Scraping Cookie and Activity Declarations from Privacy Policies

Brian Jay Tang
University of Michigan
Computer Science and Engineering
2260 Hayward St, Ann Arbor, MI 48109
bjaytang@umich.edu

Abstract

Recently, privacy laws such as the GDPR or the CCPA have been created to protect user privacy. A large portion of the legal responsibility falls on service providers to self-regulate and protect user privacy. Currently, verifying the correctness/consistency of privacy policies relies on large amounts of debugging, network traffic analysis, and consulting with privacy experts. We propose to develop a framework for verifying the correctness/consistency of privacy policies and ensuring that privacy policies are consistent with data collection during runtime. For the scope of this course, we developed (1) a crawler/scraper (2.69% false positive rate) capable of navigating online websites to reach privacy policies, cookie policies, and consent settings and (2) a fine-tuned language model (0.73 F1 score) capable of classifying cookie declarations, activity types, data collection types, and other frequently occurring privacy policy sentences.

1 Introduction

Data privacy has been a major concern at the forefront of many discussions about online technologies. One of the main challenges with online privacy is adequately and correctly informing users about the collection and processing of their data. The current "notice and consent" framework of data privacy has been the *de facto* standard for disclosing data practices to users. Unfortunately, privacy policies are often long, complicated, vague, or even inaccurate (Bui et al., 2021b), making them difficult to validate. On top of this, a vast majority of consumers (91%) skip reading them (Deloitte, 2016), likely because it would take a person 76 work days to read through all the policies they encounter in a year (McDonald and Cranor, 2008). Nevertheless, privacy policies remain an important form of disclosure for legal reasons and users.

Recently, privacy laws such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) have been created and updated to protect user privacy. While some improvements have been made (Linden et al., 2018), regulators are still struggling to enforce these protection laws on a large-scale (Naughton, 2020). As a result, a bulk of the responsibility falls on service providers to self-regulate and protect user privacy. Currently, verifying the correctness/consistency of privacy policies relies on a significant amount of debugging, network traffic analysis, and consulting with privacy professionals. Because of this labor cost and the legal ramifications, the primary focus is generally placed on compliance with regulations rather than on the usability of privacy controls and privacy policy disclosures (Waldman, 2021).

The primary focus of our project is to (1) extensively crawl the top 25k websites to search for privacy policy pages and their supplementary pages, (2) extract cookie and activity declarations from

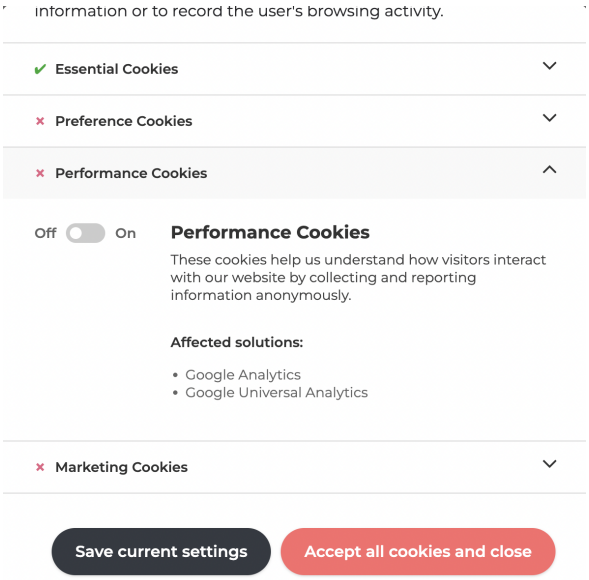


Figure 1: An example of cookie settings that a user can interact with to accept or reject specific cookie types. Prior research has shown that many of these cookie settings are inconsistent or fail to respect user's consent.

privacy policies, and (3) classify these declarations. We used NLP techniques and transformer-based language models to accomplish these tasks. This will enable creating various usable privacy tools for automating regulation and compliance, identifying and blocking intrusive cookies, and creating context-aware privacy notifications for users. Throughout this project, we will create several annotated text datasets of privacy policies, allowing other researchers to create additional tools.

Our contributions are as follows:

1. We created two datasets, one containing $2,178$ privacy-related URL annotations, and the other containing $2,578$ sentence annotations for 42 class types.
2. We developed a crawler to systematically search for web pages containing cookie declarations, cookie lists, or cookie tables. It achieves a 2.69% false positive rate (FPR).
3. We fine-tuned a classifier to classify 42 types of sentences, cookie declarations, and activities in privacy policy pages. It achieves a 0.73 F1 score.

2 Relevant Prior Work

Numerous solutions leveraging machine learning have been proposed for improving and verifying the correctness/consistency of privacy practices. Recently, PI-Extract (Bui et al., 2021a) and Polisis (Harkous et al., 2018, 2016) use DNN language models trained on privacy policies to automatically extract privacy practices, highlighting and describing the data practices. Opt-Out Easy (Bannihatti Kumar et al., 2020) extracts and classifies opt-out choices in the form of a browser extension. PrivacyCheck (Zaeem et al., 2018) extracts and summarizes information from privacy policies and presents information about data practices to its users by answering a list of 20 privacy questions. Additionally, PurPliance (Bui et al., 2021b), PolicyLint (Andow et al., 2019), and PoliCheck (Andow et al., 2020) construct data flows from privacy policies and cross-check these data flows with actual observed data practices. These systems use NLP to extract data flows from privacy-policy documents and evaluate their consistency with a dynamic analysis. While some of these privacy tools have been designed for large-scale dynamic privacy analysis, their code and UI exercising coverage is minimal.

Notably, almost all of these prior works focus on UI interaction and privacy policy compliance separately, so it is important to integrate them, especially when monitoring data collection and contexts during collection. Additionally, none of the prior works investigate using NLP techniques for extracting cookie declarations and activity contexts.

While the recent prior work by Bollinger *et al.*(Bollinger et al., 2022) relied on the descriptions of cookies specified in CMPs, our goal was to also encompass websites that implemented their own cookie preference settings and cookie declarations. One such cookie consent settings menu is presented in Fig. 1.

3 Design

Since the task of extracting cookie declarations from unstructured text is challenging, we first develop a crawler capable of locating privacy and cookie policies. Upon identifying a policy page, we convert the web page into plaintext sentences and use a classifier to detect the presence of individual cookie names, purposes, expiration dates, domains, etc. This pipeline automates crawling any website page to find policy pages, detecting the presence of individual cookie declarations, and identifying the purposes and expiration dates for specific cookies. Figure 2 demonstrates how this pipeline can enable more powerful capabilities in the future.

3.1 UI Execution and Scraping

Our tool is capable of navigating to the privacy policy and cookie policy pages of any given URL by using Playwright (pla, 2022). It uses CSS selectors and keywords to find desired pages and interactable elements. The crawler recursively interacts with site pages to reach new page contexts. After parsing the HTML, we extract the plaintext for a site's page. We plan to tokenize each word and interactive element while splitting each web page into sentences based on punctuation, headers, and div/span elements. We crawled the top $19,423$ websites according to the Tranco list (Pochat et al., 2018) and extracted $9,136$ candidate URLs. Figure 3 shows how the crawler navigates to privacy policy pages.

In the next phase of the project, our tool will collect every HTTP request/response and detect patterns resembling certain data types. For example, it decodes HTTP traffic and cookies and searches for matches in information such as IP address, location,

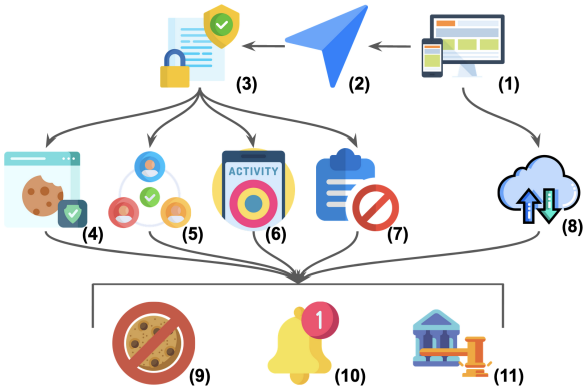


Figure 2: A demonstration of how our crawler and classifier will enable usable privacy tools. (1) A user visits a website, (2) our crawler navigates to (3) the privacy or cookie policy, (4) our classifier extracts cookie declarations, (5) 3rd party entities, (6) activities the user performs when data is collected, and (7) buttons/links to reject cookies or opt-out of data collection/tracking/selling/advertising. (8) As the user enters the website, the network traffic is monitored by a browser extension, so that our tools can (9) block cookies or data collection, (10) notify the user of harmful privacy violations or data practices, and (11) automatically audit companies' data practices.

clickstreams, previously visited URLs, page text, advertising IDs, and more. We will analyze and compare the actual data practices with the practices disclosed in the privacy policies.

3.2 Fine-Tuning Classifiers

Each input sentence contains a corresponding label for our classification task. With the annotated data, we will fine-tune a bidirectional transformer (Sanh et al., 2019) to classify each sentence embedding into labels corresponding to the cookie declaration or activity data purposes. Additionally, as these contexts may be established over multiple sentences, we will train models at 3 levels of granularity: clauses, sentences, and paragraphs. We annotated sentences according to the classes defined in Table 1.

4 Methodology

4.1 Website Crawler

Starting from any initial website URL or home page, our crawler recursively executes UI actions to reach page contexts related to keywords such as "privacy", "cookies", "privacy policy", "cookie policy", "cookie table", "cookie list", and more. It searches for navigation links by using CSS selectors, element text, and a list of keywords. This

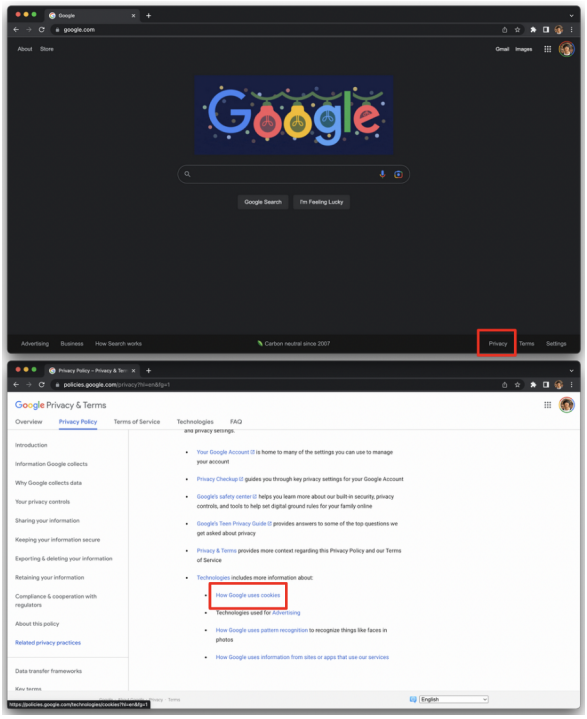
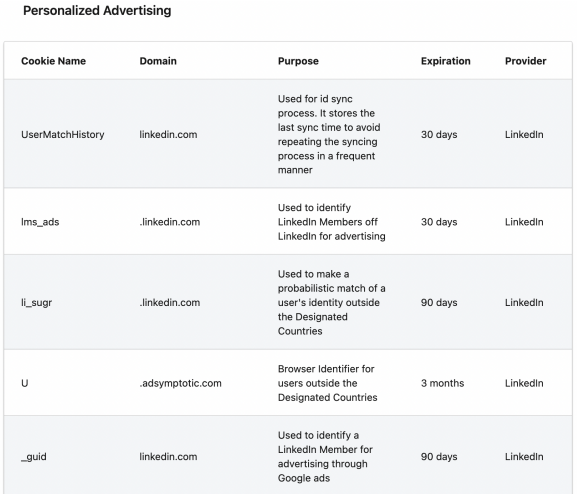


Figure 3: An example of our crawler exercising Google's UIs to navigate to the cookie policy page. It clicks on the "Privacy" link followed by the "How Google uses cookies" link.

crawler also leverages Playwright, a UI testing framework for web browsers.

4.2 Cookie Declaration Classifier

After extracting the HTML from each successfully crawled and annotated site, we attained the plaintext representation of each site. We split each privacy and cookie policy page into sentences and annotated a subset of 2,578 sentences and cookie declarations with the 42 classes defined in Table 1 in the Appendix (inter-annotator agreement of 92% of 100 URLs). In order to create a fine-grained classifier capable of distinguishing between cookie declarations and other privacy policy sentences, we constructed this set of classes to include individual cookie declarations (e.g., cookie name, expiration, or purpose) and sentences in the privacy policy relating to legal compliance, data collection methods, or 3rd party processing. We then fine-tuned a cased DistilBERT language model (Sanh et al., 2019) to classify these sentences in order to detect cookie declarations.



Personalized Advertising

Cookie Name	Domain	Purpose	Expiration	Provider
UserMatchHistory	linkedin.com	Used for id sync process. It stores the last sync time to avoid repeating the syncing process in a frequent manner	30 days	LinkedIn
lms_ads	.linkedin.com	Used to identify LinkedIn Members off LinkedIn for advertising	30 days	LinkedIn
li_sugr	.linkedin.com	Used to make a probabilistic match of a user's identity outside the Designated Countries	90 days	LinkedIn
U	.adsymptotic.com	Browser Identifier for users outside the Designated Countries	3 months	LinkedIn
_guid	linkedin.com	Used to identify a LinkedIn Member for advertising through Google ads	90 days	LinkedIn

Figure 4: An example of a list of cookie declarations in table form for LinkedIn. Listed are personalized advertising cookies with their names, domains, purpose, expiration, and 1st/3rd party provider.

5 Evaluation

5.1 Website Crawler

Overall, of the top 19,423 websites from the Tranco list (Pochat et al., 2018), we successfully crawled and discovered privacy and cookie policy pages for unique $4,283$ websites and extracted $9,136$ candidate URLs. A large portion of sites either used anti-crawling measures, timed out, were in a language other than English, or did not contain a privacy policy page. From any website URL, it is able to navigate and identify privacy & cookie policy pages with a false positive rate of only 2.69% (35 URLs). After manually annotating $2,178$ of these URLs for individual cookie declarations and cookie settings (inter-annotator agreement of 94% of 50 URLs), we found 225 (11.29%) unique pages containing individual cookie declarations and 147 (17.28%) unique pages containing cookie settings. Cookie consent mechanisms and declarations were not very prevalent on websites. Overall, few sites used cookie settings or provided specific details for the cookies they used. Most websites only vaguely described the categories and purposes of cookies used on their site without providing details or lists of these cookie names. Another common practice was for privacy policies to declare cookie names, their expiration dates, and their purposes, but not provide users with adequate consent mechanisms, simply instructing users to disable *all* cookies on their browsers. This can result in users losing the ability to properly use certain site features as a sacrifice to preserve their privacy.

5.2 Cookie Declaration Classifier

We used an 80–20 split on the annotated dataset, and our classifier achieved a 0.73 average F1 score on the test set after weighting and averaging the F1 score for each class. The accuracy on the test set was 74.42% and 92.62% on the training set. The confusion matrix containing all of the classes seen in the test set can be found in Fig. 6. We used a batch size of 4, 8 epochs, a learning rate of $1e-5$ with the Adam optimizer. Although there was a big class imbalance, for classes with more data, the classifier performed quite well. Figure 5 shows an example of the classified sentences and how cookie declarations can also appear naturally in a cookie policy.

6 Discussion

6.1 Limitations

Our project is not without several limitations. One of the biggest limitations was the dataset imbalance for our sentence classifier. Another limitation is that our preprocessing of HTML policies to plaintext sentences can sometimes result in paragraph-long sentences. Since these manual annotations were mainly performed by only one PhD student with 3+ years experience in security & privacy research, we consulted another former PhD student with 5+ years experience. This researcher validated small subsets of our annotations.

6.2 Future Directions

Going forward, we plan to improve the classifier by balancing the dataset with additional annotations. Once the classifier achieves higher performance on a test set, we will use the classifier to automatically annotate sentences and manually validate the predicted labels. The next thrust of the project will be to identify specific classes of activities and use reinforcement learning to teach the crawler to perform and identify these contexts (e.g., logging in, filling out forms, adding items to cart, etc.). Afterwards, we will use the collected network traffic to perform a large-scale analysis of website policies and any potential privacy violations or inconsistencies. Finally we will develop a tool to notify users/regulators of privacy violations as they occur while blocking these data collection activities.

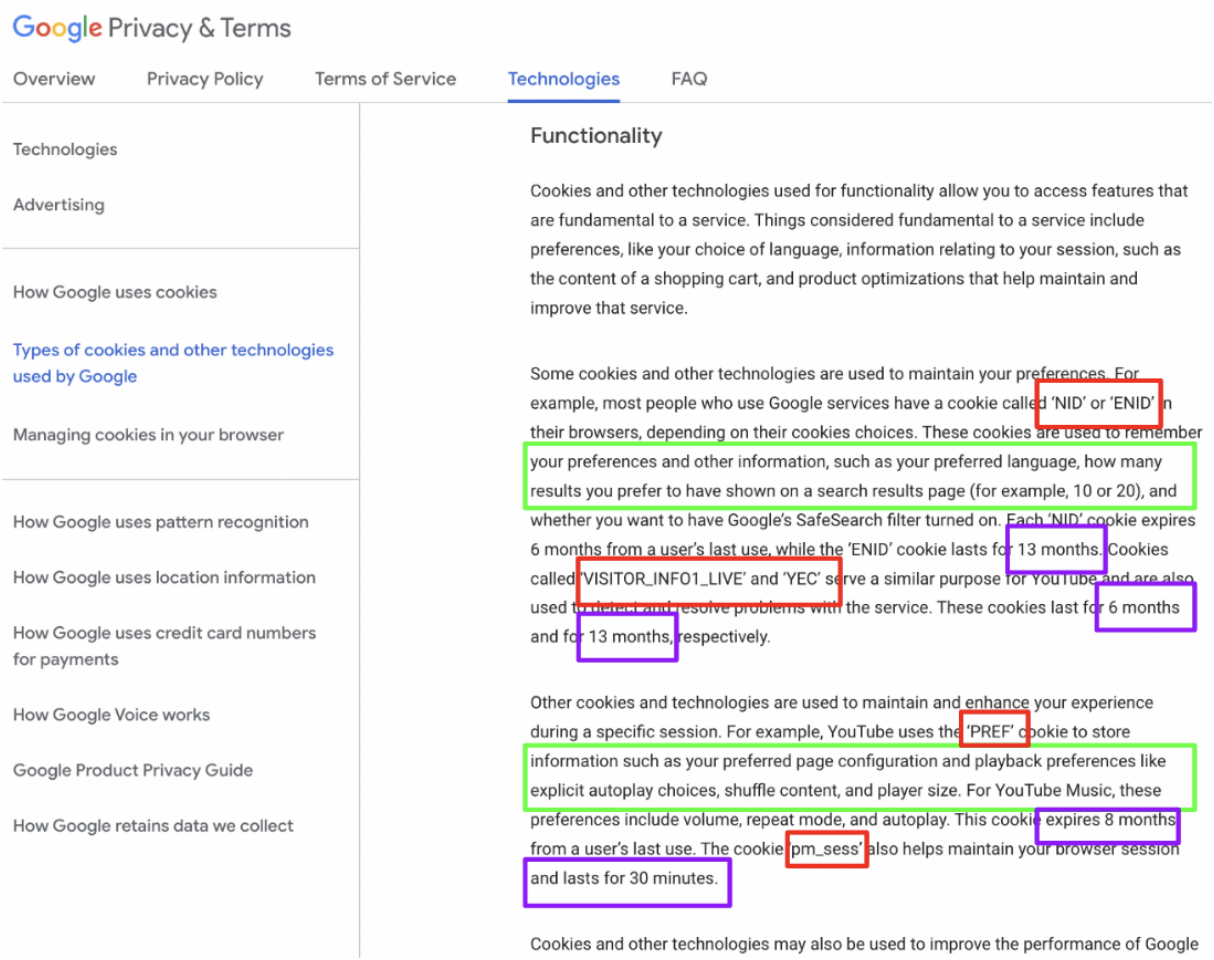


Figure 5: An example cookie policy from https://google.com with individual cookie declarations annotated. Cookie names are highlighted in red, cookie expirations are purple, and cookie purposes are green.

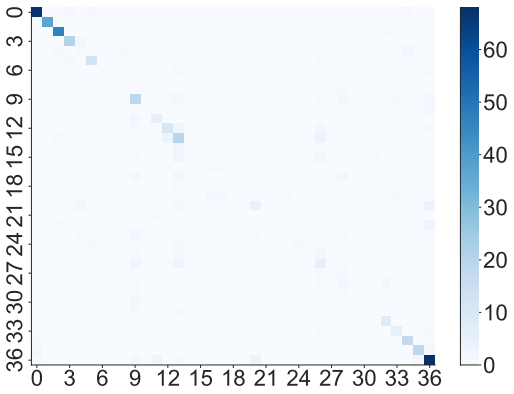


Figure 6: The test set confusion matrix for our classifier. Note the dataset imbalance.

7 Conclusion

Overall, our project sets the foundation for usable privacy tools that can automate regulation and compliance, identify and block intrusive cookies, and create context-aware privacy notifications for users. Throughout this project, we created several annotated datasets of privacy policies and URL pages, also allowing other researchers to create additional tools. We developed a crawler capable of navigating online websites to reach privacy policies, cookie policies, and consent settings with a 2.69% false positive rate. We also fine-tuned a language model to classify cookie declarations, activity types, data collection types, and other frequently occurring privacy policy sentences with a 0.73 F1 score on the test set. The code and datasets can be found at https://github.com/byron123t/cookie-tables

References

2022. Fast and reliable end-to-end testing for modern web apps | Playwright. [Online; accessed 13. Sep. 2022].

Benjamin Andow, Samin Yaseer Mahmud, Wenyu

Wang, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Tao Xie. 2019. {PolicyLint}: Investigating internal privacy policy contradictions on google play. In *28th USENIX security symposium (USENIX security 19)*, pages 585–602.

Benjamin Andow, Samin Yaseer Mahmud, Justin Whitaker, William Enck, Bradley Reaves, Kapil Singh, and Serge Egelman. 2020. Actions speak louder than words:{Entity-Sensitive} privacy policy and data flow analysis with {PoliCheck}. In *29th USENIX Security Symposium (USENIX Security 20)*, pages 985–1002.

Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, et al. 2020. Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text. In *Proceedings of The Web Conference 2020*, pages 1943–1954.

Dino Bollinger, Karel Kubicek, Carlos Cotrini, and David Basin. 2022. Automating cookie consent and GDPR violation detection. In *31st USENIX Security Symposium (USENIX Security 22)*, page TBA, Boston, MA. USENIX Association.

Duc Bui, Kang G Shin, Jong-Min Choi, and Junbum Shin. 2021a. Automated extraction and presentation of data practices in privacy policies. *Proc. Priv. Enhancing Technol.*, 2021(2):88–110.

Duc Bui, Yuan Yao, Kang G Shin, Jong-Min Choi, and Junbum Shin. 2021b. Consistency analysis of data-usage purposes in mobile apps. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pages 2824–2843.

US Deloitte. 2016. Global mobile consumer survey: Us edition. *Deloitte US*.

Hamza Harkous, Kassem Fawaz, Rémi Lebret, Florian Schaub, Kang G Shin, and Karl Aberer. 2018. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th {USENIX} security symposium ({USENIX} security 18)*, pages 531–548.

Hamza Harkous, Kassem Fawaz, Kang G Shin, and Karl Aberer. 2016. {PriBots}: Conversational privacy with chatbots. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*.

Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2018. The privacy policy landscape after the gdpr. *arXiv preprint arXiv:1809.08396*.

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

John Naughton. 2020. Data protection laws are great. Shame they are not being enforced. *the Guardian*.

Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczyński, and Wouter Joosen. 2018. Tranco: A research-oriented top sites ranking hardened against manipulation. *arXiv preprint arXiv:1806.01156*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Ari Ezra Waldman. 2021. *Industry unbound: The inside story of privacy, data, and corporate Power*. Cambridge University Press.

Razieh Nokhbeh Zaeem, Rachel L German, and K Suzanne Barber. 2018. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)*, 18(4):1–18.

Table 1: Privacy/Cookie Policy Sentence Annotations

Category	**Class**	**Index**
Cookies	Cookie Name	0
	Cookie Expiration	1
	Cookie Purpose	2
	Cookie Description	37
	Party Entity	3
	Cookie Category	4
	Cookie Domain	5
	Settings Button	6
	Preference Status	7
	About Cookies	8
	Table Header	39
Legality	Legal Compliance	9
	Copyright Statement	10
	Company Contact	11
Data collection	Data Collected	12
	Data Usage Purpose	13
	Cookie Usage	14
	Collection Activity	15
Opt-Out	Opt-Out Button	16
	Opt-Out Directions	17
	Opt-Out Status	18
	Opt-Out Effect	19
	DAA/NAI	20
	Marketing Opt-Out	21
	Do Not Sell	22
	Browser Cookies	23
Other	Section Header	24
	Company Description	25
	Product Description	26
	Other	27
	Scope	32
	Marketing	33
	Security	34
	Data Deletion	35
	Transparency Request	36
	Privacy Policy Link	38
	Language/Country	40
	Navigation Link	41
3rd Parties	Entity Name	28
	Data Usage Purpose	29
	Data Processing	30
	Opt-Out Link	31