URLs that redirected to an error page, 37.02% of websites in the sample had a privacy policy hyperlink on the landing page.

To evaluate accuracy, we consider only company websites with a privacy policy hyperlink on the landing page which we manually determine. Since our work focuses on obtaining English language privacy policies as a part of document collection and classification, we exclude the 16 company websites that have translated versions of chosen keywords in privacy policy URL from our calculation. This leaves an accuracy of 77.4% for the technique used for crawling to obtain privacy policies.

Observation		**Number**
Website URLs redirected to an error page		157
Websites without a privacy policy hyperlink on the landing page		216
Websites with a privacy policy hyperlink on the landing page	Having chosen keywords in privacy policy URL	86
	Having translated versions of chosen keywords in privacy policy URL	16
	Having other keywords in privacy policy URL or privacy policy displayed using other HTML elements	25
	Total	**500**

Table 1: Observations on a random sample of 500 company websites from our dataset.

4.2 Dead Links to Candidate Privacy Policies

While crawling for privacy policies, we recorded the URLs that led to error pages. This included both domain and candidate privacy policy URLs. We found 16,643 broken candidate privacy policy URLs that failed due to a variety of technical errors. Out of the different error types, we focus on HTTP, connection refused and value errors as these are indicative of unavailability of privacy policies. 13,261 candidate privacy policy URLs failed with HTTP errors with a majority returning 404 Not Found. There were 95 instances of connection refused error which were observed from URLs referencing localhost. 51 links pointing to candidate privacy policies had value error that comprised of incorrect or misspelled URLs on the website landing page.

Although these errors are associated with privacy policy URLs that are “candidates”, informal experiments suggest a large percentage of them are valid privacy policy URLs and hence these results help in estimating how often privacy policies are unavailable with dead links. On the whole, 1.39% of total privacy policy retrieval attempts contained dead links to candidate privacy policies.

4.3 Natural Language Discrepancies

With the rise in multilingual websites [1], we probed the availability of candidate privacy policies in as many languages as their website landing page offers. Inconsistencies were identified in which the text of a privacy policy was unavailable in certain languages. For example, in one case the website was in English with options to switch to other languages such as Spanish and Portuguese. When the privacy policy in the English version was accessed, we were directed to the document in Spanish. This excludes users non-literate in Spanish, who are unable to read the document without translation services.

To estimate such inconsistencies, we randomly sampled 500 privacy policy URLs from the 154,104 non-English language candidate privacy policies obtained at the end of language detection step. 4.4% (22) of 500 sampled presented inconsistent user experience by not displaying the privacy policy document in a preferred language set on the landing page. Figure 3 shows a heat map indicating frequencies of discrepancy found between languages offered in website landing page and candidate privacy policy page in the random sample. Each unit in the figure represents how often a particular combination occurs across the 22 websites that exhibit inconsistent use of language in landing page and privacy policy page. Another observation was that most discrepancies involved the privacy policy page being provided only in the primary language of the website.

Additionally, we detected 1.85% (2,858) of documents crawled in Latin, a language that is unsuitable for legal notice and choice in any modern jurisdiction. We informally examined the contents of these documents and found most consisted of placeholder texts, such as *Lorem ipsum* [8].

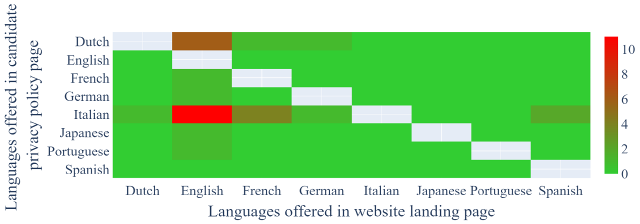

Figure 3: Heat map indicating frequencies of discrepancy between languages offered in website landing page (X axis) and candidate privacy policy page (Y axis) in a random sample of 500 URLs.

4.4 Non-privacy Policy Documents

20.71% (163,049) of candidate privacy policy documents in English were classified as non-privacy policy documents. On analysing their contents, we found them mostly belonging two groups related to unavailability of privacy policy. First, several documents had empty content, meaning that a user would be directed to a successful web page but there would be no content to read. Second, some of these documents were short or provided inappreciable information on privacy practices. However, we do not account them for unavailability of privacy policies as privacy policies may be short depending on the amount of information that company websites collect. To estimate the rate of occurrence of documents belonging to the first group, we took a random sample of 500 URLs of those documents classified as non-privacy policies and manually inspected their content. We found 6.8% (34) of documents in the random sample having empty content. Although this suggests the existence of a page for privacy policy, it does not imply availability as websites that purport to post privacy policy fail to provide content.

5 Discussion

To evaluate the unavailability of privacy policies on the web, we make use of sample proportions obtained in the previous section. In the following calculations, we use a normal model for all the sampling distribution and provide a 95% confidence interval to estimate the total proportion in percentage. For the purpose of this discussion we make two assumptions: Firstly, since a vast majority of websites have only one privacy policy, we assume that each website contains only one hyperlink to the privacy policy document on the landing page. Secondly, we assume that a large percentage of candidate privacy policies obtained are actual privacy policy documents due to the fact that about 80% of candidate privacy policies were classified as actual privacy policies by a random forest classifier with F1 score of 0.97 employed in document classification step. In Section 4.1, we found that 37.02% of websites in a random sample of 500 have a privacy policy hyperlink on the landing page. We can therefore estimate with 95% confidence that 31.8% to 42.2% of websites in the data set have a privacy policy hyperlink on the landing page. Since we extract privacy policies by searching for particular keywords in *href* attributes of websites' landing page during crawling, and with reference to Table 1 on the number of websites with the keywords in privacy policy URL on the landing page, we estimate that 20.4% to 29.6% of websites in the data set would be crawled.

A user's journey to find a website's privacy policy would generally involve accessing the privacy policy hyperlink on the website's landing page. At this point, the user would either encounter a broken link or be led to a successful page. In this data set, **0.28%** to **0.41%** of websites contained dead links to privacy policies and 20.11% to 29.18% of websites navigated to a successful privacy policy page. Once the user lands on a successful web page, the text could be only available in a language that the user is non-literate though the website landing page is provided in different language versions, or there could be no text at all. To estimate the percentage of these inconsistencies, we first calculate the ratio of English to non-English language privacy policies obtained at the end of language detection. The ratio was found to be 5:1 and assuming this remains constant, 3.35% to 4.86% of websites in the data set are estimated to have non-English language privacy policy documents. For the first case of privacy policy unavailability, we estimate the percentage of natural language discrepancies such as privacy policies of multilingual websites being unavailable in certain languages and usage of placeholder texts, to be between **0.15%** and **0.39%**. For the second case of no text in the privacy policy page, we estimate **0.16%** to **0.45%** of websites in the data set to have a privacy policy with empty content.

Adding up the percentages of unavailability of privacy policy due to several reasons listed and considering a set of 10,000 websites each containing a privacy policy hyperlink on the landing page, the range of privacy policies that would be unavailable is estimated to be between **1.62%** and **3.38%**.

6 Conclusion

We studied the unavailability of privacy policies derived from a large dataset of company domains. At every stage of document collection and classification, we encounter a number of failures in obtaining the document and estimate their frequencies. The different failures include broken links, language inconsistencies between the document and its landing page, placeholder texts and empty content in the privacy policy document page. As a result, users would either not have access to the document or a clear understanding of data practices when they choose to use a web service. As future work, we plan to explore different facets of company websites such as industry and total employee estimate provided by the dataset and their correlation to availability of privacy policies.

References

[1] 2010 madrid workshop report. https://www.multilingualweb.eu/documents/madrid-workshop/madrid-workshop-report. [Online; accessed on 28-May-2020].

[2] Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G. Shin, and Karl Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *Proceedings of the 27th USENIX Conference on Security Symposium*, SEC'18, page 531–548, USA, 2018. USENIX Association.

[3] Marco Lui and Timothy Baldwin. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea, July 2012. Association for Computational Linguistics.

[4] Aleecia M McDonald and Lorrie Faith Cranor. The cost of reading privacy policies. *Isjlp*, 4:543, 2008.

[5] Ashwini Rao, Florian Schaub, Norman Sadeh, Alessandro Acquisti, and Ruogu Kang. Expecting the unexpected: Understanding mismatched privacy expectations online. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 77–96, Denver, CO, June 2016. USENIX Association.

[6] Mukund Srinath, Shomir Wilson, and C. Lee Giles. Privacy at scale: Introducing the privaseer corpus of web privacy policies. *arXiv preprint arXiv:2004.11131*, 2020.

[7] Ali Sunyaev, Tobias Dehling, Patrick Taylor, and Kenneth Mandl. Availability and quality of mobile health app privacy policies. *Journal of the American Medical Informatics Association*, page 1–4, 08 2014.

[8] Wikipedia contributors. Lorem ipsum — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Lorem_ipsum&oldid=953944429, 2020. [Online; accessed 12-May-2020].

[9] 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. 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)*, pages 1330–1340, Berlin, Germany, August 2016. Association for Computational Linguistics.