Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text

Vinayshekhar Bannihatti Kumar*[1], Roger Iyengar*[1], Namita Nisal[2], Yuanyuan Feng[1], Hana Habib[1], Peter Story[1], Sushain Cherivirala[1], Margaret Hagan[3], Lorrie Faith Cranor[1], Shomir Wilson[4], Florian Schaub[2], Norman Sadeh[1]

[1]School of Computer Science, Carnegie Mellon University
[2]School of Information, University of Michigan
[3]Institute of Design, Stanford University
[4]College of Information Sciences and Technology, Penn State University
{vbkumar,raiyenga,sadeh}@cs.cmu.edu

ABSTRACT

Website privacy policies sometimes provide users the option to opt-out of certain collections and uses of their personal data. Unfortunately, many privacy policies bury these instructions deep in their text, and few web users have the time or skill necessary to discover them. We describe a method for the automated detection of opt-out choices in privacy policy text and their presentation to users through a web browser extension. We describe the creation of two corpora of opt-out choices, which enable the training of classifiers to identify opt-outs in privacy policies. Our overall approach for extracting and classifying opt-out choices combines heuristics to identify commonly found opt-out hyperlinks with supervised machine learning to automatically identify less conspicuous instances. Our approach achieves a precision of 0.93 and a recall of 0.9. We introduce *Opt-Out Easy*, a web browser extension designed to present available opt-out choices to users as they browse the web. We evaluate the usability of our browser extension with a user study. We also present results of a large-scale analysis of opt-outs found in the text of thousands of the most popular websites.

CCS CONCEPTS

• **Security and Privacy → Human and societal aspects of security and privacy**.

KEYWORDS

Privacy, machine learning, opt-out, privacy policy, text analysis.

ACM Reference Format:
Vinayshekhar Bannihatti Kumar*[1], Roger Iyengar*[1], Namita Nisal[2], Yuanyuan Feng[1], Hana Habib[1], Peter Story[1], Sushain Cherivirala[1], Margaret Hagan[3], Lorrie Faith Cranor[1], Shomir Wilson[4], Florian Schaub[2], Norman Sadeh[1]. 2020. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In *WWW '20: The 2020 Web Conference, April 20–24, 2020, Taipei.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

On the web, *notice and choice* primarily revolve around (1) the use of privacy policies to disclose the data practices associated with a website, and (2) the notion that users can then choose whether to interact with the website and can possibly exercise additional choices offered to them. This framework is widely perceived to be broken [10, 51]. Website privacy policies tend to be lengthy legal documents that users often struggle to understand, or simply do not read [19, 36, 38]. In spite of their cognitive inaccessibility to most web users, privacy policies often contain information about certain choices users have over the collection and use of their personal information. These choices, which we refer to collectively as *opt-outs*, allow a user to exclude themselves from data practices such as tracking by advertising networks, sharing of personal information with third parties, or being contacted by phone or e-mail.

Few users read privacy policies, people are often unaware of the existence of these opt-out choices and, as a result, fail to take advantage of them. A tool that automatically extracts and classifies opt-out choices found in the text of privacy policies could help more people make use of these choices. We present the development of such a tool, from techniques to automatically identify opt-out choices to the design, development, and evaluation of a browser extension that makes these results available to users.

Our research built on the initial observation that the privacy policy text describing opt-out choices often includes hyperlinks [43]. We initially assembled a corpus of 236 website privacy policies and hand-labeled 2,692 hyperlinks from these policies, indicating whether they represented a privacy-related opt-out mechanism. Next, we trained a logistic regression classifier to automatically detect opt-outs in privacy policy text. We also explored the potential of active learning to reduce the quantity of hand-labeled data necessary for this machine learning task. Additionally, detecting opt-outs enabled us to characterize distributions of their properties, such as the data practices that they address. We have released the corpora to the research community for further development.[1]

After establishing the feasibility of detecting opt-outs, we used our system to identify opt-outs in 6,885 privacy policies to support

*The first two authors contributed equally to the paper

[1] Our corpora are available at: https://www.usableprivacy.org/data

WWW '20, April 20–24, 2020, Taipei
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-9999-9/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

a practical level of web coverage. We discuss the distribution of different types of opt-outs across different websites – a website's popularity appears to correlate with the number and types of opt-outs it offers in its privacy policy. We further use our technique to automatically identify opt-outs in the text of privacy policies and design and develop a web browser extension, *Opt-Out Easy*, which presents users with opt-outs for the sites they visit. A small-scale between-subjects user study suggests that the extension makes a difference in helping users identify opt-out choices more quickly and in enabling them to successfully exercise the choices offered by these opt-outs.

2 BACKGROUND & RELATED WORK

Below, we briefly discuss prior work related to this research.

2.1 Regulatory Framework

Europe's General Data Protection Regulation (GDPR) grants consumers several rights pertaining to how companies can use their information. For example, Article 7 allows consumers to revoke consent for the processing of their personal data beyond fulfilling a contractual obligation or business transaction, and Article 21 provides the "right to object" to the use of personal information for direct marketing [17]. Several laws in the United States also mandate certain types of opt-out choices for consumers. At the federal level, the Controlling the Assault of Non-solicited Pornography and Marketing (CAN-SPAM) Act requires companies to provide opt-out choices for commercial and marketing email messages [62]. The California Consumer Privacy Act (CCPA) grants California residents the right to opt-out of the sale of their personal data to third parties, including for marketing purposes [9].

Opt-out choices related to targeted advertising are included in the advertising industry's self-regulatory guidelines developed by the Digital Advertising Alliance (DAA), Network Advertising Initiative (NAI), and Interactive Advertising Bureau Europe (IAB Europe) [15, 26, 44]. DAA members are required to provide consumers an opt-out mechanism for tracking-based targeted advertising [15]. IAB Europe has developed GDPR-specific guidelines for transparency and consent [27]. These industry groups also have developed opt-out tools for their members [16, 45].

2.2 Usability Issues with Opt-Outs

Prior studies have found that consumers often object to the use of their personal information for marketing purposes and desire controls over receiving marketing communications [8, 13]. Similar objections have been found related to web tracking and targeted advertising due to privacy concerns [7, 29, 60, 61]. However, consumers face multiple challenges in addressing these concerns. In a 2010 survey, McDonald and Cranor found that many people were unaware of opt-out tools related to advertising [39]. Yao et al. have found that users continue to have misconceptions and limited technical knowledge about how targeted advertising works [65].

An empirical analysis of privacy choices conducted by Habib et al. found that websites primarily provided choices through the user account settings and the privacy policy. However, the text headings under which choices were placed were inconsistent across websites, which makes finding opt-out choices difficult for consumers [21, 22]. Similarly, Sanchez-Rola et al. found that many websites they analyzed provided misleading information about choices, and that opt-outs for ad tracking were typically difficult to find or ineffective, even after the implementation of GDPR [56]. Furthermore, consumers rarely read privacy policies, which still suffer from poor readability [18]. This has negative implications for how useful current opt-out choices are.

Though broadly adopted, the guidelines and opt-out tools developed by the advertising industry have severe shortcomings. Studies have found that many websites are non-compliant with respective self-regulatory guidelines, particularly with regards to transparency [30]. Hernandez et al. observed that for the Alexa top 500 websites in the United States, fewer than 10% of shown third-party ads displayed the AdChoices icon required by DAA guidelines, and even fewer included the associated text [24]. Users also have been found to have difficulty understanding the scope of these opt-out tools, such as misinterpreting the NAI advertising opt-out tool as an opt-out for all data collection [39]. The limitations of these tools highlight the need for other technologies to enable consumers to effectively exercise their privacy preferences.

Browser extensions that block online trackers have become popular, and have been found to be effective in reducing the number of targeted ads [3]. However, they also suffer from usability issues. Depending on the extension, if users keep the default settings they may not be effectively blocking all web trackers [50]. Furthermore, some extensions use jargon that users do not understand and users may not be provided with appropriate prompts to change the extension settings when a browser extension interferes with the use of a website [33]. Prior work suggests that using these extensions does not lead users to have a better understanding of web tracking [37, 57]. In short, though users desire greater control over online tracking, current mechanisms fail to inspire engagement from users [40, 59]. We leverage the findings from this prior work to inform the design of a new browser extension which removes the burden of locating opt-out processes from users.

2.3 Programmatic Extraction of Opt-Outs

Text classification has been a well-studied problem in the field of Natural Language Processing (NLP). Classical NLP techniques focus on extracting features from text and training models like logistic regression or support vector machines (SVM) [4, 11, 35]. With the advancement of deep learning, prior work in NLP has focused on using word embeddings for text classification [20, 41, 46]. Recently, contextualized word embeddings have shown promise in achieving state of the art results on many natural language understanding problems [14, 47, 63]. We experiment with three of these techniques and compare their performance for opt-out extraction.

NLP techniques have been applied to privacy policies in the past [66, 68]. For example, Wilson et al. [64] created the OPP-115 corpus of annotated privacy policies. Recent work has focused on applying neural models to this dataset [23, 31, 34, 64]. But relatively little work has been done to automatically detect opt-out choices offered in privacy policies. Mysore Sathyendra et al. [43] used logistic regression to detect statements in web privacy policies that described data practices that a user could opt-out of. We extend this work by

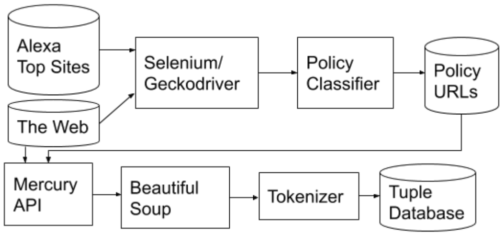


Figure 1: Privacy policy data pipeline.

examining a larger corpus. Furthermore, whereas Mysore Sathyendra et al. [43] only analyzed the text of privacy policies to identify descriptions of opt-out actions, we also utilize the HTML structure of the privacy policy page to segment the policy. This allows us to restrict our problem to detecting hyperlinks that can be used for opting-out of data practices, rather than having to consider all text in a policy. We examined how features derived from policy text, hyperlink URLs, and hyperlink anchor texts can help models determine which hyperlinks are opt-outs. Our results are significantly improved over those reported by Mysore Sathyendra et al. [43], making it possible to build a useful browser extension.

3 DATA PIPELINE

In this section we describe our data pipeline, from scraping webpages to the inputs to our ML classifiers (see Figure 1). We download webpages containing privacy policies using the Mercury Parser API [48], which renders pages and removes sidebars, ads, and other elements that are not part of the page's main content. We then use Beautiful Soup to construct a Document Object Model (DOM) tree of the page's remaining content. We then traverse the DOM tree and extract segments of text from the policy.

3.1 Finding Privacy Policies

We attempted to download privacy policies from the top 500 websites on the U.S Alexa list in the fall of 2018, using the Alexa Top Sites API [2]. Our system downloaded the homepage of each of these websites using Selenium [58] and Geckodriver [42]. Geckodriver renders the webpage which allows us to obtain content that gets loaded dynamically after the initial HTTP request. Our code then assembles a list of linked pages and downloads them. The HTML content of each webpage was classified using logistic regression (LR), to determine if it contained a privacy policy using the classifier by Zimmeck et al. [67]. Afterwards, we manually inspected all pages and removed any without privacy policies that our LR classifier mislabeled. This left us with a list of 236 unique URLs of pages containing privacy policies.

3.2 Extracting Policy Text

Many privacy policy pages include extraneous content such as navigation bars and advertisements. We used the Mercury Parser API to obtain a filtered subset of each policy page. This subset also contains content loaded after the initial HTTP request to a page. We constructed a DOM tree based on the page's content using BeautifulSoup [52] and the lxml parser [5]. Most webpages violate HTML standards [12]. Fortunately, BeautifulSoup is effective on many invalid HTML documents.

```
def recursive_tokenize(dom_subtree):
    for li in dom_subtree:
        Remove li from dom_subtree
        recursive_tokenize(li)

    for p in dom_subtree:
        Remove p from dom_subtree
        recursive_tokenize(p)

    for div in dom_subtree:
        Remove div from dom_subtree
        recursive_tokenize(div)

    nltk_sent_tokenize(dom_subtree.text)
```

Listing 1: Obtaining text segments from the DOM tree.

```
If you wish to opt-out of interest-based
advertising, click <a href="http://preference
s-mgr.truste.com">here</a>
```

Listing 2: Hyperlink with anchor text "here" [55].

The privacy policies we retrieved were not always written in complete sentences. Instead, some of these pages split up lines of text using lists or line breaks without any punctuation. This complicated the process of text segmentation. Simply running NLTK's sentence tokenizer [6] on the raw text BeautifulSoup extracted from the page resulted in malformed segments. Consider a webpage that ends a line with the word "confidence" and then starts the next line with the word "You," without any punctuation in between. BeautifulSoup will extract "confidenceYou" as the raw text, which NLTK will not split up. The same problem would occur with a list, where one item ended with the word "confidence" and the next item began with the word "You." To further complicate matters, some pages nested list, paragraph, and content division elements within each other.

To address this, we inserted a space character at every place there was a line break. We used a recursive function to traverse the DOM tree and split the text into chunks that were then run through NLTK's sentence tokenizer (see Listing 1). We call a token found using this function a *segment*. Note that a complete sentence that does not span multiple list items, paragraphs, or division elements would be a segment. A page that is not written in complete sentences would have at least one segment that does not correspond to a complete sentence.

Many websites contain hyperlinks that use part of the page's text as an anchor. The word "here" is the anchor in the example in Listing 2. We stored the URL and anchor text of hyperlinks that appeared on privacy policy pages. We also kept track of the policy segment in which a hyperlink appeared. Because NLTK's sentence tokenizer only operates over raw text, we had to replace each hyperlink's anchor text with a unique ID in order to match hyperlinks to text segments after the text was tokenized.

3.3 Annotating Data

We now had a *(segment text, hyperlink URL, hyperlink anchor text)* tuple for each hyperlink on every privacy policy page. We observed that 521 of the 3,213 hyperlinks we found linked to only 11 common third-party services. The informational webpage privacyshield.gov accounted for 80 of these 521 common third-party service links. Links to privacyshield.gov are not opt-out links. However, the remaining 441 links were opt-outs. The DAA and NAI opt-out services accounted for 259 of these 441 common third-party opt-out links. We labeled all 80 privacyshield.gov links as not being opt-out links and all 441 links to common opt-out services as being opt-out links. The tuples corresponding to the remaining 2,692 hyperlinks were manually annotated. The classifier performance results reported in Sections 4 and 5 were obtained based only on the 2,692 manually-annotated hyperlinks.

We manually labeled tuples, indicating whether or not they constituted an opt-out hyperlink. This determination was primarily based on how the policy described the link, as well as an examination of the destination page when the policy text was not sufficiently clear to make a labeling decision. All 2,692 tuples were annotated by one annotator, according to a coding manual that had been iteratively developed. A subset of 50 labeled tuples were then randomly sampled and also independently labeled by two additional annotators. Inter-rater reliability was sufficiently high (Fleiss' $\kappa = .70$).

For this task, we built an annotation tool using the Flask micro web framework [53]. This tool ran the webpages that were being annotated through a browser's rendering engine in order to enable the annotators to see hyperlinks in the context of the page and also see the page's text structured with paragraphs and headings. Segments that were repeated verbatim multiple times within a single policy or multiple different policies were filtered out, retaining only one instance of each segment. Some segments contained multiple hyperlinks. We picked a single hyperlink to go along with each segment. We treated the hyperlinks that were not picked as if they were just plain text. This left us with 2,016 tuples in our corpus, 297 of which were opt-outs.

4 IDENTIFYING OPT-OUT HYPERLINKS

We randomly assigned policies, and respectively extracted segments, to either the training, validation or test set. The training set consisted of 1,416 segments, the validation set of 258 segments, and the test set of 339 segments.

Each element contained a tuple *(segment text, hyperlink URL, hyperlink anchor text)*. All three tuple elements for the example in Listing 2 contain information that might help indicate that this segment describes an opt-out choice. We experimented with features extracted from all three tuple elements. These included features derived from segment text in the form of bags of words and bigrams, modal verbs and key phrases, and topic modeling. In addition, we tried bags of words based on the hyperlink URL and anchor text.

We ran experiments using a logistic regression model. We ran an ablation test to assess the importance of individual feature sets. The results are shown in Table 1. We note that there is a significant drop in recall when we remove our bag of words and bigrams feature set. We then trained and evaluated models using only a single feature set (see Table 2). The model that was trained and evaluated using only our bag of words and bigrams feature set performs almost as well as any combination of feature sets that we evaluated during our ablation test. This indicates that the other features do not significantly help with this task.

Table 1: Results of ablation test.

Removed Feature Set	Precision	Recall	F1
None	0.90	0.86	0.88
Words and bigrams	0.91	0.76	0.83
Modal verbs/key phrases	0.86	0.82	0.84
Topics	0.90	0.86	0.88
Hyperlink URL	0.87	0.94	0.91
Hyperlink anchor text	0.88	0.86	0.87

Table 2: Results from models that were trained and evaluated using only a single feature set.

	Validation			Test
Feature Set Used	Precision	Recall	F1	F1
Words and bigrams	0.87	0.88	0.87	0.79
Modal verbs/key phrases	0.58	0.84	0.69	-
Topic Modeling	0.25	0.92	0.40	-
Hyperlink URL	0.78	0.27	0.41	-
Hyperlink anchor text	0.56	0.45	0.5	-
BERT	0.83	0.98	0.9	-
fastText	0.90	0.76	0.82	-

We further trained and evaluated classifiers on our corpus using BERT [14] and fastText [28]. BERT is an encoder of a Transformer [63] model which uses contextualized word embeddings to achieve state of the art results on many NLP tasks. FastText is a library for text classification and word representation. FastText models require less computation than neural networks. BERT and FastText only operate over raw text. We could therefore only train our BERT and FastText models on the segment without the URL, and we could not highlihgt the anchor text to the model. Our evaluation of these classifiers is included in Table 2. The performance of the BERT model is similar to the performance of our classifier that used words and bigrams. The FastText model did not perform as well.

We decided to perform our final test using our logistic regression model that only included features from segments' words and bigrams. We chose this model over BERT because inference is less computationally intensive for LR than neural networks; explaining decisions is easier for LR than neural networks; and the LR model had higher precision than BERT. We present our results from testing this model on the test set in Table 2. The model performance on the test and validation sets are similar, indicating that this classifier will likely have similar performance on new data.

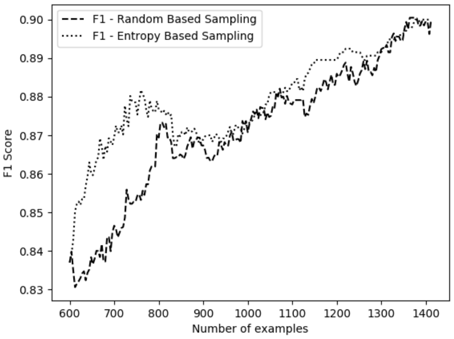


Figure 2: Comparisons of classifiers trained on data sampled randomly and data sampled based on Entropy.

4.1 Exploring Active Learning

Active learning is a semi-supervised machine learning approach in which annotators will label items that models have the highest uncertainty about. We wanted to see if active learning would reduce the number of tuples that needed to be labeled in order to build an opt-out detection classifier. First, we ran a baseline experiment in which we started with a seed of 600 tuples that were randomly selected from our training set. Next, we trained a logistic regression classifier with this seed and evaluated it on our validation set. Then we expanded the size of our sample by randomly selecting 4 of the remaining 816 tuples in our training set. Afterwards, we trained a new classifier with the 604 tuples in our sample. This process was repeated, randomly selecting 4 more training tuples to add to our sample each time. Adding a tuple to the sample represents labeling an additional piece of data and then adding it to the training set.

We then ran a similar experiment where we expanded our samples based on entropy, rather than selecting elements randomly [25]. Entropy is high when a classifier is uncertain about a prediction. Therefore, adding the tuples with the highest entropy to the training set may increase the classifier's performance more than adding tuples with lower entropy.

We repeated the experiment from the baseline, except we selected elements with the highest entropy, rather than selecting randomly. Entropy was computed using the formula:

$$H = -P_{\text{positive}} \log_2(P_{\text{positive}}) - P_{\text{negative}} \log_2(P_{\text{negative}})$$

Figure 2 shows the results of these experiments. Most classifiers trained on samples selected based on entropy performed better than classifiers trained on a sample of the same size that was selected randomly. We believe that selecting data to label based on entropy is an effective way to collect training data in this domain.

Table 3: Breakdown of corpus by category annotation.

Category	Train	Val	Test
Targeted Advertising (AD)	185	76	133
Communication (CM)	139	61	81
Cookies (CK)	90	45	44
Analytics (AN)	45	28	38
Sharing with third parties (SH)	50	29	33
Others	49	29	79

Table 4: Results from Category Classification.

	Logistic Regression		BERT
Category	**Val F1**	**Test F1**	**Val F1**
Targeted Advertising	0.75	0.79	0.73
Communication	0.83	0.85	0.86
Cookies	0.74	0.70	0.75
Analytics	0.75	0.62	0.68
Sharing with third parties	0.62	0.63	0.64
Others	0.55	0.51	0.62

5 CATEGORIZING OPT-OUT HYPERLINKS

In addition to detecting opt-out hyperlinks, we wanted to determine the types of data practices that these opt-out choices involve. The opt-out detector that we describe in Section 4 was used to help with this. We first annotated the 297 opt-out tuples in our first corpus with up to two categories of data practices that the opt-out involves. Some of our training examples had 2 categories of opt-outs. These categories are shown in Table 3. We then downloaded and filtered 388 additional policies from the Alexa top-2,000 U.S. websites. We ran these 388 policies through our opt-out detector. This provided us with 751 additional opt-out hyperlinks that we also annotated with category labels. Table 3 provides a breakdown of this corpus. If we had labeled all hyperlinks in these policies, we would have had to label 6.5 times as many hyperlinks to get the same number of tuples containing opt-out links. We acknowledge that in the process we likely missed some opt-out hyperlinks, as the performance of our overall classifier is not perfect.

We built a logistic regression classifier to automatically determine the categories of opt-outs. Features were generated by a TF-IDF vectorizer that incorporated words, bigrams, and trigrams. In addition, we built a classifier using BERT. The results of these two classifiers are presented in Table 4. Their performance is roughly similar, with F1 values typically ranging between 0.70 and 0.85, and lower values for third party sharing opt-outs. Since both the BERT model and the logistic regression model performed equally well, we chose the logistic regression model for our test set evaluation as it was faster at evaluating the classes compared to BERT. The performance of these classifiers would likely improve if one had access to a larger corpus of annotated opt-outs. It is worth remembering however that these results are for opt-out links that do not correspond to the set of 11 easily-identifiable third party services used by many sites to implement opt-out choices. When crafting simple rules to automatically detect these opt-outs and combining these rules with our classifiers, we are actually able to achieve an overall recall of 0.90 and a precision of 0.93. In our annotated corpus, the 11 easily-identifiable third party opt-out services accounted for 441 of 3,251 hyperlinks, which represents 14% of the hyperlinks. Accordingly, in determining the performance of our hybrid approach, which combines the detection of these 11 easily-identifiable opt-outs with our machine learning techniques, we considered a test set with 17%

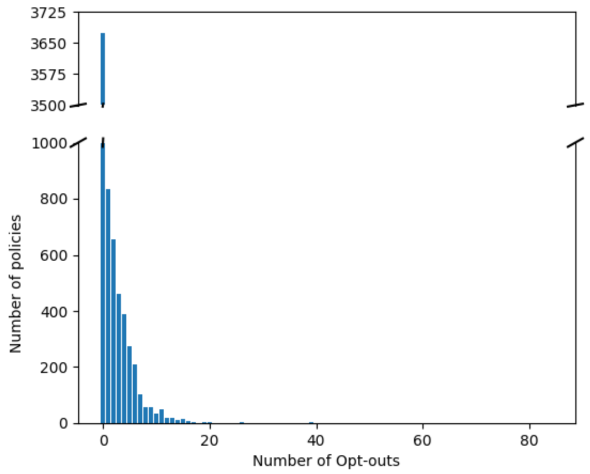


Figure 3: Number of policies vs number of opt-outs.

of the annotated data and added 74 (17% of 440) instances of the 11 easily-identifiable opt-outs, resulting in an overall precision of 0.93, a recall of 0.9 and an F1 score of 0.91.

In the remainder of this paper, we build on this hybrid approach to analyze the presence of opt-out links on several thousand top-ranked websites. We also use this hybrid approach to build and evaluate a browser extension that automatically extracts opt-out links from the text of privacy policies and presents them to users.

6 ANALYSIS OF OPT-OUT CHOICES

In this section, we use our approach to automatically analyze opt-outs disclosed in the 6,885 privacy policies displayed on The Usable Privacy Policy Explore Website.[2] Some websites linked to multiple privacy policies. At the same time, we intentionally skipped 23 websites with adult content and a small number of otherwise problematic websites (e.g., sites that created difficulties for our parser or segmenter). We segmented these policies as described in Section 3.2 and ran our hybrid approach to identify opt-outs. Below we discuss some of our findings.

Many privacy policies do not seem to have opt-outs. We observed that, at a high level, most of the analyzed privacy policies had none or at most one opt-out hyperlink, as shown in Figure 3. We proceeded to conduct a finer analysis, looking at potential correlation between the number of opt-outs found in a privacy policy and the popularity (Alexa rank) of the corresponding website.

Number of opt-outs per website based on website's Alexa rank. Given that some websites have multiple privacy policies, the results presented report the total average number of opt-outs identified across different websites (Columns 3) in Tables 5 and 6. We then find the mean number of opt-outs per site(Column 4). We find that the average number of opt-outs varies with the website's Alexa rank. This is true both when looking solely at U.S. websites (see Table 5) and also when looking at websites based on global ranking (see Table 6). Specifically, more popular websites (namely sites with low Alexa ranks) seem to offer their users more opt-outs than less popular ones (namely sites with a higher Alexa rank). This is true both when looking at U.S. rankings and worldwide rankings. It should be noted that these results are based on the analysis of these websites' privacy policies. It is always possible that some sites do not disclose all their opt-outs in their privacy policies. This being said, intuitively one would expect more popular websites to generally be more sophisticated (e.g., more complex workflows, more sophisticated privacy personnel, etc.). This in turn seems to translate into these sites also offering a greater number of opt-outs to their users.

Table 5: When looking at U.S. rankings, more popular sites also offer more opt-outs to their users

US Alexa Rank	# Policies	# Opt-Outs (normalized)	Ratio
1-200	194	669.00	3.43
200-1000	702	1,751.45	2.49
>1000	7,848	9,639.53	1.22

Table 6: When looking at worldwide rankings, more popular sites offer more opt-outs to their users.

Global Alexa Rank	# Policies	# Opt-Outs (normalized)	Ratio
1-200	121	342.2	2.82
200-1000	418	1,016.1	2.43
>1000	8,213	10,707.9	1.30

Table 7: We observe a difference in the kinds of opt-outs mentioned based on a website's Alexa ranking.

Global Alexa Rank	AD%	CM%	CK%	AN%	SH%
1-200	69.20	11.25	11.66	0.80	7.08
200-1000	56.74	10.08	19.16	7.10	6.80
>1000	54.04	10.06	21.04	8.80	5.90
Mean % of opt-outs	60.00	10.46	17.28	5.56	6.59

Distribution of Opt-Outs By Category and Website Rank. Table 7 breaks down identified opt-outs by popularity of websites and also by categories of opt-outs. As can be seen advertising opt-outs (*AD*) account overall for 60% of all detected opt-outs, following by 17% of cookie opt-outs (*CK*), 10% communication opt-outs (*CM*), about 7% third-party sharing opt-outs (*SH*), and about 6% analytics opt-outs (*AN*). The more popular websites seem to also have a greater percentage of advertising opt-outs than the less popular sites and their percentage of analytics opt-outs also seems to be significantly lower than the corresponding percentages on less popular sites. We acknowledge that these measurements are limited by the presence of opt-out hyperlinks on the privacy policies of websites.

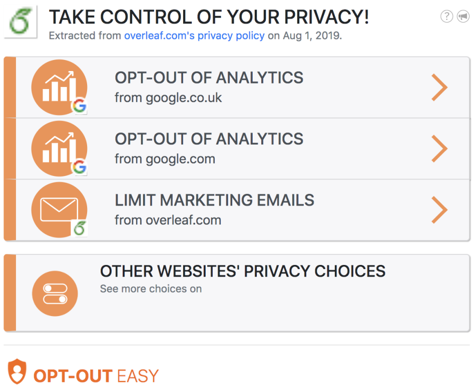
[2] https://explore.usableprivacy.org

7 BROWSER EXTENSION: OPT-OUT EASY

Building on our approach for automatically extracting and classifying websites' opt-out hyperlinks, we developed a browser extension, called Opt-Out Easy, to make it easier for users to find and access opt-outs in privacy policies. By clicking on the extension's icon, a user is presented with categorized opt-out links identified in the text of the website's privacy policy. The extension also helps users keep track of which opt-outs they have already interacted with.

We attempted to download privacy policies from the Alexa top-7,000 U.S. websites. Our pipeline described in Section 3 was used, except we did not manually verify that all policy URLs corresponded to policies. All extracted tuples were fed into the classifier described in Section 4 to determine which corresponded to opt-out choices. Next, the tuples corresponding to opt-out choices were run through the classifier described in Section 5 to determine the type(s) of data practices the choice involves. These results were stored in a MySQL database and later served to the browser extension through an API built with Django.

When the user clicks the extension's icon, the extension makes a request to our API server. The server responds with the opt-out hyperlinks for the current website, if it has already scanned the website's privacy policy for hyperlinks, otherwise the user has the option to request that the site be analyzed later. Because it would take up to a minute or two to perform the analysis in real-time and also because of cost issues, this seemed to be a reasonable compromise, as it provides for some level of user engagement even when the extension does not have results it can readily show to the user. User requests are later processed in a batch job, with results available for users who visit those sites later on. Our server only stores anonymized logs of the websites for which opt-out links have been requested. To protect users' privacy, these logs are dissociated from specific users and we make no other attempts to identify users.



TAKE CONTROL OF YOUR PRIVACY!
Extracted from overleaf.com's privacy policy on Aug 1, 2019.

OPT-OUT OF ANALYTICS
from google.co.uk

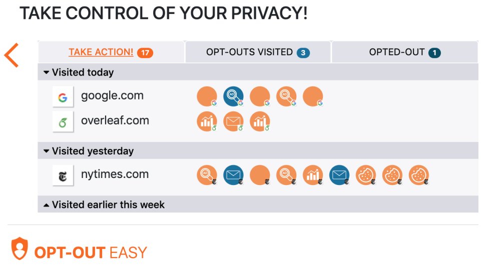
OPT-OUT OF ANALYTICS
from google.com

LIMIT MARKETING EMAILS
from overleaf.com

OTHER WEBSITES' PRIVACY CHOICES
See more choices on

OPT-OUT EASY

Figure 4: Opt-Out Easy's results for Overleaf.com.



TAKE CONTROL OF YOUR PRIVACY!

TAKE ACTION! 17 | OPT-OUTS VISITED 3 | OPTED-OUT 1

▾ Visited today
google.com
overleaf.com

▾ Visited yesterday
nytimes.com

▴ Visited earlier this week

OPT-OUT EASY

Figure 5: Opt-Out Easy's summary of opt-out hyperlinks across recently visited websites.

7.1 Browser Extension Design

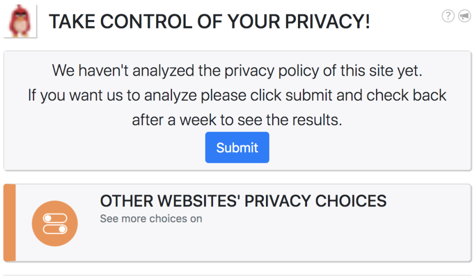
We describe the main design aspect and features of the Opt-Out Easy browser extension.

7.1.1 Opt-out Screen. The main screen users see when they click the extension's icon is the opt-out screen. It shows all opt-out choices identified in the privacy policy of the website the user is currently visiting. For a given opt-out hyperlink, the browser extension shows an icon and heading, which inform the user about the type of opt-out (e.g., targeted advertising, communication, cookies, analytics, or sharing). A favicon shown at the bottom right of the icon and additional text communicating whether the opt-out is being offered by the first party (the current website) or a third party. This helps users understand the kind and scope of the opt-out. Figure 4 shows the opt-out results after scanning the Overleaf web page.

Opt-out links that a user has already visited are shown in blue, while the links which the user has not yet visited are shown in orange. Because users are likely to forget whether or not they have already visited some opt-out choices, the feature helps them remember and saves them the trouble of revisiting opt-out choices with which they have already interacted. To further help users keep track of the actions they have taken with specific opt-outs, the extension also offers users the ability to record their action via a "tell us what you did" link. If the user decides to use this link, the extension can also remind them about the actions they have taken when they return to the website.

7.1.2 Summary of opt-outs for recently visited websites. The browser extension provides a second screen to help users keep track of opt-outs for pages they have recently visited. This screen consists of three tabs: "Take Action" shows opt-out choices for which the user hasn't yet taken any action. "Opt-Outs Visited" lists opt-out links that the user has already visited but not opted out of. "Opted-Out" lists the opt-out links that the user has visited and indicated they opted out of. These views are meant to encourage the user to take action on websites they recently visited and also help them quickly glance at all the privacy choices they have already made.

7.1.3 Information page. The extension also includes an information page (see Figure 7) that explains to users how the browser extension and the opt-out hyperlink analysis works. Clearly communicating the underlying functionality helps users understand what the extension does, helps build trust in the technology, and may also help users understand the extension's limitations (e.g., the extension could miss some opt-out links and does not show links not disclosed in the privacy policy). The extension itself is designed to be privacy friendly: it does not record any identifiable



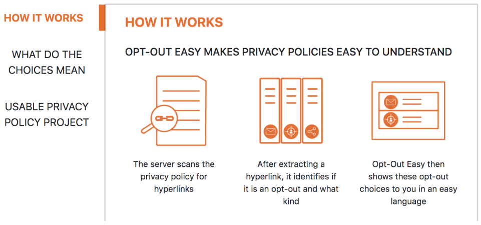
TAKE CONTROL OF YOUR PRIVACY!

We haven't analyzed the privacy policy of this site yet.
If you want us to analyze please click submit and check back after a week to see the results.

Submit

OTHER WEBSITES' PRIVACY CHOICES
See more choices on

Figure 6: Opt-Out Easy allows users to request that we scan the privacy policy of any website they want.



HOW IT WORKS

WHAT DO THE CHOICES MEAN

USABLE PRIVACY POLICY PROJECT

HOW IT WORKS

OPT-OUT EASY MAKES PRIVACY POLICIES EASY TO UNDERSTAND

The server scans the privacy policy for hyperlinks

After extracting a hyperlink, it identifies if it is an opt-out and what kind

Opt-Out Easy then shows these opt-out choices to you in an easy language

Figure 7: Opt-Out Easy's information page for users to understand how the tool works.

information about the user on the server side. We only record information about the users on the local client on which the tool has been installed.

7.1.4 Request page. Our system is currently set up to analyze privacy policies of most of the top 7,000 Alexa U.S. websites in batches. We plan to run the system once per month initially. If users want to see results for websites not included in our monthly analysis, they can use the browser extension's online request form, as shown in Figure 6. We are planning to process such requests within a week and add them to the collection of sites we analyze monthly. Over time, depending on available resources and popularity of the extension, we may increase the frequency of our analyses.

7.2 Initial Usability Evaluation

We conducted an initial usability evaluation of Opt-Out Easy to determine to what extent the extension helps users identify opt-outs, looking at effectiveness, efficiency, and overall user satisfaction.

7.2.1 Study procedures and participants. Our study employed a between-subjects design. Participants in the treatment group and the control group were asked to complete the same set of opt-out tasks with or without Opt-Out Easy, respectively. This between-subject experiment aimed to evaluate to what extent the extension helps users opt-out of data practices faster and more successfully. Follow-up interviews with all participants provide additional qualitative data to evaluate the usability of the extension.

We followed an Institutional Review Board-approved study protocol, which we detail below. We used social media posts and physical flyers to recruit potential participants to fill out a screening survey. Then, we invited eligible participants to our university campus to participate in the study. After obtaining informed consent, we first explained "opt-out" and "data practices" in layman's terms to all participants with print-out screenshots of opt-out choices. For the treatment group, we provided additional screenshots of the extension and showed participants where to access this extension in the Chrome browser. These explanations ensured that all participants had a basic understanding of the concepts and the functionality needed to work on the tasks we would assign them.

We created a list of 5 opt-out tasks on 4 major websites, covering most opt-out categories supported by the extension (see Table 9), namely advertising and email communication opt-outs. Participants used a lab computer with study accounts to complete these tasks. The accounts were preset to the same privacy settings to ensure study consistency. When describing each task to participants, we used scenario prompts without mentioning the specific word "opt-out" to minimize potential framing. For example, for the New York Times' website, we described the task as: "You just got the 10th update email from New York Times today. Now you want to stop receiving them." We recorded the computer screen when participants completed these tasks for analysis.

In the post-experiment interview, we asked participants about their (1) perceived ease of performing the tasks, (2) familiarity with the 4 websites used in the experiment, (3) previous opt-out experience on the web, and (4) intention to opt-out of data practices in the future. For the treatment group, we asked them to rate 6 usability statements about Opt-Out Easy (see Table 8) and their subjective opinions about using the extension. For the control group, we then described Opt-Out Easy to them with screenshots and asked them if they would like to use it when trying to opt-out in the future. All interviews were audio recorded and transcribed by the research team for qualitative analysis.

We recruited 8 participants for this pilot study. 4 participants were female (2 in each group), 7 had college degrees (3 in treatment group), 6 self-reported as being tech savvy (3 in treatment group). After completing all study procedures, each participant received a $15 gift card for their time.

7.2.2 Study results. To measure the effectiveness of the extension in helping users opt out of data practices, we analyzed the screen recordings to calculate if participants successfully completed each task and the time they took to do so. Note that task 5 on GAP's website contained a number of third-party advertising opt-outs and most participants struggled with it as some of these links were broken. Due to these issues in both groups, we excluded task 5 from this analysis. Also, we consider a task failed if the participant spent more than 60 seconds on it because users are unlikely to spend that much time to opt out in real life. The treatment group had an average success rate of 87.5%, while the control group's average success rate was 56.25%. Similarly, participants in the treatment group tended to opt out faster on most of the tasks, as shown in Table 9. These data show initial evidence that Opt-Out Easy is effective in helping users opt-out.

For the 6 usability statements about Opt-Out Easy, participants in the treatment group rated all statements with either positive or neutral ratings (ratings >=0). The average ratings are shown in Table 8. Their perceived future use of the extension and the

Table 8: Treatment group's rating on usability statements for Opt-Out Easy

Statements	Average Rating
This browser plugin is easy to use.	1.00
I would like to use this browser plugin in future.	0.75
The text in this browser plugin is easy to understand.	0.75
The various types of opt-outs provided by this browser plugin are useful.	2.00
I need no additional technical support to be able to use this browser plugin.	1.25
I would imagine that most people would learn to use this browser plugin quickly.	1.00

-2:Strongly disagree, -1:Slightly Disagree, 0:Neutral, 1:Slightly agree, 2:Strongly agree

easiness to understand the text in the extension received slightly lower ratings, showing space for usability improvement.

For the interview questions asked to all participants, all participants in the treatment group reported at least 4 out of 5 tasks were easy when using the extension, while most participants in the control group considered these tasks moderate. 4 participants (3 in the treatment group) had opted out of data practices on websites before, and most participants reported that they were familiar with either Google, Amazon, or both websites. These two factors may have affected their reported ease of carrying out the assigned tasks. For example, 2 participants in the control group reported that their prior opt-out experience and/or familiarity with Google made task 1 easier for them. However, 2 participants in the treatment group felt that their familiarity with these websites did not influence their reported ease of carrying out the assigned tasks. For instance, one of these two participants said "[It did] not help the most because I was using the extension." This suggests that the extension could better assist users with opting out on unfamiliar websites.

For the group-specific interview questions, we conducted a basic thematic analysis on all interview transcripts and summarized three significant themes from the qualitative data. First, *all participants in the treatment group liked some aspects of the extension*, such as the way it centralizes all opt-out choices (e.g., "I can just do it through [the] tool rather than having to hunt down the privacy practices for everything"), the minimal user effort (e.g.,"It was just like a one click kind of thing"), and the detailed information about each opt-out choice (e.g., "It broke down exactly what the tracking was").

Second, *participants in the control group showed certain frustration with the scope of data practices they encounter on the web and the level of effort needed to opt out of these practices.* One participant in the control group who stated he had not opted opt out on websites before said "I have left my data pretty vulnerable in the world, so far. Maybe it [the study] is a bit of a wake up call." Another control participant found Tasks 4 and 5 more difficult, saying "The New York Times had too many different things to click and I don't know what they were...And then the GAP one, it was buried in the privacy policy. It wasn't in account settings."

Third, *participants in both groups saw the value in a tool that streamlines the opt-out process on the web.* For the treatment group, 3 participants indicated they were likely and 1 participant stated (s)he would definitely download the extension for their own use. All of them would recommend the extension to others if there was a need, as one commented "Maybe not [recommend it] to my friends, but probably to my mom or family member who doesn't understand how to opt out." For the control group, all participants said they would like to use such a tool. Specifically, one participant in the control group initially said she would not opt out in the future but changed her mind after we described the tool, saying "That would change my previous answer to 'yes'. Rather than taking 1-3 minutes to do, if it took me 10-15 seconds, I would use it."

Table 9: The mean *time* and *success rate* of each opt-out task in the experiment by group (*n*=8). Task refers to the type of opt-out task involved with "Ad" referring to opting out of advertising and "Email" referring to opting out of email communication.

		Time (s)		Success rate	
Website	Task	Control	Treat	Control	Treat
Google	Ad.	85.50	**46.25**	0.50	**0.75**
Amazon	Ad.	142.50	**20.00**	0.50	**1.00**
Amazon	Email	**48.00**	48.50	0.75	**1.00**
NY Times	Email	104.25	**68.75**	0.50	**0.75**
GAP	Ad.	N/A	N/A	N/A	N/A

8 DISCUSSION

Because opt-out choices are often buried deep in the text of privacy policies, few people know about these choices, let alone exercise them. Overall our work shows that it is possible to (1) develop technology that can automatically identify a large percentage of opt-out choices found in the text of privacy policies and (2) develop effective user interfaces, such as the browser extension piloted in our study, to present users with available opt-out choices and enable them to more effectively make use of these choices. Below we further discuss some of the more detailed findings of our demographic study of opt-out hyperlinks and of our human subject study, including public policy considerations.

8.1 Demographics of Opt-out Choices

Results presented in Section 6 show that the number of opt-out choices found in privacy policies is relatively small. On average, websites that are not among the 1,000 most popular websites (Alexa rank over 1,000) often have just one opt-out per policy. More popular websites have more opt-outs on average. This is partly a result of these sites' complexity. Policies for sites like Amazon or Google cover multiple web properties and support very diverse data flows.

These sites are also scrutinized more, and the organizations that run them have the resources to hire privacy professionals. A more extensive study could also look at how sectoral regulations correlate with the presence of opt-out links. For instance, U.S. financial organizations are required by the Graham Leach Bliley Act to have opt-out notices [1]. Future work might also examine the jurisdictions under which different sites operate and to what extent different jurisdictions yield differences in the average number and types of opt-outs found on different sites. One benefit of the automatic classification approach presented in this paper is that it actually enables people to ask these questions and to more systematically analyse opt-out demographics within and across different categories of websites (e.g. based on popularity, based on sector, based on country where the site is hosted, and more). We hope that moving forward this type of analysis will be used to inform public policy debates. In particular, with the advent of the California Consumer Privacy Act, which requires the introduction of an opt-out for the sale of one's data, it will be interesting to extend the approach presented here and to conduct systematic studies looking at the presence of opt-out hyperlinks focused specifically on this requirement (e.g., what percentage of sites are in compliance, how compliance varies with the popularity of sites, by sector, etc.).

8.2 What Can We Learn from Our User Study?

While small, the pilot study of our Opt-Out Easy browser extension suggests that users are often unaware of available opt-out choices and lack the necessary functionality to discover and exercise these choices. Our study seems to indicate that Opt-Out Easy helps increase awareness of available opt-out choices, while also reducing the time it takes to identify opt-out hyperlinks and eventually take advantage of these choices. While a larger scale evaluation of our browser extension is needed to confirm these early findings, results of out study are encouraging. However, our study also shows that our tool only solves part of the problem experienced by users who decide to opt-out. In fact, our study, as well as prior work by Habib et al. [22], shows that it is not uncommon for opt-out hyperlinks to be broken or for the time required to take advantage of one of these links to be unreasonable. In our study, we observed the following problems at the NAI and DAA opt-out services:

(1) When users connect to these services to opt-out, they are presented with (often long) lists of trackers present on the website and have to select which tracker they want to opt-out from. Often a number of these trackers are shown as "temporarily unavailable," which would require the user to come back multiple times to complete their opt-out requests.
(2) The opt-out process tends to be painfully slow, with users complaining about the "slow progress bar" and often just giving up before the process is complete.

While our browser extension and our automated opt-out identification process cannot solve these problems, they could possibly help. Specifically, one could systematically scan websites for opt-out choices and request crowdworkers to attempt to opt-out, recording whether they succeed and how much time they need. By systematically collecting such statistics, one could help build pressure on the entities running these services. The resulting statistics could also help inform policy makers and motivate them to require minimum standards for availability and response time.

8.3 Limitations and Future work

Our corpus only includes policies for websites at the top of the Alexa list for the United States. Our classifiers thus only work on policies written in English. Future studies should examine privacy policies for non-U.S. sites and lower-ranked sites. Our corpus only contains opt-out links that use anchor tags. Non-anchor tags with Javascript event handlers that redirect users were ignored. Our classifiers for determining whether a webpage contained a privacy policy and whether a hyperlink was an opt-out had non-zero false-negative rates. Our small corpus size likely hurt our precision and recall. Future work could improve performance with additional feature engineering or training a BERT model from scratch on a large corpus of privacy policies, thereby creating a privacy policy-BERT, analogous to "Bio-BERT" [32].

Finally, we acknowledge the small sample size of the pilot study of our Opt-Out Easy extension. While we were able to mitigate this with in-depth qualitative data through post-experiment interviews, we plan to confirm our results by running a larger study.

9 CONCLUDING REMARKS

A central tenet of privacy in the U.S. revolves around the concept of "notice and choice." Unfortunately many choices, which generally come in the form of "opt-outs" are buried deep in the text of privacy policies that few people ever bother to read. The research presented in this paper shows that it is possible to develop techniques that automatically identify opt-out choices in the text of policies. We use this technology to study the demographics of opt-out choices on a corpus of 6,885 popular websites and to also develop a browser extension that automatically displays available opt-outs to users as they browse the web. Results of this research open the door to the more systematic analysis of opt-out demographics on websites and to the development of tools that empower users to effectively take advantage of available opt-outs. At the same time, our study also shows that, even when websites offer opt-outs, these hyperlinks are not always working and using them may also take more time than users have available.

10 ACKNOWLEDGMENTS

This research was mainly conducted as part of the NSF SaTC Frontier project on Usable Privacy Policies (CNS-133059) [49, 54] with additional funding provided by NSF grants CNS-1801316 and CNS-1914486, and a DARPA Brandeis grant on Personalized Privacy Assistants (FA8750- 15-2-0277). Further support was provided under a National Science Foundation's Graduate Research Fellowship Program under Grant Nos. DGE1252522 and DGE1745016. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright notation. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the NSF, DARPA, or the U.S. Government.

REFERENCES

[1] Aigbe Akhigbe and Ann Marie Whyte. 2004. The Gramm-Leach-Bliley Act of 1999: Risk implications for the financial services industry. *Journal of Financial Research* 27, 3 (2004), 435–446.

[2] Amazon Web Services, Inc. 2017. Alexa Top Sites. https://docs.aws.amazon.com/AlexaTopSites/latest/index.html. (2017).

[3] Rebecca Balebako, Pedro Leon, Richard Shay, Blase Ur, Yang Wang, and Lorrie Faith Cranor. 2012. Measuring the Effectiveness of Privacy Tools for Limiting Behavioral Advertising. In *Proceedings of the Web 2.0 Security and Privacy Workshop (W2SP)*.

[4] Eric Baucom, Azade Sanjari, Xiaozhong Liu, and Miao Chen. 2013. Mirroring the real world in social media: twitter, geolocation, and sentiment analysis. In *Proceedings of the 2013 international workshop on Mining unstructured big data using natural language processing*. ACM, 61–68.

[5] S. Behnel. 2005. lxml - XML and HTML with Python. https://lxml.de/. (2005).

[6] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python* (1st ed.). O'Reilly Media, Inc.

[7] Alexander Bleier and Maik Eisenbeiss. 2015. The Importance of Trust for Personalized Online Advertising. *Journal of Retailing* 91, 3 (2015), 390–409.

[8] Bloomberg Businessweek. 2000. Business Week/Harris Poll: A Growing Threat. (2000), 96.

[9] California State Legislature Website. 2018. SB-1121 California Consumer Privacy Act of 2018. (2018). https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201720180SB1121.

[10] Fred H Cate. 2010. The limits of notice and choice. *IEEE Security & Privacy* 8, 2 (2010), 59–62.

[11] Eugene Charniak and Mark Johnson. 2005. Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. In *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 173–180.

[12] Shan Chen, Dan Hong, and Vincent Shen. 2005. An Experimental Study on Validation Problems with Existing HTML Webpages. 373–379.

[13] Lorrie Faith Cranor, Joseph Reagle, and Mark S Ackerman. 1999. *Beyond Concern: Understanding Net Users' Attitudes About Online Privacy*. Technical Report. TR 99.4.1, AT&T Labs-Research.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). arXiv:1810.04805 http://arxiv.org/abs/1810.04805

[15] Digital Advertising Alliance. 2009. Self-Regulatory Principles for Online Behavioral Advertising. (July 2009). http://digitaladvertisingalliance.org/principles.

[16] Digital Advertising Alliance. 2019. Your AdChoices. (2019). https://youradchoices.com/.

[17] European Commission. 2016. EGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). (2016). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[18] 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)*. 18–25.

[19] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 2016. How Short is Too Short? Implications of Length and Framing on the Effectiveness of Privacy Notices. In *Proceedings of the Twelfth USENIX Conference on Usable Privacy and Security (SOUPS '16)*. USENIX Association, USA, 321–340.

[20] Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* (2014).

[21] Hana Habib, Sarah Pearman, Jiamin Wang, Yixin Zou, Alessandro Acquisti, Lorrie Faith Cranor, Norman Sadeh, and Florian Schaub. 2020. "It's a scavenger hunt": Usability of Websites' Opt-Out and Data Deletion Choices. In *CHI'20: ACM CHI Conference on Human Factors in Computing Systems*.

[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*.

[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. *arXiv preprint arXiv:1802.02561* (2018).

[24] Jovanni Hernandez, Akshay Jagadeesh, and Jonathan Mayer. 2011. Tracking the Trackers: The AdChoices Icon. (2011). http://cyberlaw.stanford.edu/blog/2011/08/tracking-trackers-adchoices-icon.

[25] Alex Holub, Pietro Perona, and Michael C Burl. 2008. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 1–8.

[26] IAB Europe. 2011. EU Framework for Online Behavioural Advertising. (April 2011). https://www.edaa.eu/wp-content/uploads/2012/10/2013-11-11-IAB-Europe-OBA-Framework_.pdf.

[27] IAB Europe. 2019. GDPR Transparency and Consent Framework. (2019). https://iabtechlab.com/standards/gdpr-transparency-and-consent-framework/.

[28] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, 427–431.

[29] Hyejin Kim and Jisu Huh. 2017. Perceived Relevance and Privacy Concern Regarding Online Behavioral Advertising (OBA) and Their Role in Consumer Responses. *Journal of Current Issues & Research in Advertising* 38, 1 (2017), 92–105.

[30] Saranga Komanduri, Richard Shay, Greg Norcie, and Blase Ur. 2011. AdChoices? Compliance with Online Behavioral Advertising Notice and Choice Requirements. *A Journal of Law and Policy for the Information Society* 7 (2011).

[31] 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*.

[32] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746* (2019).

[33] Pedro Leon, Blase Ur, Richard Shay, Yang Wang, Rebecca Balebako, and Lorrie Cranor. 2012. Why Johnny can't opt out: a usability evaluation of tools to limit online behavioral advertising. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 589–598.

[34] Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. 2018. Towards Automatic Classification of Privacy Policy Text. (2018).

[35] Larry M Manevitz and Malik Yousef. 2001. One-class SVMs for document classification. *Journal of machine Learning research* 2, Dec (2001), 139–154.

[36] F Marotta-Wurgler. 2015. Does "notice and choice" disclosure regulation work? An empirical study of privacy policies. *Michigan Law: Law and Economics Workshop* (2015). https://www.law.umich.edu/centersandprograms/lawandeconomics/workshops/Documents/Paper13.Marotta-Wurgler.Does%20Notice%20and%20Choice%20Disclosure%20Work.pdf

[37] Arunesh Mathur, Jessica Vitak, Arvind Narayanan, and Marshini Chetty. 2018. Characterizing the use of browser-based blocking extensions to prevent online tracking. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS)*.

[38] Aleecia M. McDonald and Lorrie F. Cranor. 2008. The Cost of Reading Privacy Policies. *I/S: A Journal of Law and Policy for the Information Society* 4, 3 (2008), 540–565.

[39] Aleecia M McDonald and Lorrie Faith Cranor. 2010. Americans' Attitudes About Internet Behavioral Advertising Practices. In *Proceedings of the Workshop on Privacy in the Electronic Society (WPES)*.

[40] William Melicher, Mahmood Sharif, Joshua Tan, Lujo Bauer, Mihai Christodorescu, and Pedro Giovanni Leon. 2016. (Do Not) Track Me Sometimes: Users' Contextual Preferences for Web Tracking. *Proceedings on Privacy Enhancing Technologies* 2016, 2 (2016), 135–154.

[41] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2017. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405* (2017).

[42] Mozilla. 2019. Geckodriver. https://github.com/mozilla/geckodriver. (2019).

[43] Kanthashree Mysore Sathyendra, Shomir Wilson, Florian Schaub, Sebastian Zimmeck, and Norman Sadeh. 2017. Identifying the Provision of Choices in Privacy Policy Text. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2774–2779. https://doi.org/10.18653/v1/D17-1294

[44] Network Advertising Initiative. 2018. NAI Code of Conduct. (2018). https://www.networkadvertising.org/sites/default/files/nai_code2018.pdf.

[45] Network Advertising Initiative. 2019. Opt Out of Interested-Based Advertising. (2019). http://optout.networkadvertising.org/.

[46] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[47] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365* (2018).

[48] Postlight Labs, LLC. 2019. Mercury Web Parser. https://mercury.postlight.com/web-parser/. (2019).

[49] Usable Privacy Policy Project. 2017. Usable Privacy Policy project website. https://usableprivacy.org/. (2017).

[50] Enric Pujol, Oliver Hohlfeld, and Anja Feldmann. 2015. Annoyed Users: Ads and Ad-Block Usage in the Wild. In *Proceedings of the Internet Measurement Conference*.

[51] Joel R Reidenberg, Travis Breaux, Lorrie Faith Cranor, Brian French, Amanda Grannis, James T Graves, Fei Liu, Aleecia McDonald, Thomas B Norton, Rohan Ramanath, N. Cameron Russell, Norman Sadeh, and Florian Schaub. 2015.

Disagreeable privacy policies: Mismatches between meaning and users' understanding. *Berkeley Tech. LJ* 30 (2015), 39.

[52] L. Richardson. 2004. Beautiful Soup. https://www.crummy.com/software/BeautifulSoup/. (2004).

[53] A. Ronacher. 2010. Flask. http://flask.pocoo.org/. (2010).

[54] Norman Sadeh, Alessandro Acquisti, Travis D Breaux, Lorrie Faith Cranor, Noah A Smith, Fei Liu, Florian Schaub, and Shomir Wilson. 2013. The Usable Privacy Policy Project: Combining Crowdsourcing, Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About. *Tech. report CMU-ISR-13-119, School of Computer Science, Carnegie Mellon University,Pittsburgh, PA 15213, USA* (December 2013).

[55] Salesforce.com, inc. 2018. Salesforce DMP Privacy. https://www.salesforce.com/products/marketing-cloud/sfmc/salesforce-dmp-privacy/. (12 June 2018).

[56] Iskander Sanchez-Rola, Matteo Dell'Amico, Platon Kotzias, Davide Balzarotti, Leyla Bilge, Pierre-Antoine Vervier, and Igor Santos. 2019. Can I Opt Out Yet?: GDPR and the Global Illusion of Cookie Control. In *Proceedings of the ACM Asia Conference on Computer and Communications Security.*

[57] Florian Schaub, Aditya Marella, Pranshu Kalvani, Blase Ur, Chao Pan, Emily Forney, and Lorrie Faith Cranor. 2016. Watching Them Watching Me: Browser Extensions' Impact on User Privacy Awareness and Concern. In *Proceedings of NDSS Workshop on Usable Security (USEC).*

[58] Selenium project. 2004. Selenium. https://www.seleniumhq.org/. (2004).

[59] Fatemeh Shirazi and Melanie Volkamer. 2014. What Deters Jane from Preventing Identification and Tracking on the Web?. In *Proceedings of the Workshop on Privacy in the Electronic Society (WPES).*

[60] Joseph Turow, Jennifer King, Chris Jay Hoofnagle, Amy Bleakley, and Michael Hennessy. 2009. Americans Reject Tailored Advertising and Three Activities That Enable It. https://ssrn.com/abstract=1478214.143.

[61] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, Useful, Scary, Creepy: Perceptions of Online Behavioral Advertising. In *Proceedings of the Symposium on Usable Privacy and Security (SOUPS).*

[62] U.S. Federal Trade Commission. 2009. CAN-SPAM Act: A Compliance Guide for Business. https://www.ftc.gov/tips-advice/business-center/guidance/can-spam-act-compliance-guide-business. (2009).

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems.* 5998–6008.

[64] 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

[65] Yaxing Yao, Davide Lo Re, and Yang Wang. 2017. Folk Models of Online Behavioral Advertising. In *Proceedings of the Conference on Computer-Supported Cooperative Work and Social Computing (CSCW).* 1957–1969.

[66] L. Yu, X. Luo, X. Liu, and T. Zhang. 2016. Can We Trust the Privacy Policies of Android Apps?. In *DSN '16.*

[67] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N. Russell, and Norman Sadeh. 2019. MAPS: Scaling Privacy Compliance Analysis to a Million Apps. *Proceedings on Privacy Enhancing Technologies* 2019 (07 2019), 66–86. https://doi.org/10.2478/popets-2019-0037

[68] S. Zimmeck, Z. Wang, L. Zou, R. Iyengar, B. Liu, F. Shaub, S. Wilson, N. Sadeh, S. M. Bellovin, and J. Reidenberg. 2016. Automated Analysis of Privacy Requirements for Mobile Apps. In *NDSS '16.* https://www.ndss-symposium.org/ndss2017/ndss-2017-programme/automated-analysis-privacy-requirements-mobile-apps/