Measuring the Privacy Dimension of Free Content Websites through Automated Privacy Policy Analysis and Annotation

Abdulrahman Alabduljabbar
University of Central Florida
Orlando, FL, USA
jabbar@knights.ucf.edu

David Mohaisen
University of Central Florida
Orlando, FL, USA
mohaisen@ucf.edu

ABSTRACT

Websites that provide books, music, movies, and other media free of charge are a central piece of the web ecosystem, although they are vastly unexplored, especially for their security and privacy risks. In this paper, we contribute to the understanding of those websites by focusing on the comparative analysis of their privacy policies, a primary channel where service providers inform users about their data collection and use. To better understand the data usage risks associated with such websites, we study 1,562 websites and their privacy policies in contrast to premium websites. We uncover that premium websites are more transparent in reporting their privacy practices, particularly in categories such as *"Data Retention"* and *"Do Not Track"*, with premium websites are 85.00% and ≈70% more likely to report their practices in comparison to the free content websites. We found the free content websites' privacy policies to be more similar to one another and generic in comparison to the premium websites' privacy policies. Our findings raise several concerns, including that the reported privacy policies may not reflect the data collection practices used by service providers, and various pronounced biases across privacy policy categories. This calls for further investigation of the risks associated with the usage of such free content websites and services through active measurements.

CCS CONCEPTS

• **Information systems → World Wide Web**; **Web mining**; • **Security and privacy → Usability in security and privacy**.

KEYWORDS

Free Content Websites, Privacy Policy, Natural Language Processing, Web Security

ACM Reference Format:
Abdulrahman Alabduljabbar and David Mohaisen. 2022. Measuring the Privacy Dimension of Free Content Websites through Automated Privacy Policy Analysis and Annotation. In *Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25–29, 2022, Virtual Event, Lyon, France.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3487553.3524663

WWW '22 Companion, April 25–29, 2022, Virtual Event, Lyon, France

ACM ISBN 978-1-4503-9130-6/22/04…$15.00
https://doi.org/10.1145/3487553.3524663

1 INTRODUCTION

One very important class of websites on the web is for services that provide content for free, relying on the online ad ecosystem for generating revenues and for their operation. Those websites are in contrast to premium websites, which provide the same type of content by charging for them through a monthly subscription or a pay-per-use business model. Both types of websites provide various content, including software, books, movies, music, etc. They are very popular [5–8, 12, 13, 20], attracting significant traffic towards them and are placed high on websites ranking.

The security of free content websites is a central issue and a concern in the broad treatment of web security, and has recently attracted the attention of the research community. For instance, it is long believed that free content websites are a source of lurking dangers, as they are several times more likely to host malicious scripts that would expose users to significant risks. Moreover, free content websites, by design, are less likely to be maintained, making it significantly more likely for their software (e.g., web platform) to go unpatched for discovered vulnerabilities [14, 21, 33]. This, in turn, opens boundless opportunities for adversaries to exploit such vulnerabilities, take control over the websites, and serve malicious content to their visitors. Even worse, many of these websites use invalid digital certificates [15], allowing adversaries to create fake websites to impersonate them, and deliver malicious content to their users without even gaining control over the original website.

All those issues are concerning, and have resulted in several initiatives to systematically analyze and understand the characteristics of those websites in terms of the security they offer. The main finding in the relevant literature is that free content websites offer significantly degraded security guarantees than those offered by premium websites. Given this clear gap in the security characteristics, one would be concerned with the privacy assurances of those websites as well. That is, how do the privacy assurances and guarantees of those websites compare to premium websites?

While the question of privacy might seem initially arbitrary, it is indeed an intelligent question that is nicely fitting in this context and stems from a deep understanding of the ecosystem those websites fall in. In particular, as free content websites are prone to compromise, due to their poor security qualities, they expose their operators to significant liabilities. Such liability is best exposed in legal documents that define the boundaries of responsibilities of those websites and their operators. In the context of those websites, such a legal document is known as the privacy policy.

The privacy policy is a legal document that defines and discloses how the website, represented by its operator, collects, uses, discloses, and manages a website visitor's data. Exploring the privacy policy characteristics of those websites will shed light on behaviors

and perhaps intents of the websites owners. For instance, advertisements can be exploited for data leakage, in addition to running malicious scripts on the user device [27]. Moreover, service providers may reportedly collect and sell users' data to increase their profit margin. The absent of this insight on policies with respect to those practices, makes it difficult to understand them, and their actual analysis is desired to estimate the real risk.

In this work, we utilized a classifier for automated privacy policy annotation, achieving a baseline annotation F_1 score of 91%. Then, we used the model for uncovering the privacy policies reporting discrepancy between the free content and premium websites. Towards this goal, our analyses uncover that premium websites are more transparent in reporting their privacy practices. This is more evident in categories such as *"User Choice"*,*"Data Retention"* and *"Do Not Track"*, with premium websites are 51.33%, 85.00%, 69.92%, more likely to report their practices in comparison to the free content websites. Moreover, the premium websites' privacy policies are more concise and to-the-point, where 58.96% of the free content websites' segments are assigned to at least one of the categories, in comparison with 64.33% of their premium counterparts (+5.37% difference). Further, we investigate the privacy policy uniqueness and similarity to other policies in our dataset. The free content websites' privacy policies have ≈11% higher similarity scores in comparison to the premium websites. Our results highlight that the reported privacy policies by free content websites may not accurately represent the service provider's data collection practices, shedding light on additional risk dimensions to free content websites.

Contributions and Findings. With a list of 1,562 free content and premium services websites obtained from the top results of Google, DuckDuckGo, and Bing search engines, the privacy policies are extracted and analyzed toward understanding the service providers' reporting practices across the following verticals.

(1) **Privacy Policy Reporting (§5.1).** We analyze the free content websites to understand the reporting practices of data collection, uncovering that premium websites are more transparent in reporting collection, sharing, and retention practices.
(2) **Privacy Embedded Information (§5.2).** Through a segment and word-level analysis, we find that premium websites are more concise and are to-the-point, while free content websites' privacy policies are less likely to contain useful information regarding the privacy policy practices.
(3) **Generic Privacy Policies (§5.3).** Through similarity analysis of policies, we show that the free content websites tend to use generic privacy policy templates, with a 33.05% increase in similarity score in comparison with premium websites.

Organization. In section 2, we present the related work. An in-depth overview of the utilized privacy policy annotator is provided in section 3. We discuss our compiled dataset in section 4. Our results and discussion are in section 5. The concluding remarks and future work are in section 6.

2 RELATED WORK

The literature work that best aligns with our research is categorized into two primary research directions: (i) website analysis and (ii) privacy policy analysis. In the following, we will review recent and central work that corresponds to both of these directions.

Websites Analysis. Websites are advancing rapidly in terms of content and user base growth, with significant enhancements in the intricacy and diversity of their components. Nonetheless, the interaction between these components results in a variety of risks.

One of the important analysis modalities of websites has been their digital certificates [3, 4, 15]. For instance, Chung *et al.* [15] addressed this issue by presenting an in-depth analysis of websites' certificates in the online Public Key Infrastructure and found that most websites' certificates are invalid. Moreover, they discussed the source of invalid websites' certificates and contended that end-user devices created all the invalid certificates of the websites, and these certificates were reproduced regularly with new self-signatures.

Libert *et al.* [28] examined the privacy-compromising policies of one million prominent websites. They evaluated policies, e.g., for data leakage, to identify potential data breaches, and the study concluded that nine out of ten online websites shared user data with third-party services without the user's permission. Moreover, Lavrenovs *et al.* [23] used a similar dataset and presented a detailed evaluation of Alexa's top-million websites' security. The research made four types of requests to each website in order to access HTTP header information and investigate the existence of web security-related response header variables. They discovered that 29.1% of HTTPS servers had invalid TLS (Transport Layer Security) configurations, and only 17.5% of websites used the HTTP Strict Transport Security policy. These results raise concerns about the security policies of such famous websites' security protocols.

Creating environments to assess the security flaws in web-based services is a challenging task. Alsmadi *et al.* [9] proposed a component-based testing framework to evaluate the security flaw in web applications by checking numerous invalid inputs and used this mechanism to assess website behavior owing to such inputs. As a result, the security of web applications is strengthened by eliminating the invalid inputs (i.e., rejecting invalid inputs), which are central features of the attack surface. The authors also offered multiple ways to identify invalid inputs and uncovered several vulnerabilities.

Privacy Policy Analysis. Websites' privacy policies inform users about their processes for data collection and processing. These websites are responsible for providing information regarding collecting, storing, and managing users' data. However, their privacy policies may be unclear, and some users may not comprehend those policies even when they review them carefully due to the lack of experience in understanding the technical languages used in such policies. Therefore, it is crucial to assess these policies to overcome various concerns, including readability and comprehensibility.

The early studies on automatic privacy policy analysis and understanding emphasize machine-readable policies to verify privacy policies of web-based services. For instance, the Platform for Internet Content Selection (PICS) [16] framework is one of the earliest works presented to verify the privacy policies of web-based services. Additionally, the Platform for Privacy Preferences (P3P) [18] was established to offer online users with a machine-readable language for articulating privacy policies. Typically, privacy policies contain machine-readable languages. However, natural language is preferred for privacy policies making natural language processing

(NLP) techniques an ideal tool for extracting legal information from documents and fully comprehending the privacy policies.

Ammar*et al.* [10] initiated the research on information extraction in privacy policies with a pilot study where they categorized the information disclosure in those policies into two classes: (1) to law enforcement authorities, and (2) the account deletion policies. Furthermore, they show that natural language analysis is a viable choice for such a task. Similarly, Constante *et al.* [17] executed a rule-based identification of users' data collected by online services and used NLP to assess the identification performance. They extended their rule-based technique with a machine learning-based approach for analyzing whether a privacy policy provides enough information on various privacy features of the evaluated websites.

Zimmeck *et al.* [38] presented a browser extension that retrieves the analyses of policy by utilizing NLP techniques applied over a repository of policies. Other studies [11, 37] analyzed the manually annotated privacy policies and discovered significant inconsistencies in data collecting and sharing policies. Harkous *et al.* [22] employed a dataset that contains 130K privacy policies to train a privacy-centric language model and presented an automated framework to analyze the privacy policy. In their study, the authors proposed model produced 88.4% accuracy in structured requests and 82.4% accuracy in the top-3 responses of free-form queries.

Wilson *et al.* [35] presented OPP-115, a baseline privacy policy dataset that contained nine classes and was annotated by skilled law students. The OPP-115 contained text in paragraphs form, and these paragraphs are classified into nine categories. The study used the Paragraph2Vec embedding [24] and three machine learning classifiers: (1) Logistic Regression (LR), (2) Support Vector Machine (SVM), and (3) Hidden Markov Model (HMM). The proposed classification model produced an average 0.66 micro F_1 score.

Liu *et al.* [30] used the OPP-115 dataset with new embedding, classifiers, and classification granularity. Unlike Wilson *et al.* [35], Liu *et al.* [30] used the TF-IDF weighting scheme at the sentence and paragraph granularity and used two machine learning and one neural network-based classifiers: (1) LR, (2) SVM, and (3) Convolutional Neural Networks (CNN). To evaluate the performance of the classifiers, they used the micro F_1 score and produced 0.66 for the sentence-based and 0.78 for the paragraph-based technique.

In an earlier study, Liu *et al.* [29] used unsupervised learning approaches on the OPP-115 dataset to analyze policies. Unsupervised learning approaches do not require a labeled dataset, proving beneficial in understanding privacy policies cost-effectively. As such, the authors used a Non-negative Matrix Factorization (NMF) technique [25] to create a lexicon for each category based on expert-defined mappings between subject models and categories.

Although the previous literature emphasizes the potential of the proposed models, the proposed techniques are incapable of achieving good accuracy on benchmark datasets. Therefore, Alabduljabbar *et al.* [1, 2] introduced TLDR, which used a variety of text representations and machine learning algorithms to address the technical gap in the literature by boosting the accuracy. They further presented a case study by analyzing the Alexa top 10,000 websites' privacy policies using TLDR's pipeline.

This Work. This work is a *case study*, in essence, where the goal is to understand the data collection and practices embedded in the privacy policies of free content websites. The goal is to uncover and report the differences between the stipulations of the privacy policies of the free content websites and their premium counterparts. For achieving this discovery, we use TLDR's pipeline.

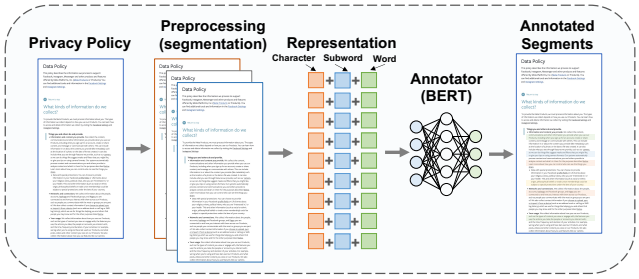

Figure 1: An overview of TLDR's pipeline. The processed segments are represented using different feature representation techniques, and then fed to the corresponding category classifier for multi-label classification.

3 PRIVACY POLICY ANNOTATION PIPELINE

To understand the differences between the free content websites and their premium counterparts with respect to their privacy policy and data usage reporting, we leverage TLDR [2], a pipeline of automated machine and deep learning models that was designed to extract privacy and data collection practices directly from the policies. The overview of TLDR pipeline is shown in Figure 1. In TLDR, the segments are first preprocessed, then various WordPiece [36] representations are applied to extract deep representative features. Afterward, a classifier of Bidirectional Encoder Representations from Transformers (BERT) is used to predict the corresponding labels of each segment in a multi-label classification setting.

Upon establishing a baseline for our learning model by training and validation, we proceed to examine the reporting practices of free content and premium websites regarding first and third data parties collection and tracking (the complete details and configurations of the training evaluation of TLDR are provided in [2]).

3.1 Training Set: Ground Truth Annotation

In their seminal work, Wilson *et al.* [35] employed a taxonomy for segment labeling, which we show in Figure 2, and describe in Table 1 (for the main categories/class labels). Based on this taxonomy, the privacy policies are divided into high-level and low-level categories, where each category includes key information about the policy practices, such as "first-party data collection", "third-party information sharing", and "user tracking", among others. For our automated annotator, we used the high-level categories as a class label. It is worth noting that the annotation within a segment's phrase can be applied to the entire section. For example, each privacy policy category's existence or absence is given a binary label (positive or negative) in the segment for generalization.

3.2 Privacy Policy Preprocessing

Our study follows the same segmentation protocol as that of Wilson *et al.* [35]. In their work, the authors defined *segments* in each privacy policy and identified each segment by the separator ("|||"). Moreover, the privacy policies in OPP-115 were saved as Hypertext

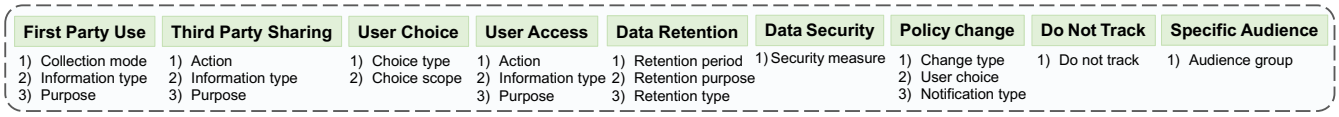

Figure 2: The taxonomy used by Wilson *et al.* [35] in categorizing the privacy policy practices and labeling each segment. We consider the high level nine categories in the process of utilizing the classifier.

Table 1: Privacy policies' high-level categories. The classifier is trained on these categories, classifying each segment as positive and negative in the context of each category.

Category	Description
1st Party Use	How and why data is collected by a service provider.
3rd Party Sharing	How data is collected and shared with third parties.
User Choice	Whether users have control over their data.
User Access	How users can access, edit, or delete their data.
Data Retention	How long the stored user data is retained.
Data Security	Methods of securing and protecting user data.
Policy Change	If/how a provider informs users about policy change.
Do Not Track	If/how a provider honors online and ad tracking.
Specific Audiences	(e.g., children, Europeans, or California residents).

Markup Language (HTML) files. In addition, each segment comprises multiple *sentences*, and these sentences typically consider one or more areas of the service provider's privacy protocols.

Segment Representation. Our study employs WordPiece [36] to present segments to BERT [19] for annotation as shown in Figure 1 for the general learning pipeline. We used WordPiece to preprocess the original segment. Technically, it generates a set of words, subwords, and characters for a given amount of characters. WordPiece is also favored over other techniques in the literature because it can naturally help breaking unrecognized words into subwords to address the out-of-vocabulary problem by predefining a dictionary. The candidate word in WordPiece is divided into characters and mapped to the relevant embedding when subwords are not recognized within the predefining dictionary.

Learning Algorithms. BERT [19] is a transformer-based language model based on attention provided by the transformer architecture [34]. This language model contains two layers of encoders and decoders, where each encoder and decoder includes six layers. Each encoder and decoder layer can learn the contextual links between words in a specific context. Moreover, BERT outperformed traditional machine learning and deep learning models in various tasks, such as text classification, information retrieval, and named entity identification. In this study, each segment is preprocessed using WordPiece [36], which uses words, subwords, and character-level terms matching methods to employ the BERT model.

3.3 Experimental Setup and Baseline

Training and Validation. To train our annotation pipeline of the privacy policies, we followed the same process followed by the authors of the original TLDR in [2]. In particular, the OPP-115 dataset is split into *training* and *validation* sets, with document-based splitting (that is, a whole document is considered as a sample), where 80% of the documents are used for training the classifier while the remaining 20% of the documents are used for validation. We recall that each segment is represented using WordPiece [36] representation, and forwarded to the BERT model for binary classification.

Table 2: TLDR's performance (F_1) using the best performing word representations and learning algorithms on OPP-115.

Category	TLDR	Wilson [35]	Harkous [22]	Liu [30]
First party	0.94	0.75	0.79	0.81
Third party	0.89	0.7	0.79	0.79
User choice	0.85	0.61	0.74	0.70
User access	0.91	0.61	0.80	0.82
Data retention	0.87	0.16	0.71	0.43
Data security	0.88	0.67	0.85	0.80
Policy change	0.95	0.75	0.88	0.85
Do not track	1.00	1.00	0.95	1.00
Specific audiences	0.94	0.70	0.95	0.85
Overall	0.91	0.66	0.83	0.78

Table 3: An overview of the collected dataset.

Type	Books	Games	Movies	Music	Software	Overall
Free Content	154	80	331	83	186	834
Premium	195	113	152	86	182	728
Total	349	193	483	169	368	1,562

The BERT model is trained with learning rates of $[5 \times e^{-5}, 3 \times e^{-5}, 2 \times e^{-5}]$, and ten training epochs. The best performing BERT model is obtained with 1,000 features and $2 \times e^{-5}$ learning rate.

Metrics. The performance of the machine learning algorithms is measured through widely used confusion metrics, such as precision, recall, and F_1 scores. The precision metric is used to determine the correct positives identified by the classifier and answer the question "*How many segments labeled as positive are correct?*". The precision value is calculated as $P = TP/(TP + FP)$, where TP denotes true positives and FP shows false positives, which refers to negative segments that were inaccurately classified as positive by the learning model. Similarly, the recall metric determines "*how many positive segments were properly categorized?*" and is measured as $R = TP/(TP + FN)$, where FN shows false negatives, denotes positive segments that were incorrectly labeled as negative. In other words, the classifier considered these positive segments as negative and classified them incorrectly. F_1 measure is used to evaluate the overall performance of the classifier, which uses the values of both precision and recall. F_1 measure is derived as $F_1 = 2 \times (P \times R)/(P + R)$.

Results. Trained on the manually annotated dataset, TDLR is then used as an oracle for annotating the privacy policies of the free content and premium websites. To this end, we report the best performing evaluation results of TLDR on the OPP-115 in Table 2. For comparison, we also provide the results on the same dataset for the techniques proposed by Wilson *et al.* [35] and Liu *et al.* [30] (using F_1 as a measure) which shows that TLDR is superior.

4 FREE CONTENT WEBSITES DATASET

For our analyses, we prepared a list of 1,562 free content (834) and premium (728) websites. We considered two primary factors to select the websites for the analyses: (1) Choosing the most popular

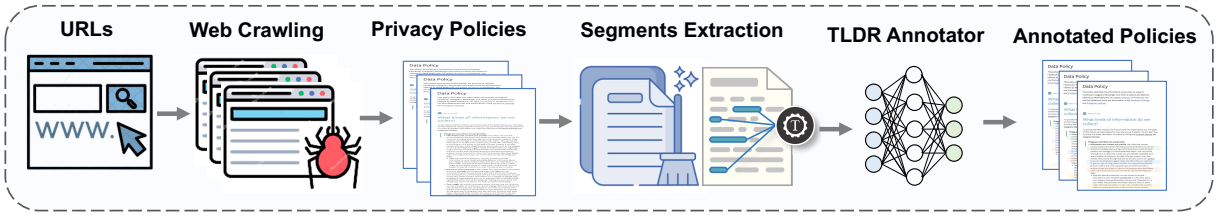

Figure 3: Our data collection and segment extraction pipeline, including crawling the website structure and searching for the privacy policy. Once found, paragraphs are extracted and preprocessed to extract the policy segments.

Table 4: An overview of the crawled privacy policies showing the number of retrieved and validated privacy policies and the average number of segments and words per policy for free content and premium websites. TP=Total Policies, VP=Valid Policies, TS=Total Segments, AS=Avg. Segments, TW=Total Words, AW=Avg. Words.

Free Content Websites							
Group	URLs	TP	VP	TS	AS	TW	AW
Books	154	89	55	3,825	69.55	149,079	2710.53
Games	80	39	23	1,382	60.09	57,266	2489.83
Movies	331	213	84	5,285	62.92	278,077	3310.44
Music	83	54	28	3,022	107.93	119,546	4269.50
Software	186	90	62	3,090	49.84	103,159	1663.85
Overall	834	485	252	16,604	65.89	707,127	2806.06
Premium Websites							
Group	URLs	TP	VP	TS	AS	TW	AW
Books	195	161	121	5,399	44.62	258,859	2139.33
Games	113	94	74	5,430	73.38	278,582	3764.62
Movies	152	137	99	5,384	54.38	282,355	2852.07
Music	86	73	50	3,617	72.34	154,944	3098.88
Software	182	160	124	7,749	62.49	328,250	2647.18
Overall	728	625	468	27,579	58.93	1,302,990	2784.17

websites, such as those that appear in Google, DuckDuckGo, and Bing search results, and (2) maintaining a balanced dataset. We also individually and manually inspected and annotated each website in our dataset. Furthermore, the websites are then divided into five distinct categories depending on their content: (1) books, (2) games, (3) movies, (4) music, and (5) software. The distribution of the eventually utilized dataset is shown in Table 3.

Privacy Policy Extraction. We first start by crawling the privacy policies of each website among the free content and premium websites in our dataset. Selenium [32], an automated browser testing framework that enables extensions to mimic user interaction with a web browser/web server, is used for this task by passing the appropriate user-agent as a parameter to the HTTP requests. Subsequently, as shown in Table 4, we extracted the privacy policies of 1,110 websites from this list. In order to be able to obtain the privacy policies from those websites, we traverse all of their accessible pages starting with the home directory using the scan capability of Selenium. Finally, the privacy policies are retrieved by examining the pages of each website that include various terms, such as "privacy policy", "privacy terms", or "privacy statement".

The linked HTML with the privacy policy is kept intact for processing once identified. We notice that the remaining websites among those in our initial set are either in a foreign language or their privacy policy is not directly obtained from their structure using our aforementioned heuristic.

For our analysis, a python library called BeautifulSoup [26] was used to extract all paragraphs using the HTML paragraph tag ($< p >$). It is important to note that the BeautifulSoup library has been widely used for parsing HTML and XML documents. Our candidate segments are based on the extracted paragraphs.

Upon extracting the segments, all segments containing fewer than ten words were discarded because such segments generally include introduction phrases and do not contribute significant information about the privacy and data collection practices. The remaining segments are then linked to the extracted privacy policy for category analysis. Figure 1 illustrates the order and steps of the crawling and cleaning process of a website.

Validation and Filtering. For validation and as a form of sanity check, we thoroughly examined the extracted policies to determine the correctness of the extraction process. We found that 64.86% of the policies were correctly extracted. Consequently, we only considered the correctly extracted policies for accurate analysis.

Data Preprocessing & Representation. The extracted segments, 44,183 in total, are further preprocessed and represented in a manner similar to that described in section 3.2 for the OPP-115 segment preprocessing. Figure 3 depicts the data preprocessing and representation processes in more detail.

5 RESULTS AND DISCUSSION

After extracting the privacy policies from the free content and premium websites, we apply the pretrained TLDR model on the OPP-115 dataset to annotate and classify the segments of each website. In this section, and towards the main goal of this study, we will measure and discuss the main differences between free and premium websites using the following dimensions: (1) privacy policy reporting and transparency, (2) the usefulness of the privacy policy information with respect to each policy category, and (3) whether the policies are reused among free and premium websites.

5.1 Privacy Practices Reporting

Understanding the reporting practices of collecting data and information by different websites is critical for user understanding of the risks associated with using a service (*i.e.,* data leakage and privacy), particularly when such a service is provided free of charge. Upon passing the different filtered privacy policies into our pipeline, we collect annotated policies and aggregate the number of websites that contain each policy category. Table 5 shows the percentage of the websites containing various privacy policy categories for free

Table 5: The percentage of websites with positive segments per category for free and premium websites.

	Books			Games			Movies			Music			Software			Overall		
Category	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff
First Party Use	81.82	98.35	+16.53	95.65	95.95	+0.29	84.52	94.95	+10.43	100.00	90.00	-10.00	85.48	95.97	+10.48	86.90	95.73	+8.82
Third Party Sharing	80.00	95.87	+15.87	91.30	89.19	-2.12	85.71	87.88	+2.16	96.43	88.00	-8.43	79.03	87.10	+8.06	84.52	89.96	+5.43
User Choice	63.64	79.34	+15.70	73.91	78.38	+4.47	34.52	83.84	+49.31	64.29	82.00	+17.71	53.23	75.00	+21.77	52.38	79.27	+26.89
User Access	52.73	70.25	+17.52	13.04	60.81	+47.77	67.86	67.68	-0.18	57.14	80.00	+22.86	33.87	57.26	+23.39	50.00	65.81	+15.81
Data Retention	38.18	52.07	+13.88	30.43	67.57	+37.13	22.62	57.58	+34.96	53.57	70.00	+16.43	25.81	50.81	+25.00	30.95	57.26	+26.31
Data Security	80.00	69.42	-10.58	65.22	85.14	+19.92	73.81	72.73	-1.08	71.43	74.00	+2.57	48.39	76.61	+28.23	67.86	75.00	+7.14
Policy Change	72.73	76.86	+4.13	65.22	77.03	+11.81	77.38	76.77	-0.61	92.86	70.00	-22.86	54.84	62.10	+7.26	71.43	72.22	+0.79
Do Not Track	14.55	18.18	+3.64	0.00	24.32	+24.32	13.10	31.31	+18.22	28.57	24.00	-4.57	8.06	14.52	+6.45	12.70	21.58	+8.88
Specific Audiences	80.00	67.77	-12.23	60.87	82.43	+21.56	71.43	86.87	+15.44	78.57	74.00	-4.57	50.00	65.32	+15.32	67.86	74.15	+6.29

Table 6: The percentage of highlighted segments from free and premium websites of each category.

	Books			Games			Movies			Music			Software			Overall		
Category	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff
First Party Use	24.78	35.21	+10.43	8.79	27.81	+19.02	25.93	36.62	+10.69	29.36	35.18	+5.82	23.20	32.41	+9.21	25.76	32.91	+7.15
Third Party Sharing	16.13	17.44	+1.31	6.07	16.10	+10.03	18.57	16.46	-2.11	16.67	16.11	-0.57	13.74	16.96	+3.22	16.00	15.77	-0.23
User Choice	5.96	6.69	+0.73	3.62	6.26	+2.64	6.02	6.37	+0.35	5.70	6.99	+1.30	4.90	5.17	+0.26	5.70	6.12	+0.42
User Access	3.78	3.40	-0.38	0.63	3.22	+2.59	3.47	3.54	+0.07	3.29	3.27	-0.02	2.67	3.16	+0.49	3.23	3.14	-0.09
Data Retention	2.56	2.12	-0.44	0.82	2.62	+1.80	2.08	1.83	-0.25	2.32	2.35	+0.02	2.49	1.81	-0.68	2.43	1.89	-0.53
Data Security	3.59	2.93	-0.67	3.62	3.09	-0.53	2.95	2.33	-0.62	2.79	2.17	-0.62	4.03	2.81	-1.22	3.39	2.62	-0.76
Policy Change	2.78	2.49	-0.29	1.90	1.89	-0.01	2.34	2.85	+0.51	2.13	2.08	-0.05	2.00	3.51	+1.51	2.22	2.65	+0.44
Do Not Track	0.44	0.37	-0.08	0.00	0.44	+0.44	0.67	0.30	-0.37	0.47	0.44	-0.03	0.28	0.30	+0.02	0.45	0.31	-0.13
Specific Audiences	7.06	9.14	+2.08	3.80	7.61	+3.80	10.10	8.58	-1.52	6.17	10.09	+3.92	5.99	7.43	+1.44	7.34	8.37	+1.03
All Categories	59.62	70.16	+10.54	27.36	60.41	+33.05	63.47	67.44	+3.98	60.96	69.29	+8.33	53.41	64.33	+10.91	58.96	64.33	+5.37

Table 7: The percentage of highlighted words from free and premium websites of each category.

	Books			Games			Movies			Music			Software			Overall		
Category	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff	Free	Prem	Diff
First Party Use	28.69	41.78	+13.09	9.96	35.48	+25.52	46.30	28.32	-17.98	40.75	37.29	-3.46	30.00	39.33	+9.33	40.45	31.41	-9.04
Third Party Sharing	22.50	20.09	-2.41	6.07	21.17	+15.10	21.79	23.66	+1.87	18.14	21.67	+3.53	17.21	19.15	+1.94	19.15	21.04	+1.88
User Choice	6.53	6.13	-0.40	3.56	6.81	+3.25	6.46	6.50	+0.04	7.43	7.10	-0.33	5.62	6.28	+0.66	6.29	6.42	+0.13
User Access	5.07	3.95	-1.12	0.71	4.33	+3.62	3.76	3.74	-0.02	4.30	3.97	-0.34	2.94	3.16	+0.22	3.56	3.96	+0.40
Data Retention	3.78	2.66	-1.12	0.73	3.80	+3.07	2.37	2.92	+0.55	3.11	3.40	+0.29	3.15	2.13	-1.02	2.39	3.40	+1.01
Data Security	4.70	2.89	-1.81	4.89	4.12	-0.77	2.59	3.41	+0.82	2.10	3.68	+1.58	5.18	2.33	-2.84	2.72	4.29	+1.58
Policy Change	3.35	2.54	-0.81	2.21	2.20	-0.01	2.76	3.18	+0.42	2.20	2.07	-0.14	3.05	6.14	+3.10	3.07	2.84	-0.23
Do Not Track	0.53	0.27	-0.26	0.00	0.49	+0.49	0.22	0.66	+0.44	0.34	0.42	+0.08	0.35	0.26	-0.09	0.24	0.49	+0.25
Specific Audiences	9.39	8.42	-0.96	5.04	9.66	+4.61	8.52	16.25	+7.73	9.36	9.68	+0.32	8.36	9.08	+0.72	8.44	10.71	+2.26
All Categories	71.09	73.09	+2.00	29.98	73.13	+43.15	73.90	74.52	+0.61	72.83	73.79	+0.96	64.81	69.61	+4.81	69.37	71.01	+1.64

and premium content. The comparison is shown for each category of free content websites, books, games, movies, music, and software, as well as for the overall combined set of categories. In our analysis, we consider both the per-category and overall trends.

In our per category analysis, we notice the diversity in behaviors covered in our results when comparing the positive segments, compared to the overall behavior. While generally the privacy policies are well articulated to cover all aspects of a privacy policy by premium websites to a higher degree than those in the free content websites, we notice that the premium websites perform significantly worse for "Books" on two privacy categories (*"Data Security"*; with 80% in free vs. 69.42% in premium, and *"Specific Audience"*; with 80% vs 67.77%, respectively) and "Music" on five categories (*"First Party Use*; with 100% in free vs. 90% in premium, *"Third Party Sharing"*; with 96.43% in free vs. 88% in premium, *"Policy Change"*; with 92.86% in free vs. 70% in premium, *"Do Not Track"*; with 28.57% in free vs. 24% in premium, and *"Specific Audience"*; with 78.57% in free vs. 74% in premium), while performing marginal worse for "Games" on one category (*"Third Party Sharing"*; with 91.30% in free vs. 89.19% in premium) and for "Movies" on three categories (*"User Access"*; with 67.86% in free vs. 67.86% in premium, *"Data Security"*; with 73.81% in free vs. 72.73% in premium, and *"Policy Change"*; with 77.38% in free content websites vs. 76.77% in premium websites).

One explanation for the performance difference between books and music in both free and premium website categories, in contrast to games, movies, and software, is perhaps to limit the responsibility of website concerning data security, targeted audience, and general use, to avoid legal battles as those categories of content seem to have been the most targeted content with lawsuits pertaining to stricter classifications and regulations of copyrights.

By the same token, and with the exception of the aforementioned privacy categories for the content categories, the premium content outperformed the free content websites on every privacy category, with margins ranging from 0.29% (*"First Party Use"* in "Games") to as high as 47.77% (*"User Access"* in "Games"). This shows that, despite the occasional detailed and well-annotated language of the free content website, they still are lax with their policy, and not pronouncing the various essential elements that guard the use and provide remedies for abuse of users' data.

As we pointed out earlier through our per category analysis, the premium content websites generally are more comprehensive in reporting their data collection, sharing, and retention practices (last group in Table 5). This is more evident in categories such as

"User Choice",*"Data Retention"* and *"Do Not Track"*, with premium websites being 51.33%, 85.00%, 69.92%, more likely to report their practices in comparison to the free content websites.

Our measurements and experimentation evaluation show that among "Games" free content websites, 0% report their user tracking practices, in comparison with 24.32% of their premium counterparts. The same observation can be made for *"User Access"*, with 13.04% and 60.81% of free and premium content websites reporting information regarding this category, respectively.

Key Takeaway: Overall, the premium websites' privacy policies are more elaborate and transparent in reporting their data collection, sharing, and retention practices. This pattern is persistent, although shown to deviate in favor of the free content websites in two groups and for only a few privacy policy categories. Moreover, in extreme cases, 0% of the "Gaming" free content websites report their user tracking practices, which is alarming. Reporting practices are essential for users' awareness of risks associated with the service. In contrast to the premium websites, the lack of such reporting in free content websites highlights the high risk associated with their usage, given the lack of policy-level guarantees.

5.2 Privacy Policies Embedded Information

Privacy policies are lengthy statements that can be overwhelming for ordinary users to read and comprehend. Therefore, it is essential for these statements to be to-the-point, and not to add unrelated information that may confuse users. This is particularly understood, given the often usage of indirect language exploited by service providers to hide their privacy practices in complex language framing. This, in turn, calls for an in-depth and fine-grained analysis. In Table 6, we show our results of such analysis by reporting the percentage of segments (*i.e.,* paragraphs) annotated by TLDR and assigned to one of the nine categories. Overall, 58.96% of the free content websites' segments are assigned to at least one of the categories, in comparison with 64.33% of their premium counterparts (+5.37% difference). While this difference (percentage) might not seem significant, it has an interesting implication: that the presence (or absence) of language cues in a segment is sufficient to topically drift the annotation of the document with respect to a given class label, which supports our initial claim concerning indirect and overly (and intentionally) complex language framing.

Taking the results forward, we further analyze the micro differences across categories. We notice that despite the small increase, the gap is much larger for "Books", "Games", and "Software" websites. For instance, 27.36% of the "Games" free content websites' segments are assigned to at least one privacy policy category, in comparison with 60.41% of the premium websites' segments (120.79% increase). Moreover, the highlighted words by TDLR for "Games" free content websites are 43.15% less than their premium counterpart, as shown in Table 7.

Analyzing the highlighted information per category, we observe that privacy policies practices are covered widely in premium websites in comparison with their free counterparts. For instance, the *"First Party Use"* privacy policy is highlighted within 24.78% of "Books" free websites' segments, in comparison with 35.21% of premium websites' segments. In cases where free websites' segments, as shown earlier, have a higher highlighting ratio, we notice that the difference is marginal (-1.52% only). However, word-wise, this margin becomes non-trivial, with "Movies" free websites' highlighted *"First Party Use"* words being significantly higher than the highlighted words for premium "Movies" websites (46.30% vs. 28.32%, with -17.98% difference). This may be a byproduct of free content websites privacy policies using generic privacy reporting templates that are not necessarily reflective of their specific practices, as we show later in more detail in subsection 5.3.

Key Takeaway: From the word and segment level analysis, we find distinctive patterns among each type: the premium websites' privacy policies are richer, providing more to-the-point information In contrast, free content websites' privacy policies are less likely to contain useful information regarding privacy policy practices.

Table 8: The similarity in (%) between the privacy policies of each group for free content and premium websites.

Group	Free Content	Premium	Diff	% Diff
Books	53.96	50.74	3.22	6.15
Games	57.77	56.74	1.04	1.81
Movies	67.92	48.66	19.26	33.05
Music	62.03	55.65	6.38	10.84
Software	52.26	42.12	10.14	21.48
Overall	54.38	43.45	10.93	22.34

5.3 Privacy Policy Content Reuse

Despite their importance, many websites may adapt generic privacy policy templates that are not necessarily reflective of their actual privacy practices. Understanding the importance of having customized privacy policies for the website-provided services, we investigate the privacy policy uniqueness. For that purpose, we calculate the similarity between each privacy policy and other privacy policies in our dataset. In particular, we used Pysimilar [31], a python library for computing the similarity between two strings by using TF-IDF vectorizer and the cosine similarity metric to compute a similarity score between two documents.

Table 8 shows the average similarity (as a percentage; the cosine similarity scaled up to 100) among websites' privacy policies in our dataset. Notice that, across all categories, the free content websites' privacy policies have higher similarity scores in comparison to the premium websites. This is more evident in categories such as "Movies", with free content websites' privacy policies average similarity score being more than 33% in comparison with its premium counterpart. Overall, the average similarity score of free websites' policies is ≈11% more than premium websites similarity score (*i.e.,* 54.38% for free websites vs. 43.45% for premium websites).

Key Takeaway: Free content websites are more likely to use generic privacy policies templates, with ≈33.05% increase in similarity score in comparison with premium websites for "Games" category. The usage of generic templates in free websites indicates that the reported privacy policies may not reflect the actual data collection practices used by the websites' owners (service providers).

6 CONCLUSION AND FUTURE WORK

The Internet is the most widely used medium for marketing, promotion, and communication in the digital era, especially when offering

traditional and digital content. Moreover free content websites that offer publicly accessible free content have grown in popularity in recent years. We explored the privacy policies reporting practices of free and premium content websites, unveiling that the premium content websites are more transparent in reporting their privacy practices, particularly in categories such as *"Data Retention"* and *"Do Not Track"*, with premium websites are 85.00% and 69.92% more likely to report their practices. Our findings also uncover that free content websites' privacy policies are similar to one another and are generic, with ≈11% higher similarity scores.

Toward a safe and secure web environment, we highlight that free content websites would highly benefit from consistent monitoring and management, particularly with the lack of data collection and sharing practices. Our observations in this study raise concerns regarding the safety of using such free services, especially when such usage could put users at risk, and call for an in-depth analysis of their actual risks, ramifications, and remedies.

Acknowledgement. This research was supported by Global Research Laboratory (GRL) Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT of Korea (NRF-2016K1A1A2912757). A. Alabduljabbar is also supported in part by the Saudi Arabian Cultural Mission (SACM).

REFERENCES

[1] Abdulrahman Alabduljabbar, Ahmed Abusnaina, Ülkü Meteriz-Yildiran, and David Mohaisen. 2021. Automated Privacy Policy Annotation with Information Highlighting Made Practical Using Deep Representations. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS'21)*. 2378–2380.

[2] Abdulrahman Alabduljabbar, Ahmed Abusnaina, Ülkü Meteriz-Yildiran, and David Mohaisen. 2021. TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights. In *ACM Workshop on Privacy in the Electronic Society (WPES'21)*. ACM, 103–118.

[3] Abdulrahman Alabduljabbar, Runyu Ma, Sultan Alshamrani, Rhongho Jang, Songqing Chen, and David Mohaisen. 2022. Poster: Measuring and Assessing the Risks of Free Content Websites. In *Network and Distributed System Security Symposium, (NDSS'22), San Diego, California, April 24-28, 2022*. The Internet Society.

[4] Abdulrahman Alabduljabbar, Runyu Ma, Soohyeon Choi, Rhongho Jang, Songqing Chen, and David Mohaisen. 2022. Understanding the Security of Free Content Websites by Analyzing their SSL Certificates: A Comparative Study. In *In Proceedings of the The 1st International Workshop on Cybersecurity and Social Sciences (CySSS'22), Nagasaki, Japan, May 30 - June 3, 2022*. ACM.

[5] Sultan Alshamrani, Mohammed Abuhamad, Ahmed Abusnaina, and David Mohaisen. 2020. Investigating Online Toxicity in Users Interactions with the Mainstream Media Channels on YouTube. In *The 5th International Workshop on Mining Actionable Insights from Social Networks (MAISoN'20)*. 1–6.

[6] Sultan Alshamrani, Ahmed Abusnaina, Mohammed Abuhamad, Anho Lee, Dae-Hun Nyang, and David A. Mohaisen. 2020. An Analysis of Users Engagement on Twitter During the COVID-19 Pandemic: Topical Trends and Sentiments. In *Proceedings of the 9th International Conference on Computational Data and Social Networks (CSoNet'20)*. Springer, 73–86.

[7] Sultan Alshamrani, Ahmed Abusnaina, Mohammed Abuhamad, Daehun Nyang, and David Mohaisen. 2021. Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in youtube. In *Companion Proceedings of the Web Conference 2021*. 508–515.

[8] Sultan Alshamrani, Ahmed Abusnaina, and David Mohaisen. 2020. Hiding in Plain Sight: A Measurement and Analysis of Kids' Exposure to Malicious URLs on YouTube. In *Third ACM/IEEE Workshop on Hot Topics on Web of Things*. 1–6.

[9] Izzat Alsmadi and Fahad Mira. 2018. Website security analysis: variation of detection methods and decisions. In *Saudi Computer Society National Computer Conference (NCC'18)*. 1–5.

[10] Waleed Ammar, Shomir Wilson, Norman Sadeh, and Noah A Smith. 2012. Automatic categorization of privacy policies: A pilot study. *School of Computer Science, Language Technology Institute, Technical Report CMU-LTI-12-019* (2012).

[11] 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 *USENIX Security Symposium*. 585–602.

[12] Anam Bhatti, Hamza Akram, Hafiz Muhammad Basit, Ahmed Usman Khan, Syeda Mahwish Raza, and Muhammad Bilal Naqvi. 2020. E-commerce trends during COVID-19 Pandemic. *Future Gen. Comm. and Nets* 13, 2 (2020), 1449–1452.

[13] Timm Böttger, Ghida Ibrahim, and Ben Vallis. 2020. How the Internet reacted to Covid-19: A perspective from Facebook's Edge Network. In *ACM Internet Measurement Conference (IMC'20)*. 34–41.

[14] Miguel Carvajal, José A García-Avilés, and José L González. 2012. Crowdfunding and non-profit media: The emergence of new models for public interest journalism. *Journalism practice* 6, 5-6 (2012), 638–647.

[15] Taejoong Chung, Yabing Liu, David R. Choffnes, Dave Levin, Bruce MacDowell Maggs, Alan Mislove, and Christo Wilson. 2016. Measuring and Applying Invalid SSL Certificates: The Silent Majority. In *ACM Internet Measurement Conference (IMC'16)*. 527–541.

[16] World Wide Web Consortium et al. 2003. Platform for Internet content selection (PICS). *http://www. w3c. org/PICS/* (2003).

[17] Elisa Costante, Jerry den Hartog, and Milan Petkovic. 2012. What Websites Know About You. In *Data Privacy Management and Autonomous Spontaneous Security*. 146–159.

[18] Lorrie Faith Cranor. 2002. *Web privacy with P3P - the platform for privacy preferences*. O'Reilly.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*. 4171–4186.

[20] Anja Feldmann, Oliver Gasser, Franziska Lichtblau, Enric Pujol, Ingmar Poese, Christoph Dietzel, Daniel Wagner, Matthias Wichtlhuber, Juan Tapiador, Narseo Vallina-Rodriguez, Oliver Hohlfeld, and Georgios Smaragdakis. 2020. The Lockdown Effect: Implications of the COVID-19 Pandemic on Internet Traffic. In *ACM Internet Measurement Conference (IMC'20)*. 1–18.

[21] Kathryn Greenhill and Constance Wiebrands. 2012. No library required: the free and easy backwaters of online content sharing. *VALA 2012: eM-powering eFutures* (2012).

[22] 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 *USENIX Security Symposium*. 531–548.

[23] Arturs Lavrenovs and F. Jesus Rubio Melon. 2018. HTTP security headers analysis of top one million websites. In *International Conference on Cyber Conflict*. 345–370.

[24] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. In *International Conference on Machine Learning (JMLR Workshop and Conference Proceedings, Vol. 32)*. 1188–1196.

[25] Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for Non-negative Matrix Factorization. In *Neural Information Processing Systems*. MIT Press, 556–562.

[26] Leonard. 2020. Beautiful Soup. https://www.crummy.com/software/BeautifulSoup/

[27] Zhou Li, Kehuan Zhang, Yinglian Xie, Fang Yu, and XiaoFeng Wang. 2012. Knowing your enemy: understanding and detecting malicious web advertising. In *ACM conference on Computer and communications security*. 674–686.

[28] Timothy Libert. 2015. Exposing the Hidden Web: An Analysis of Third-Party HTTP Requests on 1 Million Websites. *CoRR* abs/1511.00619 (2015).

[29] Frederick Liu, Shomir Wilson, F. Schaub, and N. Sadeh. 2016. Analyzing Vocabulary Intersections of Expert Annotations and Topic Models for Data Practices in Privacy Policies. In *AAAI Fall Symposia*.

[30] Frederick Liu, Shomir Wilson, Peter Story, Sebastian Zimmeck, and Norman Sadeh. 2018. Towards automatic classification of privacy policy text. *School of Computer Science Carnegie Mellon University* (2018).

[31] Pysimilar. 2022. Computing the similarity between two string/text. https://pypi.org/project/pysimilar/

[32] Selenium. 2020. SeleniumHQ Browser Automation. https://www.selenium.dev/

[33] Ronald Snijder. 2010. The profits of free books: an experiment to measure the impact of open access publishing. *Learn. Publ.* 23, 4 (2010), 293–301.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[35] Shomir Wilson, Florian Schaub, et al. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In *Annual Mtng of Association for Comp. Linguistics*.

[36] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144* (2016).

[37] Razieh Nokhbeh Zaeem, Rachel L. German, and K. Suzanne Barber. 2018. PrivacyCheck: Automatic Summarization of Privacy Policies Using Data Mining. *ACM Trans. Internet Techn.* 18, 4 (2018), 53:1–53:18.

[38] Sebastian Zimmeck and Steven M. Bellovin. 2014. Privee: An Architecture for Automatically Analyzing Web Privacy Policies. In *Proceedings of the 23rd USENIX Security Symposium,*. USENIX Association, 1–16.