TOWARDS A HOLISTIC AND COMPARATIVE ANALYSIS OF THE FREE CONTENT WEB: SECURITY, PRIVACY, AND PERFORMANCE

by

ABDULRAHMAN ALABDULJABBAR
M.S. University of Glasgow, United Kingdom, 2011
B.S. King Saud University, Saudi Arabia, 2008

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
in the Department of Computer Science
in the College of Engineering and Computer Science
at the University of Central Florida,
Orlando, Florida

Summer Term
2022

Major Professor: David Mohaisen

ABSTRACT

Free content websites that provide free books, music, games, movies, etc., have existed on the Internet for many years. While it is a common belief that such websites might be different from premium websites providing the same content types in terms of their security, a rigorous analysis that supports this belief is lacking from the literature. In particular, it is unclear if those websites are as safe as their premium counterparts. In this dissertation, we set out to investigate the similarities and differences between free content and premium websites, including their risk profiles. Moreover, we analyze and quantify through measurements the potential vulnerability of free content websites. For this purpose, we compiled a dataset of free content websites offering books, games, movies, music, and software. For comparison purposes, we also sampled a dataset of premium content websites, where users need to pay for using the service for the same type of content. For our modality of analysis, we use the SSL certificate's public information, HTTP header information, reported privacy and data sharing practices, top-level domain information, and website files and loaded scripts. The analysis is not straightforward, and en route, we address various challenges, including labeling and annotation, privacy policy understanding through a highly accurate pre-trained language model using advanced ensemble-based classification technique at the sentence and paragraph level, and data augmentation through various sources. This dissertation delivers various significant findings and conclusions concerning the security of free content websites. Our findings raise several concerns, including that the reported privacy policies may not reflect the data collection practices used by service providers, and pronounced biases across privacy policy categories. Overall, our study highlights that while there are no explicit costs associated with those websites, the cost is often implicit, in the form of compromised security and privacy.

To my family.

ACKNOWLEDGMENTS

This work would not have been possible without the support of so many individuals whom I would like to take this opportunity to acknowledge.

First and foremost, I am extremely grateful to my supervisor, Prof. David Mohaisen, for his invaluable advice, tremendous encouragement, continuous support, and guidance during my research journey and Ph.D. study. I would like to also thank him for his availability outside of the office, for the walks and coffee breaks that made my experience pursuing the Ph.D. at UCF wholesome.

I would like to extend my sincere thanks to my doctoral dissertation committee members, Prof. Cliff Zou, Prof. Wei Zhang, and Prof. Sung Choi Yoo, for their valuable feedback and suggestions for improvement at every step of my doctoral milestones.

I am also grateful to my friends and collaborators in the Security and Analytics Lab (SEAL): Afsah, Ahmed, Hisham, Jinchun, Mo (Abuhamad), Mohammad, Necip, Ran, Rhongho, Saad, Soohyeon, Sultan, and Ulku. I would like to also thank my collaborators outside SEAL: Prof. DaeHun Nyang, Prof. Songqing Chen, and Runyu Ma.

Last but not least, I would like to thank my family: my parents, my wife, my daughter "Jana", my sisters, and my brothers for their support and encouragement throughout my Ph.D. journey.

I would like to finally thank the sponsors of the work reported in this dissertation: Prince Sattam Bin Abdulaziz University in Alkharj, Saudi Arabia, and the Saudi Arabian Cultural Mission (SACM) in Washington, D.C., USA.

In addition, part of this dissertation was also supported by the National Research Foundation of South Korea (grant number NRF-2016K1A1A2912757), CyberFlorida (Collaborative Seed Grant), and NVIDIA (GPU Grant).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1: INTRODUCTION

Online services and websites are categorized into two broad categories based on their monetization options: free and premium content websites. While the free content websites provide content free of charge, and are typically sustained by proceeds of advertisements and user donations [30, 61, 64, 138], the premium websites offer services through fees, e.g., subscriptions or pay-as-you-use models. The premium websites ensure a very high level of quality of service as a result of well-designed websites that are well-maintained through dedicated engineering and operational efforts. In contrast, free content websites are believed to lack such a high expectation for the quality of service and are often user-driven.

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 [9, 103, 104]. 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 [30, 61, 64, 138]. 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 [35], 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. In this

dissertation, we study the fundamental differences between free and premium content websites, exposing the vulnerability of free content websites to exploitation and data leakage. To this end, this dissertation makes three different contributions in three different thrusts:

Thrust 1. *Measuring and Modeling the Free Content Websites in the Wild*: In this thrust, we systematically analyze a dataset of 834 free content websites and 728 premium websites on both domain- and content-level analysis, coupled with security analysis across various dimensions. In particular, on domain-level, we examine the domain name system features, creation time, and SSL (Secure Sockets Layer) features as measures of intent. For the content-level analysis, we examine the HTTP (Hypertext Transfer Protocol) request, page size, loading time, content type, which all are measures of website complexity. For security analysis, we examine both the website- and component-level detection and vulnerability using two major off-the-shelf tools, including *VirusTotal* [149] and *Sucuri* [143]. Risk-wise, we found that free content websites are 19 and 2.64 times more likely to be malicious than premium websites at the page-level (38% vs. 2%) and the file-level (45% vs. 17%), respectively.

Thrust 2. *Understanding the Security of Free Content Websites by Analyzing their SSL Certificates*: In this thrust, we investigate the validity of the SSL certificate for both free and premium services. In particular, we focus on understanding the fundamental differences between free and premium content websites in three directions: (i) Errors within the SSL certificate, including unmatched client name, expired certificate, or invalid/vulnerable information and content, (ii) SSL certificate issuer organization analysis, including the most commonly used certificate providers, such as *Cloudflare Inc.*, *Let's Encrypt*, *DigiCert Inc*, and the SSL certification issuer countries distribution analysis (*e.g.* United States, United Kingdom, and Belgium), and (iii) SSL certificate signature algorithm analysis (*e.g. SHA256 with RSA*, *SHA256 with ECDSA*, and *SHA1 with RSA*).

Thrust 3. *A Comprehensive Analysis and Measurements of Free Content Websites' Infrastructure and HTTP Headers*: In this thrust, we investigate the security-related features and encoding

related features by extracting and analyzing the HTTP response header parameters. Through our analysis, we uncover that the most distinguishable features between free and premium content websites are the security and encoding-related HTTP response headers, with premium websites adapting large-scale security configurations, reducing the potential attacks and threat surfaces. On the other hand, our analysis shed light on free content websites practices that can be exploited, including the usage of the third party *iframes* and *redirections*, allowing HTTP communication channels, and transferring data and packets without secure encryption.Our observations and findings confirm that free and premium content websites are indeed distinguishable, with recurring patterns among their corresponding websites. We raise several concerns regarding using free content websites as we unveil that free content websites are more relaxed in their security configurations and protocol adaptation. Analyzing these features uncover the main differences and similarities between the infrastructural and behavioral aspects of free content and premium websites, which is highly beneficial toward accurate modeling of their associated risks.

Thrust 4. *Measuring the Privacy Dimension of Free Content Websites through Automated Privacy Policy Analysis and Annotation*: In this thrust, we investigate several annotation techniques for a practical automation of policy annotation. The goal of our annotation is to provide users with easy-to-interpret high-level annotations on whether various privacy policies they encounter in their daily life meet certain requirements with respect to a broad set of privacy and security expectations. Our pipeline, called TLDR, employs advances in deep representation and machine learning. Then, we utilized TLDR to uncover 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. Further, we investigate the privacy policy uniqueness and similarity to other policies in our dataset. The free content websites' privacy policies have $\approx$11% higher similarity scores in comparison to the premium websites. Then, we leverage the implemented TLDR pipeline to analyze Alexa [16] top-10,000 websites and compare their privacy policy reporting practices

to the free content and premium websites' practices. Analyzing Alexa top-10,000 websites will uncover the commonly used data collection and privacy practices, and how the general practices compared to their free content and premium websites. This allows for better understanding of the trade-off between the provided services and the privacy of the user.

In summary, this dissertation explores the fundamental differences between free and premium content websites, exposing the vulnerability of free content websites to exploitation and data leakage. Our findings highlight that the free content websites are, in general, not safe to use, lack transparency in data collection and information sharing practices, and more likely to contribute toward malicious attacks. These, among other findings, shed light on the high risks associated with the usage of free content websites.

Organization. This dissertation is organized as follows: We review the literature and outline notable related works in chapter 2. An in-depth analysis of the differences and similarities between free content and premium websites across various dimensions: domains, content, and security, is discussed in chapter 3. In chapter 4, we explored the unique SSL certificate characteristics of premium and free content websites to understand their commonalities, differences, and the potential risks associated with expired or invalid SSL certificates. In chapter 5, we explore the structural and fundamental differences between free and premium websites by analyzing their HTTP response headers attributes and flags, in addition to the hosting domain infrastructural behavior. Then, We propose TLDR, a pipeline to automatically and accurately categorize each segment in the privacy policy to its corresponding high-level content category, to explored the privacy policies reporting practices of free content and premium websites in chapter 6. Finally, the concluding remarks of the dissertation are in chapter 7.

CHAPTER 2: LITERATURE REVIEW

In this chapter, we discuss the related literature, starting with work related to websites analysis and malicious web content analysis. Followed by the literature related to online services and SSL certificates measurements. Then, we present the related work of HTTP header analysis and online infrastructure analysis. Finally, we discuss the literature related to privacy policy annotation and analysis.

Websites Analysis and Malicious Web Content Analysis

Websites Analysis. Websites are continuously evolving in terms of content and usage, paralleled by an increase in the complexity and richness of their components. However, with such an evolution, various security risks emerge as a result of the interplay between those components.

Several studies analyzed the most popular websites' privacy policies [3, 4, 8, 15, 76, 86]. For instance, Libert *et al.* [86] evaluated the privacy-compromising practices employed by a million popular websites, e.g., data leakage. They concluded that roughly nine out of ten websites were sharing user data with third-party services without user consent. Using a similar dataset, Lavrenovs *et al.* [76] conducted a comprehensive assessment of the security of Alexa top-million websites. They initiated four types of requests to each website to obtain the HTTP header information and examined the presence of the web security-related response header variables, e.g., Strict-Transport-Security, Content-Security-Policy, X-XSS-Protection, X-Frame-Options, Set-Cookie, and X-Content-Type. They show that 29.1% of HTTPS requests have incorrect TLS (Transport Layer Security) configurations, and the HTTP Strict Transport Security policy is implemented in only 17.5% of the websites. These findings shed light on the worrisome state of the security policies followed by such popular websites.

Exploring environments to evaluate the security flaw in web applications, Alsmadi *et al.* [15] designed a component-based testing mechanism for a variety of invalid inputs and used this mech-

anism to investigate websites behavior, including security. Since the invalid input consistent part of the attack surface, the security of the online services and web applications is strengthened by eliminating those inputs (i.e., reject invalid inputs). To do so, they proposed several methods for detecting invalid inputs, uncovering a large number of SQL injection vulnerabilities.

Websites Security Analysis. Several studies [5, 7, 35, 163] have been conducted in the area of online services analysis to determine the security, privacy, and risk assessment of their websites. For example, as an SSL certificate is often used to evaluate the safety of transferring resources between servers and clients' machines, Chung *et al.* [35] discussed the importance of analyzing the SSL certificate information to demonstrate the risk associated with a website. They investigated over 80M certificates to show that invalid certificates can be used to track users' devices. Another study by Zhang *et al.* [163] explored the patterns of reissuing and revoking the certificates of Alexa Top 1 Million domains over six months. They found that only 28% of those websites reissued their certificate in response to a widespread vulnerability, and 13% of them revoked their SSL certificates three weeks after a vulnerability was disclosed.

In the area of malicious website detection, several studies [72, 145, 150] have been proposed to analyze phishing websites by extracting different features. For example, Barraclough *et al.* [22] proposed a machine learning technique that utilizes blacklist-based, web content-based, and heuristic-based approaches to detect phishing websites with a 99% accuracy. Similarly, a survey by Basit *et al.* [23] discussed several methods of detecting malicious and phishing websites using various machine and deep learning techniques where they differ in depth and comprehensive of the extracted features. One study by Rao *et al.* [121] proposed a website classification method based on various heuristic features. Their method identifies phishing websites based on the URL extracted features, the website content source code, and other third-party features from the search engine results, WHOIS database, and Alexa page ranking. In order to detect or label this kind of website, VirusTotal [149] is one of the online tools that has been heavily used in the literature for malicious website labeling and detection. However, a study performed by Peng *et al.* [114] investigated the

effectiveness of VirusTotal in detecting phishing websites by implementing 66 fake websites mimicked from PayPal and IRS to perform the checking process on them. The authors claimed that there is a shortage and inconsistency between the scanning engines in detecting those websites.

Malicious Web Content Analysis. Recent studies have shown that adversaries are capable of embedding malicious codes within *JavaScript*, *GIF*, or *Redictection* components of the websites [96, 112, 135, 144, 154]. The security (and safety) of end-users depend significantly on detecting and preventing such malicious content, which has also been studied. To do so, researchers have leveraged various features of web applications, including URL (Uniform Resource Locator) domain components, webpage content, HTTP headers, and loaded scripts, and used them to detect malicious web applications [46, 69]. It has also been shown that a promising feature set is the HTTP header information [86], where McGahagan *et al.* [68] leveraged 672 of those features to build a system for malicious website detection. To examine the feasibility of using components and content (i.e., files and scripts) as features for detection, the authors conducted a comprehensive evaluation of different webpage content features, engineering 17 new features that can improve malicious websites detection performance.

One important yet unexplored aspect of websites is the interplay between advertisements deployed on them and their associated maliciousness. Li *et al.* [83] investigated a variety of malicious online advertising and marketing methods, e.g., malware propagation, click frauds, etc. Their study used a large-scale dataset of ads-related web traces, showing the existence of malicious advertisement practices in hundreds of high-ranked websites. To examine the effectiveness of malicious advertisement detection, Masri *et al.* [96] evaluated three tools, *VirusTotal*, *URLVoid*, and *TrendMicro*, showing URLVoid to provide the best performance.

Another prominent threat that has been explored is the distribution of malicious content on free download portals [56]. Such portals can be maliciously utilized for distributing harmful software to end-user devices. Rivera *et al.* [125] conducted a systematic analysis of PUP (Potentially Unwanted Programs) and malware obtained using free download portals, showing that, on average,

8% to 26% of the downloaded content are either PUP or malicious.

Machine learning algorithms have also been widely used for effectively detecting malicious websites [95]. However, machine learning approaches for malicious websites are impaired by two key challenges, features selection, and evasion. To address the feature selection problem for machine learning-based malicious websites detection, Singh and Goyal [136] argued for coupling the feature selection with overhead performance and accuracy in their analysis. Detection evasion, the other issue, is often associated with intrinsic features of today's web ecosystem, including the usage of redirection and hidden iFrames. In this domain, Liu and Lee [88] proposed Convolutional Neural Network-based malicious content detection based on a screenshot of the webpage, and their model was shown effective shown effective.

This Work. We explore the fundamental structural differences between free content and premium website, understanding their behavioral patterns for future characterization. Moreover, we assess the maliciousness of the free content websites in contrast with premium websites. Our findings show worrisome increasing trends in the portion of malicious content within free content websites. To proactively address this concern, we model the risks associated with these websites through easy-to-obtain performance features and identify up to 86.81% of the risky websites verified against ground-truth.

Online Services Analysis and SSL Certificates Measurements

Online Services Analysis. Online services and web applications are evolving in terms of development and utilization. However, with the evolution of their capabilities, different components in these applications can be compromised, invalidating some security aspects and putting their users at risk, a topic that has been of increasing interest.

For instance, one of the security aspects that is not thoroughly studied in the literature is the validity of the websites' certificate [35]. To address this gap, Chung *et al.* [35] proposed the first in-depth

analyses of the invalid certificates in the web public key infrastructure (PKI). The study shows that the vast majority of certificates in the web PKI are invalid. Their study also investigated the source of the invalid certificates, showing that the invalid certificates were mostly generated from end-user devices, with periodic regeneration of new self-signed certificates.

SSL Certificate Measurements. SSL certificate has been widely studied in the literature for online risk and vulnerability analysis of websites [10, 26, 34, 35, 24, 36, 164, 102]. The following is an overview of some recent studies on the topic.

For instance, Meyer *et al.* [100] analyzed the SSL certificate information and content to differentiate between phishing websites and benign websites. Their analysis showed that phishing websites do not, in general, replicate the issuer and subject information but reuse certificates of compromised servers. Moreover, Huang *et al.* [66] analyzed forged SSL certificates on the web. Their analysis unveiled that 0.2% of the studied SSL connections were tampered with forged SSL certificates, where most of them were related to antivirus software and corporate-scale content filters.

Towards accurate malicious SSL certificate detection, Ghafir *et al.* [57] studied the command and control communication channels of malicious SSL certificate services and their generated patterns and traces. Their experimental evaluation highlighted the successful detection of malicious SSL certificates using blacklisting information, the associated IP addresses, and practices. More recently, Wang *et al.* [155] statically and dynamically analyzed the SSL certificates to extract potential exploitation and vulnerabilities within Android applications. Their analysis showed that 11.07% of the studied applications are prone to man-in-the-middle and phishing.

In this work, we explore the SSL certificate-based fundamental structural differences between free content and premium websites, by understanding their patterns for modeling and characterization.

This Work. We explore the fundamental and structural differences between free content and premium websites within their Secure Sockets Layer (SSL) certificate content. In particular, we focus on the validity analysis, issuer analysis, and signature analysis of the SSL Certificates.

HTTP Header Analysis. HTTP response header attributes have been intensively used to measure websites' security and performance. A study by Mendoza *et al.* [98] analyzed the inconsistencies of HTTP security response configuration between mobile and desktop in the top 70,000 websites. Such inconsistencies can lead to vulnerabilities and some possible attack scenarios that compromise the security and privacy of web users. These inconsistencies emerge due to the weak development and deployment of website configuration to hold on both the desktop and mobile browsers.

Another work by Gadient *et al.* [54] explored the use of HTTP response headers with a focus on the security-related features to study the safety of web communications on 3,376 mobile applications. Querying 9,714 URLs, they collected the corresponding HTTP response headers by performing the HTTP GET request. Their findings conclude that 93% of the security-related header attributes are not enabled, which may cause data leakage or arbitrary code execution. As a result, they uncover the lack of important security configurations among Android apps.

Moreover, Lavrenovs and Melon [76] analyzed the HTTP response headers of Alexa top-1M websites. The authors focused on exploring the security-related attributes, such as the Strict-Transport-Security, and X-XSS-Protection features. They further found the correlations between deploying the HTTP header features and the popularity ranking of a website. The study unveils that popular websites are more likely to apply HTTP response security headers.

Benefiting from the aforementioned HTTP response headers extracted information, several works showed the capabilities of machine learning for malicious website detection, utilizing HTTP header information and associated attributes. For example, recent studies by Laughter *et al.* and McGahagan *et al.* [68, 75] have used the HTTP response header attributes as features to detect whether a website is associated with malicious activities. Those studies showed that various features, such as a website's URL and HTTP request length, can be used to detect malicious requests with an accuracy of 96.9%.

Online Infrastructure Analysis. Content owners deploy their websites on private or public infrastructures. Private infrastructures enable custom security optimizations and full authority on the website infrastructure. However, they may lack scalability. On the contrary, public infrastructures benefit from a large scale of servers and broad coverage of geo-location. As a result, they can provide a reliable, scalable, and secure infrastructure for website hosting. In online infrastructure analysis, Content Delivery Network (CDN) plays an essential role in delivering the website content to clients at long distances, enabling fetching the cached content from the closet edge servers in a CDN rather than downloading the content from the origin servers.

Due to its importance, many studies analyzed the CDN performance. For instance, Saverimoutou *et al.* [129] highlighted the effects of CDN choice and configuration on web browsing quality. They conducted a 12-month measurement campaign on Alexa's top-10,000 websites and discovered that CDNs reduced the average page load time by 43.1% for *HTTP2* and 38.5% for *QUIC*. Furthermore, Mangili *et al.* [94] analyzed the overtime performance of CDN, and they concluded that the CDNs largely reduced the overall traffic exchanged between network nodes. In addition, CDN organizations provide various security and hosting options. Gillman *et al.* [58] provided an overview of the CDNs protection options for hosted websites. Further, Liang *et al.* [85] observed the usage of insecure back-end communication and widespread use of invalid certificates within the CDNs protocol configuration. They further proposed a lightweight solution to enhance the ability of CDNs to protect websites.

This Work. We investigate the HTTP header and CDN infrastructural-level differences between free and premium content websites. Our findings show that the major differences between these two groups are security-related, with premium websites emphasizing more on user privacy and security, in comparison with free content websites.

Privacy Policy Annotation and Analysis

Privacy Policy Historical Regulations. The current privacy regulations for privacy policies date

back to the infancy of the Internet. The initial steps for introducing privacy laws have been motivated by the fact that technological advances will significantly impact human rights, and society as a whole. As a result, one of the earliest actions was taken by the Council of Europe by recognizing the new threats introduced by advances in computing systems, in 1968 [140]. Subsequently, the Organisation for Economic Co-operation and Development (OECD), concerned by the growing trends in data leaving their jurisdiction by traveling out of the borders of member countries, has advocated for stricter laws requiring conventions for the protection of personal data taking into account the growing automatic processing capabilities. This OECD effort has resulted in Convention 108—the Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, which was introduced in 1981 [140]. To address the mounting concerns around the fairness, accuracy, and privacy of the private information collected about consumers concerning their credit, the Fair Credit Reporting Act (FCRA) was introduced in the United States in the late 1960s, and became a law in 1970. The law allowed consumers to review their credit files and correct errors, if they exist. In subsequent years, the United States Department of Health and Human Services drafted a policy named the Fair Information Practices (FIPs) in 1973 [140]. The committee after the FIPs also led structuring the Privacy Act in 1974 [140]. Today, FIPs outline various principles for the collection and use of private data. Those principles outline guidelines concerning the collection limitation, data quality, collection purpose, use limitation, security safeguarding, the openness of the collection, individuals participation, and accountability.

General Data Protection Regulation. In 2018, the European Union's (EU) General Data Protection Regulation (GDPR), which was proposed in 2016, was introduced into a law [87, 140]. GDPR enforced businesses to reveal how consumers' data is handled and established requirements for using and sharing the collected data. Although GDPR is enforced in the EU, it still imposed fines for all misconduct whether the business is in the EU territory or not.

Such a broad enforcement of GDPR put other countries in motion as well, where major businesses have updated their privacy policies to address the GDPR law. For example, after the introduction of

GDPR to mitigate any legal consequences, the word count of the privacy policy for Instagram has increased from 2,981 to 4,221, and from 2,881 to 5,617 in case of Wikipedia. However, this word count increase was not the only added complexity to the privacy policies: the language introduced to address aspects of GDPR have been "protectionist" of the businesses in the first place, resulting in an increased reading level, where high level of expertise is needed to read and comprehend the meaning of those privacy policies.

While the main purpose of GDPR and its enforcement is to simplify the process of data management for the consumers, recent studies demonstrated just the opposite: the significant increase in the complexity of policies as measured by the reading level and word count has made it more challenging for the ordinary users [139, 156].

Privacy Policy Annotation. Service providers are required to prepare a legal document, a privacy policy, revealing how they collect, use, disclose, and manage data from their clients in accordance with the various privacy laws [105, 111]. Particularly, the PII of users that is covered by the privacy law [2] is governed by the same law for their protection. The definition of PII is broad, and includes everything that can be used to infer an individual's identity, such as name, address, date of birth, marital status, contact information, medical history, travel itinerary, or even intentions to acquire goods and services (cookies, browser history, etc.) [2]. The protections provided to users and to their PII differ significantly from a country to another, as the governing laws change. However, those protections have been the subject of an intense recent public debate in light of the disclosure of various recent security breaches, data monetization practices, and user manipulation events (e.g., the 2016 US presidential election [71, 124], and the Cambridge Analytica scandal [33, 39, 99]).

Privacy Policy 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. 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

Table 2.1: Summary of the related work. The table shows the best performing method of each study. All datasets are manually annotated, however, the OPP-115 dataset is annotated by expert law students, and therefore is used in this study. TLDR in the table below is our work, which we use to study the privacy policies of free content websites.

Reference	**Year**	**Representation**	**Annotation Level**	**Learning**	**Dataset**	F_1 **score**
Ammar *et al.* [17]	2012	n-grams	Word	LR	57 policies	0.77
Constante *et al.* [42]	2012	POS tag	Word	-	12 policies	0.83
Constante *et al.* [43]	2012	Bag-of-Words	Document	Ridge	64 policies	0.90
Zimmect *et al.* [166]	2014	TF-IDF + bi-gram	Document	Naive Bayes	50 policies	0.90
Wilson *et al.* [158]	2016	Paragraph2Vec	Segment	SVM	OPP-115	0.66
Harkous *et al.* [63]	2018	Custom (fastText)	Segment	CNN	OPP-115	0.83
Liu *et al.* [90] (1)	2018	TF-IDF	Sentence	LR	OPP-115	0.66
Liu *et al.* [90] (2)	2018	TF-IDF	Segment	SVM	OPP-115	0.78
TLDR (1)	2021	WordPiece	Segment	BERT-Segment	OPP-115	0.91
TLDR (2)	2021	WordPiece	Segment	BERT-Document	OPP-115	0.91

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) [40] framework is one of the earliest works presented to verify the privacy policies of web-based services. Additionally, the Platform for Privacy Preferences (P3P) [44] 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. Table 2.1 summarizes the efforts in this direction, compared with our work in terms of used techniques, annotation level, dataset, and performance (measured by the F_1 score).

Ammar*et al.* [17] 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.* [42] 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.* [166] presented a browser extension that retrieves the analyses of policy by utilizing NLP techniques applied over a repository of policies. Other studies [18, 162] analyzed the manually annotated privacy policies and discovered significant inconsistencies in data collecting and sharing policies. Harkous *et al.* [63] 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.* [158] 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 [78] 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.* [90] used the OPP-115 dataset with new embedding, classifiers, and classification granularity. Unlike Wilson *et al.* [158], Liu *et al.* [90] 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.* [89] 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 [79] to create a lexicon for each category based on expert-defined mappings between subject models and categories.

While the literature highlights the potential of the directions and techniques, it falls short by not delivering high accuracies on accepted benchmarks.

This Work. We investigated the technical gap in the literature by employing numerous text representations and machine learning techniques from the previous studies. With an accurate ensemble, we retrieve the data collection and privacy practices of a website, and automatically select the paragraphs highlighting the topic of interest. Then, we applied the TLDR pipeline 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.

CHAPTER 3: NO FREE LUNCH: MEASURING AND MODELING THE FREE CONTENT WEBSITES IN THE WILD

The lax level of expectations for functional and security qualities, the user-driven content, and the extensive utilization of third-party advertisements on free content platforms introduce various risks. For example, advertisements on these websites can be exploited for data and information leakage, or even the distribution and execution of malicious scripts on the user device [84, 21]. Moreover, the lack of strict maintenance operation rules in free content websites allows for various risks: web frameworks used in free content websites are rarely updated, allowing for the exploitation of old unpatched vulnerabilities and exposing their users to various levels of risk.

However, untested hypotheses and widely unverified beliefs aside, are free content websites different from premium websites delivering the same type of content? Do free content websites differ in their structure, content, and security properties from premium websites? Do these websites come with a hidden cost to users, outweighing the perceived benefits, i.e., being free? To answer these questions, we proceed with a systematic analysis of a carefully assembled dataset that curates 834 free content websites and 728 premium websites. Our study combines both domain- and content-level analysis, coupled with security analysis across various dimensions. For the domain-level analysis, we examine the domain name system features, creation time, and SSL (Secure Sockets Layer) features as measures of intent. For the content-level analysis, we examine the HTTP (Hypertext Transfer Protocol) request, page size, loading time, content type, which all are measures of website complexity. For security analysis, we examine both the website- and component-level detection and vulnerability using two major off-the-shelf tools, *VirusTotal* [149] and *Sucuri* [143].

Our analysis concludes that there are significant, fundamental, and intrinsic differences between free content and premium websites delivering the same type of content. Among other interesting findings, we report that free content websites are exclusively vastly distributed across TLDs (Top-level Domains), although using common SLDs (Second-level Domains). Moreover, they

frequently change their domains, likely to evade blacklisting, and are more often associated with invalid SSL certificates. Content-wise, free content websites tend to require significantly fewer HTTP requests for smaller requested page sizes, although at a penalty of significant load time due to extensively employing redirection with more script objects. Risk-wise, we found that free content websites are 19 and 2.64 times more likely to be malicious than premium websites at the page-level (38% vs. 2%) and the file-level (45% vs. 17%), respectively.

We leverage our insight from those analyses to generalize and extrapolate through modeling the risk of free content websites. We define risk using pure performance metrics, and we were able to group the risky websites with very high accuracy (more than 86%).

Summary of Completed Work

This paper delivers an in-depth analysis of the differences and similarities between free content and premium websites of the same content types across various dimensions: domains, content, and security. Enabled by a feature-rich analysis, we build a machine learning-based approach to score the risk of free content websites with high accuracy.

1. **Free Content Websites Curation.** We assembled a list of more than 1,500 free content and premium websites offering the same type of content. The websites are obtained from the top search results of Google, DuckDuckGo, and Bing search engines. The websites are then crawled to obtain their content, including scripts, images, HTML (HyperText Markup Language), CSS (Cascading Style Sheets), etc.

2. **Domain-level Analysis.** To examine the domain-level features of free content websites, we analyze three aspects: their TLD (Top-level Domain), SSL certificates, and creation date. As a result, we found a significant increase in the number of free content websites, in contrast to a decrease in the number of newly created premium websites. Moreover, we observe more frequent domain name dynamics in free content websites than in premium websites, and almost

one-third of the free content websites operated using an invalid or unmatched SSL certificate.

3. **Content-level Analysis.** To examine the content-level features, we analyze three aspects: the HTTP requests, page size, and average load time. Among other findings, we observe that the premium websites contain significantly more images, and their size, on average, is three times the size of free content websites. Interestingly, however, we found the load time appears comparable due to various intrinsic design choices, including the utilization of scripts and redirection to deliver advertisements, which are more prevalent in free content websites.

4. **Free Content Websites Risk Analysis.** We leverage two popular off-the-shelf tools, *VirusTotal* and *Sucuri* to assess the security risks associated with the free content websites. Our analysis shows that free content websites are significantly more likely to be associated with maliciousness than premium websites. However, the discovery of premium websites detected as malicious is quite interesting and calls for further exploration.

5. **Risk Modeling.** Both the performance and security metrics analysis highlight significant differences between free content and premium websites. Moreover, their risk profiles are vastly different from one another. Motivated by the differences in their features, we build a simple machine learning algorithm that utilizes easy-to-obtain domain- and content-level features to predict the risk of a website. We report a promising accuracy of 86.81% for modeling the risk of free content websites.

Dataset Overview

In the following, we highlight the approach we followed in creating our dataset, including initial selection and associated criteria, manual annotation, crawling, and augmentation.

Websites Selection. We compiled a list of 1,562 free content and premium websites that we use as our key instrument for conducting our measurements. In selecting the websites, various constraints for representation. In particular, the following criteria are utilized in selecting our websites:

Table 3.1: An overview of the collected dataset. The collected URLs are associated with five different categories, and belong to free content and premium websites. Overall, 1,562 websites were crawled for the purpose of this study.

Category	Free Content Websites			Premium Websites		
	URLs	Files	Avg. Files	URLs	Files	Avg. Files
Books	154	7,073	45.93	195	17,840	91.49
Games	80	6,439	80.49	113	11,314	100.12
Movies	331	9,821	29.67	152	10,738	70.64
Music	83	6,059	73.00	86	7,225	84.01
Software	186	11,561	62.16	182	18,742	102.98
Overall	834	40,953	49.10	728	65,859	90.47

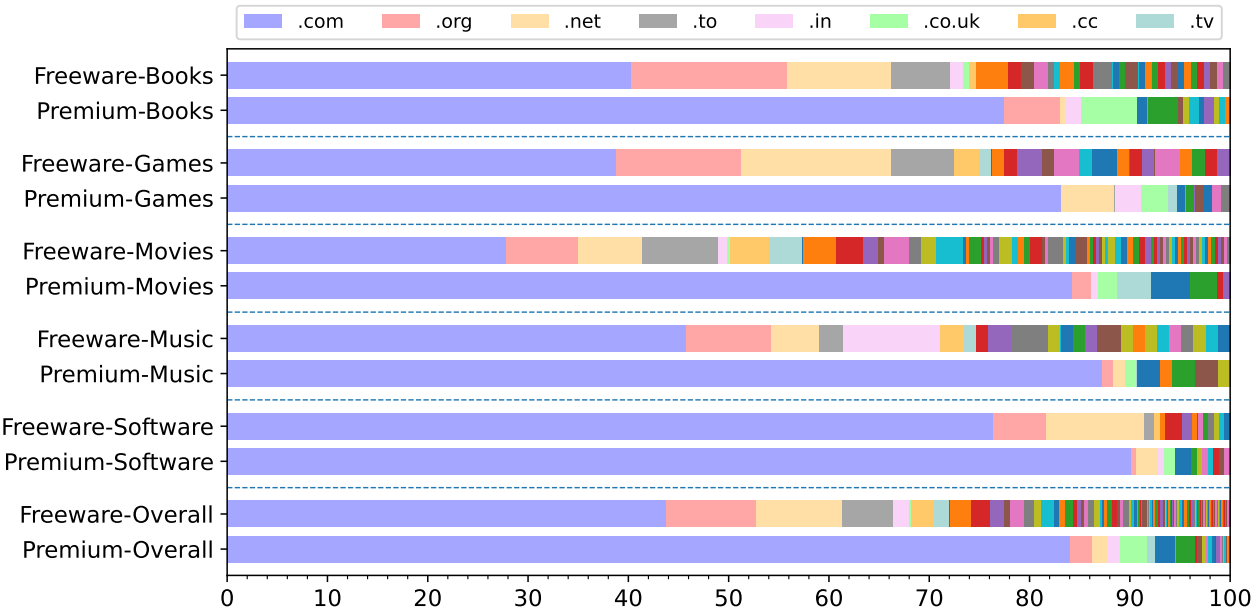


Figure 3.1: TLD distribution of free vs. premium websites. The free content websites are more distributed among the TLDs, in contrast to premium.

1. **Popularity:** Each website has to be among the most popular websites on the web. Given that those websites may not necessarily in the most popular websites, we use search engines' results as a proxy for estimating their popularity. A website is considered popular if it appears in the top results by at least one of the used search engines: Google, DuckDuckGo, and Bing.

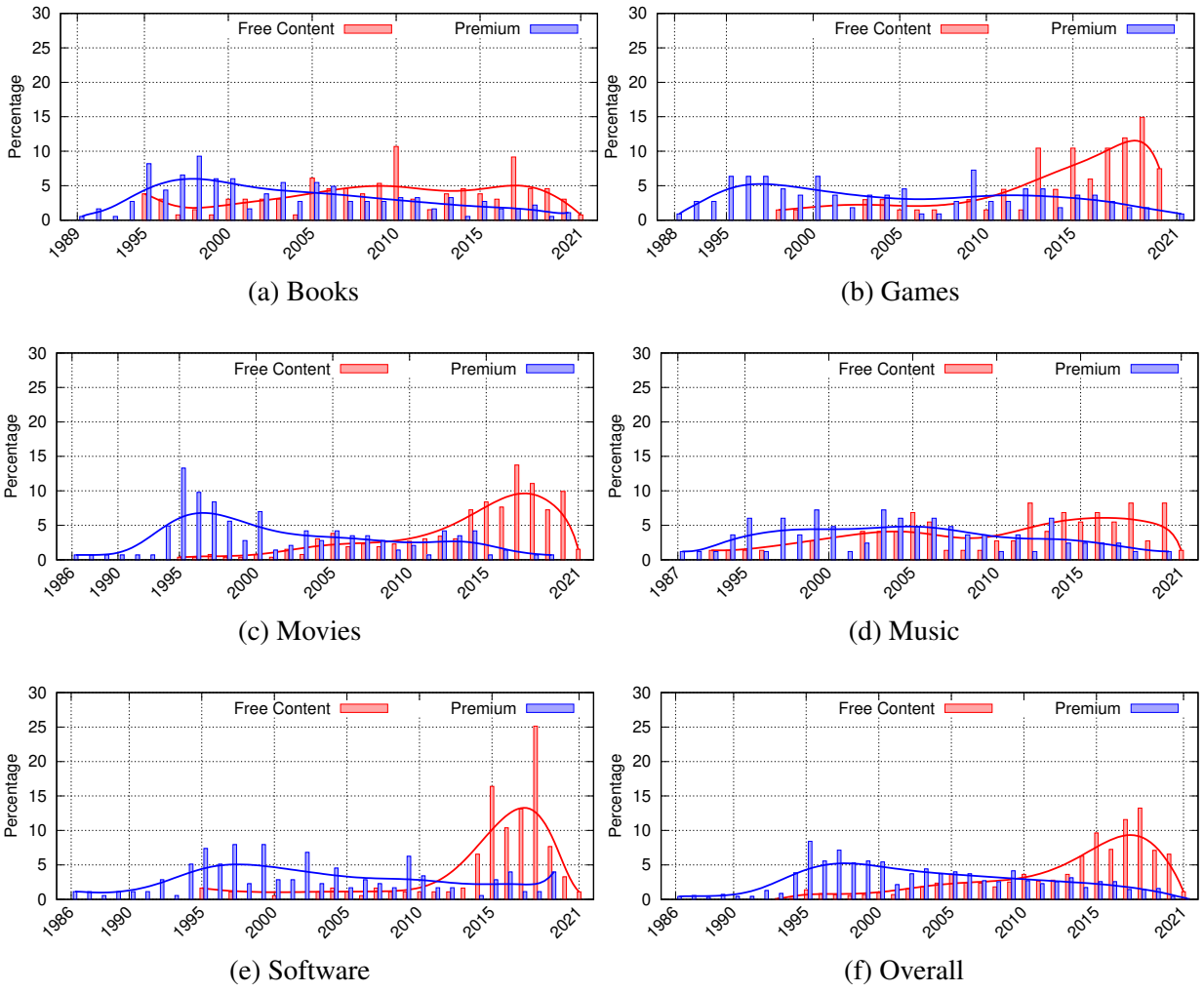


Figure 3.2: The domain creation year comparison between free content and premium website. By comparing the trend across the various content types, we observe the significant upwards trend of free content domain creation compared to premium websites.

2. **Balanced Representation:** In composing our overall dataset, we ensure that our dataset is balanced per category. To that end, we expand our queries until we achieve close-to-balanced representation across categories for both the free content and premium websites.

We seed the search with various search queries from which we obtain our candidate websites to include in the study upon manual analysis. Terms used for querying the search engines include combinations of "*free*", "*books*", "*ebooks*", "*download*", "*games*", "*movies*", "*music*", "*soft-*

ware", "*premium*", "*buy*", "*shop*", "*store*" etc. Example queries include: "*free books*", "*free online games*", "*watch movies free*", "*listen to music for free*", "*buy books online*", "*online game store*", "*shop ebooks online*", etc. Upon retrieving the top 100 query results for each query example, we filtered out the redundant results.

Upon initially selecting the unique websites for inclusion in our dataset, we proceed by manually examining and labeling each of them as either premium or free content websites. Each of the websites is then categorized, also manually, into one of five groups based on the type of content the website mainly provides: books, games, movies, music, or software.

Websites Crawling. To understand the risks associated with free content websites, we crawled each website's content (i.e., files) using PyWebCopy [117], a python package for cloning websites and downloading their associated files. The obtained files are then used for the risk analysis and modeling, as they reflect the behavior of the provided services. Our dataset is then augmented with various attributes categorized into two broad groups, the domain-level attributes (TLD, domain creation information, SSL certificate information) and content-level attributes (HTTP request information, page size, load time, and content type).

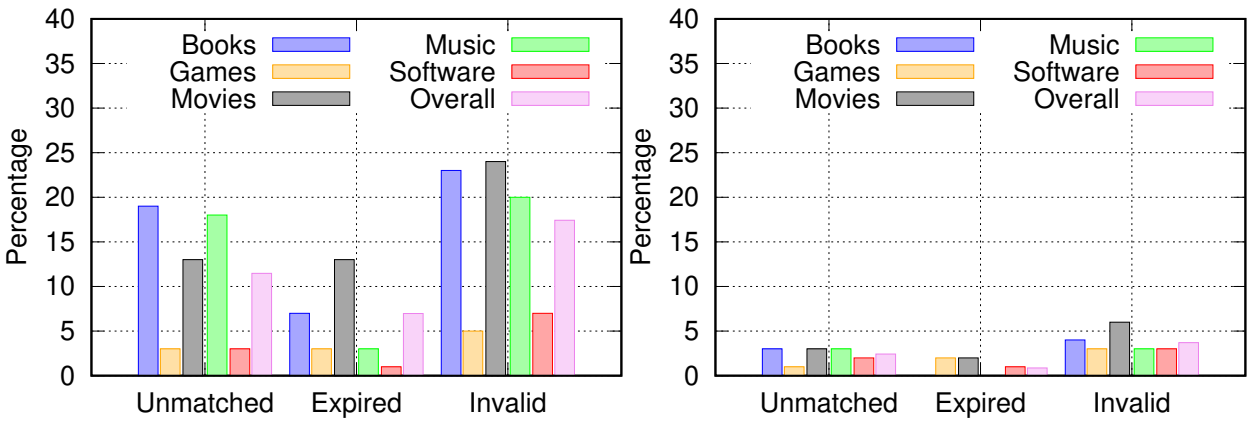
High-level Characteristics. Table 3.1 shows the distribution of the collected dataset. Notice that the average files crawled from premium websites are significantly larger than the average files for free content websites.

Shortcomings. While we do not use any explicit popularity list of websites, e.g., popular Alexa list of the one million domains, for our free and premium contents websites selection, and use the search engine results as a proxy for estimating the popularity, we note that this "proxy method" is an implicit popularity list, depending on the used terms for querying it. As various issues pertaining to the representation of characteristics [130] and bias due to external manipulation (e.g., search engine optimization and poisoning [82, 92, 153] or search personalization [73]) might skew our dataset and findings. However, we believe that most of those issues are equally likely to happen in both types of websites, making the findings rather meaningful as a mean for comparison.

In order to understand the fundamental differences between free content and premium websites, we conduct two types of analyses: domain-level analysis and content-level analysis. Domains are the gateways to websites, and they are rich in information that can be utilized to understand their intent. Supplementing the domain-level features with content-level features improves the visibility into the websites intent. In the following, we provide our analysis results based on both of those features groups.

Domain-level Analyses. The domain-level analysis provides a view of the website as an infrastructure across owner information, creation date, and the used TLD. We pursue such an analysis to contrast the free content and premium websites based on their associated domains.

Top-level Domains Analysis. The TLD is one of the highest level domains in the hierarchical domain name system, followed by the SLD (Second-level Domain); in `example.com`, `example` is the SLD, and `com` is the TLD. Recently, the number of TLDs has grown significantly with the introduction of the new generic TLDs (gTLDs), although *'.com'*, *'.net'*, *'.org'*, and *'.edu'* remain the most prominent [142]. In this work, we investigate the distribution of free content and premium websites among the TLDs, shown in Fig. 3.1. We found that *'.com'* is the most prominent TLD domain, with 44% and 84% of free content and premium websites using *'.com'*, respectively. However, interesting, we found that the total number of unique TLDs used by the premium websites in our dataset to be only 24, while this number is 98 domains in the free content websites. We note that this widespread distribution could be triggered by the mechanisms employed for malicious website blocking implemented by major browsers and systems. For instance, Chrome and Firefox rely on user reports when using safe browsing service [60] to collect and block malicious websites. To evade blocking, free content websites change their domain name periodically. However, free content operators maintain the same SLD and migrate their websites to other TLDs to retain the existing users. We observe that some free content websites use more than one TLD to evade blocking.



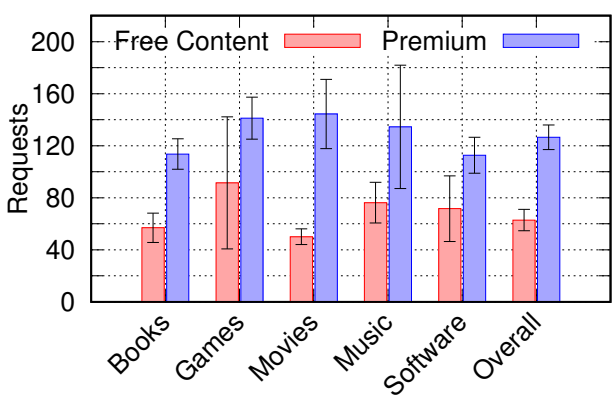
(a) Free content websites

(b) Premium websites

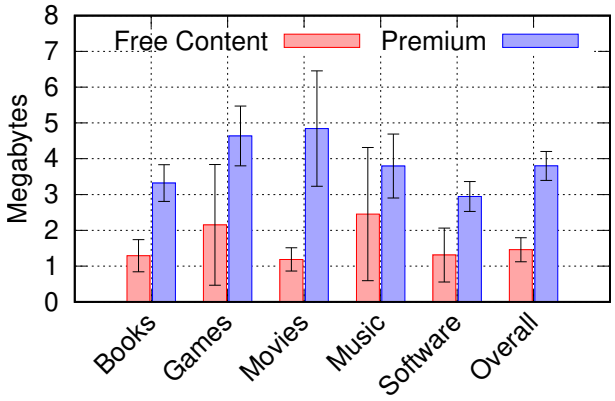
Figure 3.3: The SSL certificate analysis results. We observe that almost 36% (sum of the pink bars) of the free content websites have problematic SSL certificates (unmatched, expired, or invalid) compared to 7% in premium websites.

Domain Name Creation. We examine the website creation dates, where we observe an increasing trend in the number of newly created free content websites, in contrast to the declining number of newly created premium websites, as shown in Fig. 3.2. This growing trend, particularly in the period of 2015–2021, motivates us to examine and understand the risks associated with using online free content websites. To further support that, we found from the TLDs analysis that free content websites tend to change their domain name periodically to avoid content blocking or blacklisting.

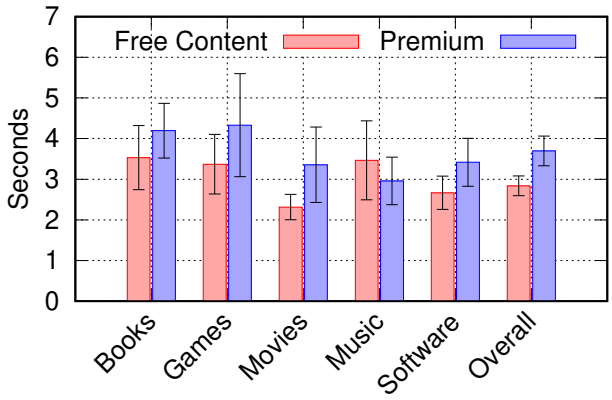
SSL Certificate Analysis. HTTP is responsible for transferring a website content, e.g., HTML, from the web server to the user browser. However, this protocol is not secure, and the transferred data can be exposed to unauthorized access. Therefore, most websites have moved to use the secure version of HTTP (HTTPS), which implements an encryption mechanism to protect the transferred content. From our analysis, we found that 36% of the free content websites have invalid HTTPS compared to only 7% of the premium websites. Moreover, we found that 26% of free content websites still allow HTTP (insecure) access, whereas 0% of the premium websites allow HTTP access. SSL certificate is a digital authentication method that authenticates the identity of a website and provides the HTTPS with an encrypted connection. The SSL certificate is a critical component



(a) The average HTTP requests per page.



(b) The average page size in MB.



(c) The average page load time in seconds.

Figure 3.4: Page-related comparison between the free content and premium websites. Despite having different page sizes, the free content and premium websites average comparable page load times, indicating other reasons than size that affect time.

of a website to secure user data and protect them against, e.g., phishing.

In this work, we investigate the validity of the SSL certificate for both free content and premium websites. In particular, we study three aspects: (i) unmatched hostname in the certificate, (ii) expired certificate, and (iii) invalid/fabricated certificate. Fig. 3.3 shows that, in total, 36% of the free content websites have issues with their certificates (i.e., 11.5% unmatched name, 7% expired, and 17.5% invalid certificate), compared to a total of only 7% of the premium websites with problems in their associated SSL certificates. This is more noticeable in the *"Movies"*, *"Books"*, and *"Music"* categories. As shown, the free content websites are more likely to have issues with their

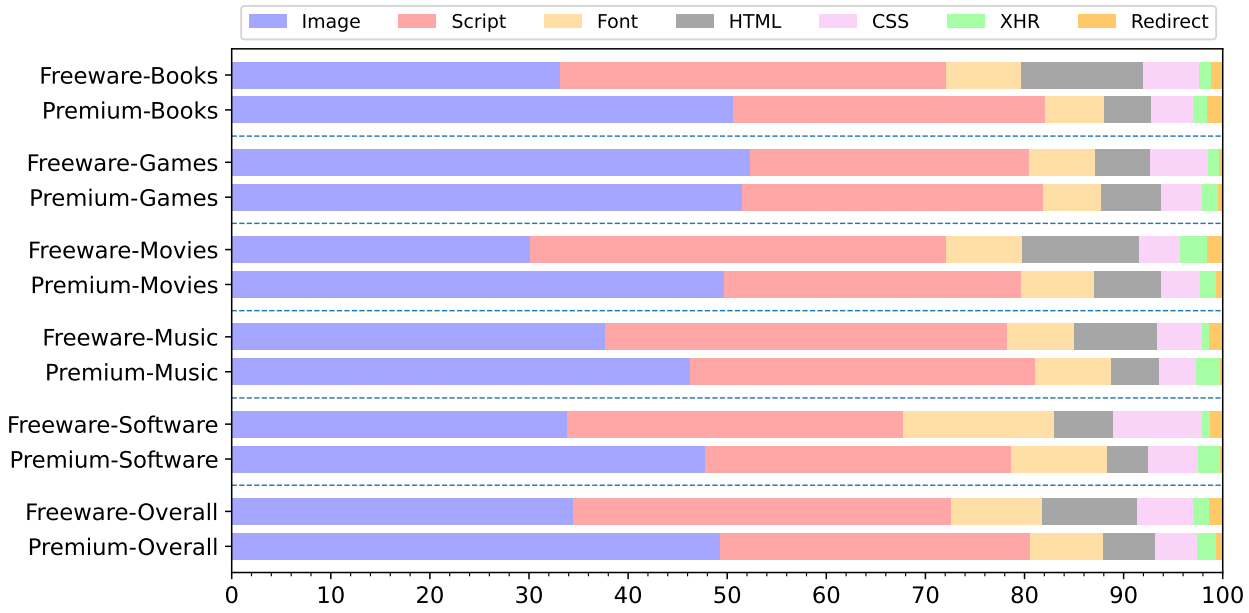


Figure 3.5: Content-type comparison between the free content and the premium websites. We observe some differences in the website file types, notably in the *images* type.

SSL certificate. This may be contributed to that free content operators are not renewing the SSL certificate, unwilling to increase their operational cost. Nonetheless, this practice leads to potential risks regarding user information and data privacy.

Key Takeaways. Through domain-level analyses, we found that (i) the free content websites are newer, and their growth has been increasing significantly in recent years, whereas the premium websites' growth is decreasing, with fewer websites introduced every year, (ii) the free content websites are more distributed across the TLDs as they change their domain to avoid malicious website blocking mechanisms, (iii) the free content websites are more likely to have invalid or expired SSL certificate. These findings complement our analysis concerning the safety of using free content websites and the risks associated with them.

Content-level Analyses. In order to gain insight into the content-level features of free content and premium websites, we analyze the extracted files in both types of websites. In this analysis, we focus on four features: number of HTTP requests, page size, page load time, and content type.

HTTP Requests. HTTP requests are made by clients to request access to resources on servers (e.g., HTML files, CSS, images), and their numbers per page are an indication of the complexity of the requested page. Fig. 3.4a shows the average number of HTTP requests made for free content and premium websites. We observe that a client would initiate almost twice the number of requests to access a premium website compared to accessing a free content website. This is quite anticipated, given that the premium websites pages are larger in size. However, we observe that the average page size in premium websites is 3x the free content websites, whereas the number of the HTTP requests is only 2x more, indicating that visiting a free content page requires more HTTP requests for the same amount of data. That could be a result of redirection, where each redirection triggers one or more independent HTTP requests and consumes more time for loading. Notice that in Fig. 3.4a, four out of the five categories have significant statistical differences between free and premium content websites. This is more prevalent when considering all websites. This indicated that the average HTTP request per website page is indeed a feature that can be leveraged to distinguish between free and premium content websites.

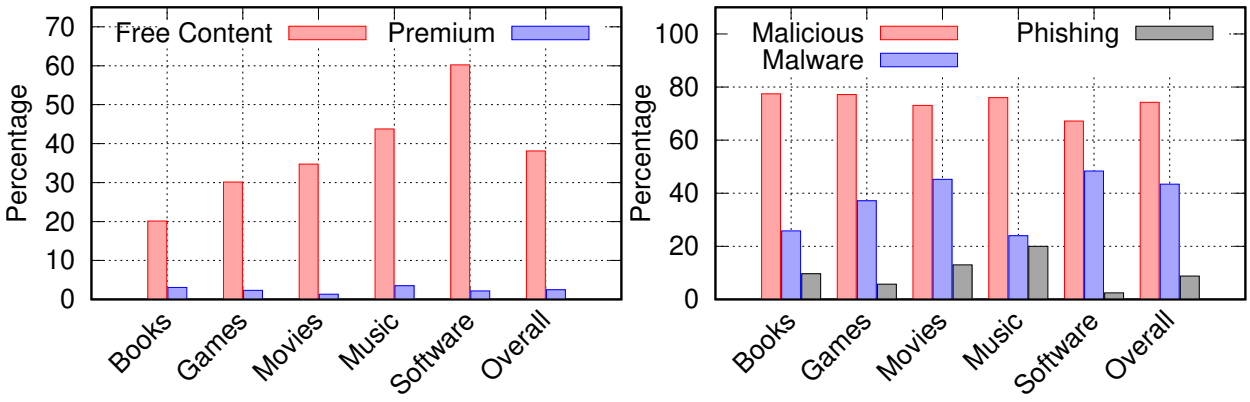
Average page size. According to the page weight report by HTTP Archive [20], the average page size of the top one million websites is around 2.07 MB. How far is the size of the average page that belongs to either category? To answer this question, we examine the average page size of the free content and premium websites, with the results reported in Fig. 3.4b. We observe that the free content websites follow the normal distribution of the page sizes reported by the HTTP Achieve [20], while the premium websites have an average homepage size of 3.9MB, three times the average size of a free content page. A potential explanation might be that the free content websites rely on redirecting users to other websites content or advertisement websites, as we demonstrate later, instead of including and presenting content in the free content page body. Similar to our previous observation, Fig. 3.4b shows significant statistical differences between free and premium websites for all the categories, except for “Music”.

Average page load time. We define the page load time as the time it takes the page to be loaded

fully and measured to understand additional aspects of websites complexity. Fig. 3.4c shows the average page load time, calculated using the SolarWinds Pingdom API (Application Programming Interface) [115], for both the free content and premium websites. While the average size of the premium websites is three times the average free content page size, we notice that the average load time is comparable across them, indicating aspects beyond the size that affect the load time, i.e., degraded performance and extensive usage of redirection. We note that for the average page load time, the significant statistical difference is much lower and overlapping for free and premium content websites. This indicates that the average page load time can not be an independent feature for distinguishing between free and premium content websites.

Content type. The page size does not seem to fully explain the complexity and loading time of websites, which calls for a deeper analysis of the websites content. The content type is another statistical feature about the website's content at the component level (i.e., files). These components include *Image* (*GIF*, *PNG*, *JPEG*), *JavaScript*, *Text*, *HTML*, *CSS*, *XHR*, and *Redirection*. We found that *Image* is the most common component, followed by *JavaScript*, whereas the *Redirection* content is the least common among these components. Fig. 3.5 shows the average distribution (%) of the different components in the free content and premium websites. Overall, the premium websites have 15% more images than free content websites. However, we notice the extensive usage of *Redirection* in the free content websites, as it is often a method to deliver advertisements and mislead the filtering algorithm. We found that the (rounded) ratio of the redirection in free content compared to premium pages to be 6 (software), 7 (music), 3 (movie), 1 (games), 1 (books). Overall, free content websites redirect twice as much as premium sites, have twice the HTML, 1.5 times the CSS, and 1.23 times the JavaScript.

Key Takeaways. Our content-level analyses shed light on the main differences between free content and premium websites. We found the following. (i) The premium websites have almost twice the number of requests as free content websites and three times the average size of free content websites pages. (ii) Nevertheless, the average homepage load time is comparable for free content



(a) Websites labeled as malicious.

(b) Malicious free content websites.

Figure 3.6: The potential maliciousness of free content and premium websites.

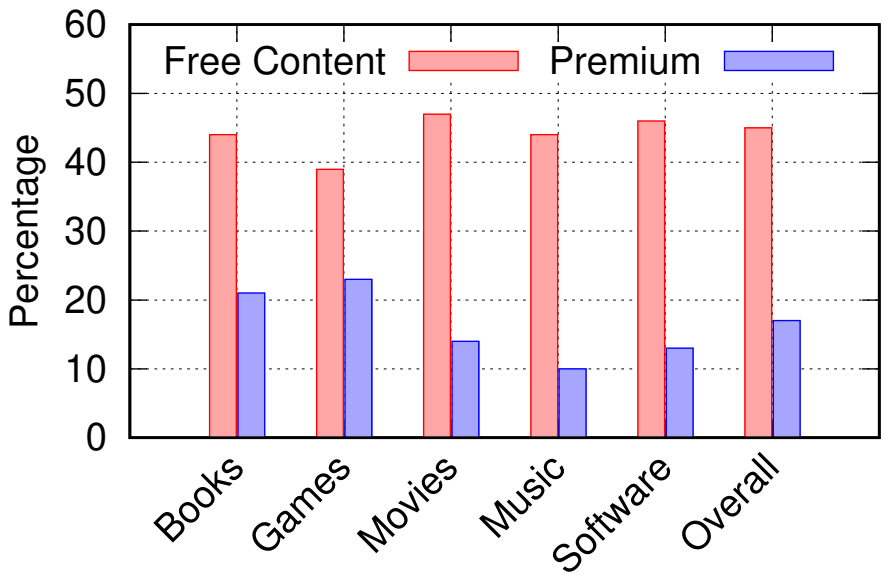


Figure 3.7: The malicious files detected by *VirusTotal*: free content vs. premium.

and premium websites. (iii) Content type-wise, the free content websites have a higher portion of *redirection* components, as they are a primary method to deliver advertisements.

The analysis we conducted so far considered the performance and non-security characteristics of free content and premium websites, which highlight clear differences that contribute to both direct and indirect costs. One of the most important and obvious metrics to measure the cost of free content websites is by understanding their security and associated risk. In this section, we conduct this analysis, focusing on indicators of threat, such as maliciousness of URLs, files, and associated vulnerabilities. Towards automating the discovery of such risks, we also report the results of a machine learning-based tool that shows the risk boundaries of websites based on features obtained from the risk analysis.

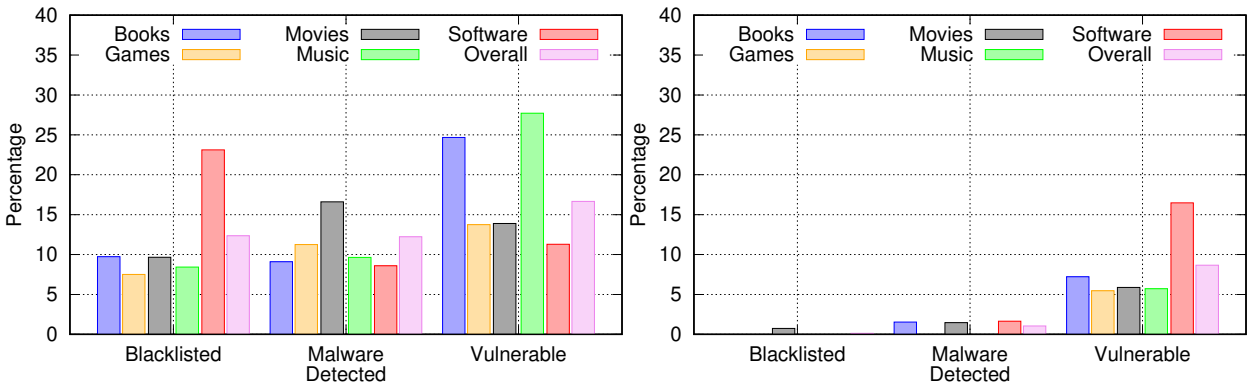
Risk Assessment. The study of the maliciousness and vulnerabilities of both services websites, by shedding examining how they potentially affect users experience, safety, and security, is important. Motivated by that, we define the risk of a website using several metrics, namely: (i) containing malware, (ii) running malicious scripts, (iii) exploiting user device's resources, or (iv) containing vulnerabilities, outdated software versions, or unpatched frameworks.

To assess the risk of each type of websites without reinventing the wheel, we leverage two public APIs: *VirusTotal* [149] and *Sucuri* [143] for harmful behavior analysis. *VirusTotal* is an online service that aggregates the scanning results of more than 70 scanning engines and can be used for scanning files and URLs alike. On the other hand, Sucuri is a service that tests websites against several known malware, viruses, blacklisting lists, vulnerabilities, outdated frameworks, and malicious code.

Malicious URLs Detection and Annotation. Using *VirusTotal API*, we extracted malicious activities associated with the website URL, shown in Fig. 3.6. We notice that there is a noticeable discrepancy between free content and premium websites in terms of maliciousness. In particular, Fig. 3.6a shows that 38% of the free content websites are considered malicious by *VirusTotal*, compared to only 2% of the premium websites. A significant number of those detected websites

Table 3.2: The distribution of malicious files for different file formats in free content and premium websites. We observe that a large portion of *".gif"* files are labeled as malicious in both cases, although almost twice as much (percentage) in free content.

	Category	.gif	.html	.png	.js	.php	.woff	.jpg	.eot	.woff2	.svg	.ttf	.log	.css
Free Content	Books	28%	1%	0%	2%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Games	7%	13%	0%	1%	3%	0%	0%	0%	0%	0%	0%	0%	0%
	Movies	40%	6%	1%	2%	0%	1%	0%	1%	0%	1%	1%	1%	0%
	Music	26%	6%	0%	1%	4%	3%	0%	3%	3%	0%	2%	0%	0%
	Software	11%	30%	4%	1%	1%	4%	0%	2%	4%	2%	3%	1%	0%
	Overall	26%	11%	2%	1%	2%	2%	0%	2%	4%	1%	2%	1%	0%
Premium	Books	19%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Games	21%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Movies	9%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Music	21%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Software	5%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
	Overall	15%	1%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%



(a) Free content websites

(b) Premium websites

Figure 3.8: Assessing the maliciousness of the free content and premium websites. We show the percentage of the websites labeled as blacklisted, malware, and vulnerable.

(≈74%) were labeled as malicious (Fig. 3.6b), a website created to promote scams, attacks, and frauds. We also notice that a significant portion of the free content URLs is detected as malicious, ranging from 20% (*"Books"* websites) to 60% (*"Software"* websites). In contrast, premium websites have a very low detection rate, ranging from 1% to 4% only.

Malicious File Detection and Formats Analysis. In order to understand the behavior of a given

service (i.e., content providers), it is essential to analyze the behavioral characteristics of the executable scripts hosted by the service. These scripts are forwarded to the end-user as files, including images, JavaScript codes, HTML, among other formats, and are often rendered or executed on the user's device. Analyzing the scripts and website files is critical, as recent studies [37, 141, 161, 70] have shown that such content can be exploited, leading to information and data leakage, in addition to abusing the resources of the end-user device. In order to understand the risks of free content websites, we leverage *VirusTotal API* for malicious file identification. In contrast, Fig. 3.7 shows the percentage of malicious files detected by the *VirusTotal API* in the free content and premium websites. While the number of URLs that pertain to the premium websites and are labeled as malicious is only 2%, the number of their files labeled as malicious was 17%.

We notice that the trend persists overall, although magnified: 45% of the free content websites had files that have been labeled as malicious (compared to 17% in premium). To better understand this observation, we investigate the distribution of the format of the malicious files, where the comparative results are shown in Table 3.2.

Based on Table 3.2, we report that the majority of malicious files have *'.gif'* and *'.html'* formats. This is a result of either (i) the *'.gif'* and *'.html'* files containing malicious embedded scripts, or (ii) the VirusTotal engines considering the *'.gif'* files as malicious content in general (i.e., potential false positives). It is worth noting that we manually inspected the *'.gif'* files, and found that the majority of the malicious-labeled *'.gif'* files are advertisement-related content.

Websites' Vulnerability and Blacklisting. In order to analyze the potential exploitable vulnerabilities and blacklisting, we leveraged Sucuri API [143] to obtain information of domains activities for both types of services. As a result, we found that 12% of the free content websites were detected as *containing malware*, compared to only 1% of their premium counterparts, as shown in Fig. 3.8. Moreover, we found the free *"Movie"* websites have the highest percentage of malware detection (16.67%), as shown in Fig. 3.8a.

We also scanned the websites for vulnerabilities and found that the free *"Books"* and *"Music"*

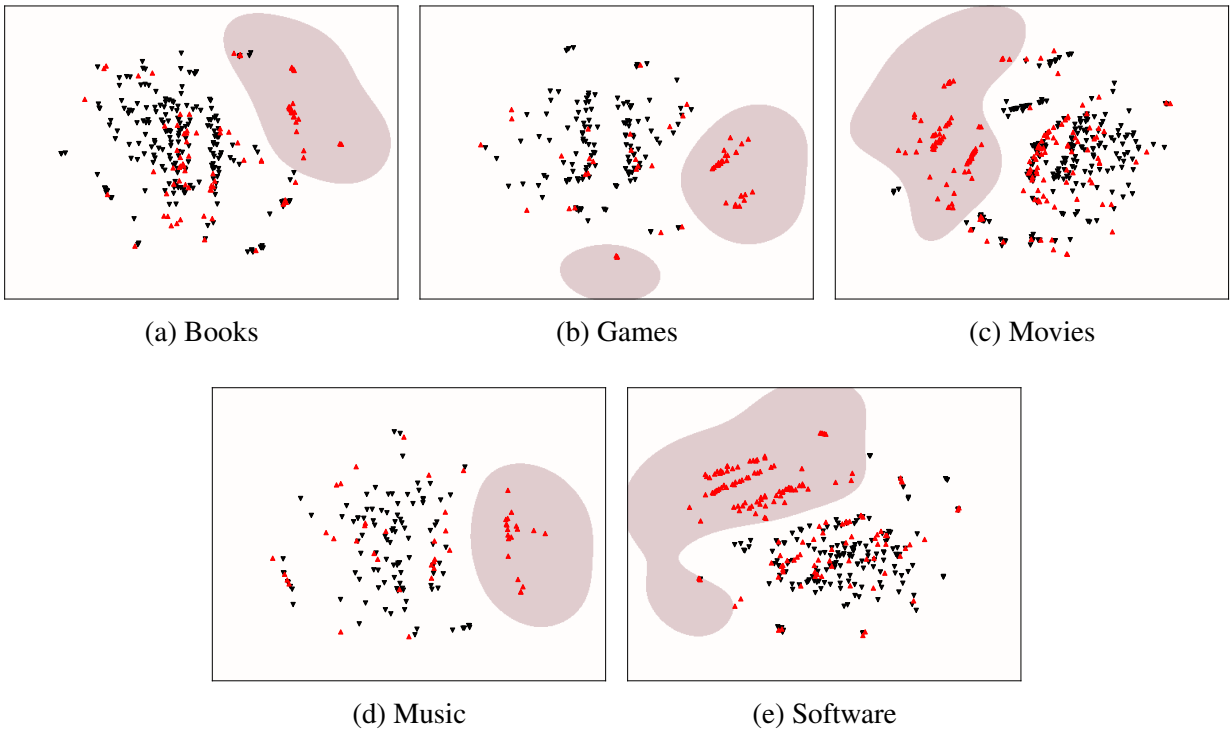
Table 3.3: The description of the website's characterization features. The features are extracted from three sources, (i) The website's content, (ii) The website's public information, (iii) The website's SSL certificate information. We include the characteristics extracted from VirusTotal and Sucuri APIs for risk characterization and potential detection. *[c]*: categorical feature, *[b]*: boolean feature (T/F), *[n]*: numerical feature, *[p]*: percentage feature.

#	Type	**Description**	#	Type	**Description**
1	[c]	TLD name used in the website URL	14	[p]	The percentage of Redirect content in the website
2	[b]	Website's name not match SSL certificate	15	[b]	Website's domain is detected by VirusToal API
3	[b]	Website contains expired SSL certificate	16	[b]	Website is detected as malicious
4	[b]	Website's SSL certificate cannot be verified	17	[b]	Website is detected as containing malware
5	[n]	Average number of HTTP requests	18	[b]	Website is detected as phishing
6	[n]	Average content size of a given website	19	[b]	Website's files detected as malicious
7	[n]	Website's web page load time	20	[b]	Website's URL detected by Sucuri API as malicious
8	[p]	Percentage of images in the website	21	[b]	Website's URL is blacklisted by Sucuri scanning engines
9	[p]	Percentage of script files in the website	22	[b]	Sucuri API discovered vulnerability within the website
10	[p]	Percentage of Fonts content in the website	23	[n]	The lifetime of the website's IP address
11	[p]	Percentage of HTML files in the website	24	[b]	Website is using/used Cloudflare as a CDN
12	[p]	Percentage of CSS files in the website	25	[b]	Website is using/used Akamai Tech as a CDN
13	[p]	Percentage of XHR content in the website			

websites have the highest vulnerabilities overall. Despite the low reporting rate in the premium websites, 17% of *"Software"* were labeled as vulnerable, a higher portion than in free content websites (12%), which is quite surprising. According to *Sucuri* reports, a high percentage of the legitimate *"Software"* websites vulnerabilities are due to outdated framework versions, which is common in *"Software"* services websites.

In terms of blacklisting, Fig. 3.8a shows that 12% of the free content websites were blacklisted by the *Sucuri* scanning engines, including Google, McAfee, Yandex, Norton, ESET, and AVAST engines. We observe that the *"Software"* free content websites have a significantly higher percentage of blacklisted URLs (23.12%) compared to other categories, which all had at most 12% blacklisting rate. One reason for this behavior is the fact that these websites are changing their domain names frequently using a different TLD.

Key Takeaways. To assess the risks associated with free content websites, we leveraged *VirusTotal* and *Sucuri* APIs for analyzing the maliciousness of domain and files of both service types. Our analyses show worrisome trends among free content websites, including (i) free content websites are more likely to be associated with maliciousness at a domain-level (38% of the free content



(a) Books (b) Games (c) Movies

(d) Music (e) Software

Figure 3.9: The decision boundary of the risk-free and risky websites. A risky website is a website with potential malicious intention. Notice that this malicious behaviour can be characterized (i.e., determined) using a support vector machine.

websites), and (ii) they are more likely to be associated with maliciousness at the file-level (45% of them). These trends are not limited to maliciousness, which led to high blacklisting, but include exploitable vulnerabilities that can expose visitors to leakage attacks. Our analysis also unveils that 17% of the free content websites are vulnerable.

Risk Modeling. The insights that we have provided thus far are intriguing, although we are left with a key question: how much of these insights can be generalized and extrapolated across sites of the same population and type for risk assessment? To answer this question, we report on our effort to identify *risky* websites using simple machine learning algorithms. To do so, we first present our definition of what constitutes a "risky" website, our features utilized for automatically determining those risky websites, and the results of a learning algorithm utilized for automatically extrapolating the risk definition using the previously defined features.

Risky Websites. A website in our analysis is considered risky if it is associated with any of the following:

1. *Malicious Domain.* Websites that are associated with URLs responsible for malicious activities are considered risky.
2. *Malicious Files.* Upon visiting a website, multiple scripts are executed on the host. As such, we consider any website with malicious files, regardless of its *VirusTotal* label, as a risky website.
3. *Blacklisted URLs.* Blacklisting can occur due to (i) massive user reporting, or (ii) previous maliciousness by the website (e.g., scam attacks). As such, we consider all blacklisted websites as risky.
4. *Vulnerable Websites.* Websites that are identified as vulnerable by *Sucuri* are considered risky, for the potential exploitability.

We note that the risk modeling is not limited to free content websites. We also consider any free content and premium website with one or more of the aforementioned aspects as a risky website and otherwise a risk-free website.

Website Features. To model the risks associated with each service, we leverage the aforementioned extracted features as a representation. In particular, Table 3.3 shows the superset of potential features that we use to represent each online service, including SSL certificate, page size, load time, TLD, and website content features. Additionally, we include three more features extracted using *SecurityTrails* [133]: (i) the lifetime of a service IP address, (ii) whether a website is using or previously used Cloudflare as a Content Delivery Network (CDN), and (iii) whether a website is using or previously used Akamai Tech as a CDN.

Hold-out Data. The data obtained by *VirusTotal* and *Sucuri* (#15–#25) in Table 3.3 is held out, and is only used to model validation. This allows us to utilize easy-to-obtain website quality metrics

that do not require access to third-party information to model the website risk. We envision that our lightweight modeling, in contrast to third-party risk data, would be more practical, since the third-party labels are determined based on reporting and expensive analyses accumulated over a period of time. Solely relying on third-party tools, such as *VirusTotal* to identify risks would exclude a significant number of websites, including those newly created for free content.

Risk Boundaries. Considering the aforementioned features, we visualize the boundaries between *risky* and *risk-free* websites, shown in Fig. 3.9. In particular, we use the t-distributed stochastic neighbor embedding (t-SNE) visualization technique [147] to plot the features of the websites. t-SNE is a technique that maps high-dimensional data into two or three dimensions best suited for visualization using nonlinear mapping. The process is initiated by measuring the similarity between each datapoint within the high-dimensional and low-dimensional object of two or three dimensions. The similarity can by represented as a conditional probability where similar objects have higher probability than dissimilar points. The conditional probability is defined as:

$$p_{j|i} = \frac{\exp\left(-\left\|x_i - x_j\right\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\left\|x_i - x_k\right\|^2 / 2\sigma_i^2\right)}$$

Where; x_i and x_j are the two datapoints to compute their similarity. $p_{j|i}$ is the conditional probability that x_i would pick x_j as its neighbor. σ_i^2 is the Gaussian variance of a given number of neighbors (perplexity). Then, using a support vector machine, we estimate the risk boundaries, shown in the red-shaded area in Fig. 3.9. Based on the validation, we find that the riskiest websites are clustered together, as they share different website features. Our modeling is capable of identifying risky websites with an accuracy of 86.81% (calculated against samples that fall out of the boundary for the class of interest), despite some limitations (e.g., potential false positives among our sampled websites).

Key Takeaways. We address the need for lightweight risk modeling of free content websites using a representation of 17 generic and file-related features. Our modeling is shown effective, with multiple advantages, and producing an accuracy of 86.81%.

Shortcomings

While curating our dataset, and to have a fair comparison between the free and premium websites, we manually examined and categorized the websites into categories. Regarding the popularity of websites, we only collected the top-ranked "most popular" search results of the most popular search engines (Google, DuckDuckGo, and Bing). This is done by observing the top returned results when searching for websites of different categories on the online search engines. We performed that using keywords such as ("free software," "watch movies for free," "download free books," etc.) We define a free content website as any website that provides content (books, games, movies, music, or software) for free. Whereas premium websites require the user to pay some amount of money to access or buy the content (including websites with any type of subscription or pay-as-you-use websites). The labeling mechanism was firstly conducted using the search engines' keywords (e.g., free). Then, a second step was performed manually by inspecting each website to accurately assign the websites into either free or premium categories, in addition to the type of the provided contents. We crawled the homepage of the websites only for our analysis.

Summary & Concluding Remarks

Free content websites are an interesting element of the makeup of the web today, and their characteristics are not rigorously analyzed nor understood in contrast to other websites that offer the same content. This paper provides the first look into a comparative analysis of such websites across various domain- and content-level dimensions, as well as their risk profiles. Our curated datasets offer valuable resources for exploring this uncharted space, and our findings shed light on the fundamental differences between free content websites in contrast to premium websites. We believe that our analysis in this paper only "scratches the surface" of this important problem and calls for further explorations and actions. For instance, our domain- and content-level analyses have been only focused on easy-to-obtain metadata features and did not consider the in-depth features, e.g., linguistic, network topology information, regional information, deep content type, and organiza-

tion attributes (e.g., in the case of SSL certificates; signing authorities, and hosting infrastructure). All of these dimensions could shed more light on the characteristics of such websites and constitute our future work. Finally, we notice that our analysis utilizes a single snapshot of those websites, and we did not consider the temporal dimension of their characteristics, which would be a very interesting yet challenging aspect to explore.

CHAPTER 4: UNDERSTANDING THE SECURITY OF FREE CONTENT WEBSITES BY ANALYZING THEIR SSL CERTIFICATES: A COMPARATIVE STUDY

There has been a recent explosion in popularity and usage of online services and web platforms that deliver content (music, movies, books, etc.). This significant growth in such platform's popularity is in part attributed to the convenience of their use [50, 45, 28, 74, 55, 118, 27, 106, 48]. Generally speaking, websites delivering such content are categorized into two groups based on their monetization options: free and premium. The free websites provide free physical or virtual services and are typically run by donations or advertisements [64, 61, 30, 138]. On the other hand, the premium services are either subscription-based or pay-as-you-use. The latter category is, in most cases, strictly mentored to ensure quality, while the free services lack high level of monitoring as they may be user-driven.

The reliance of free content websites on advertisements and user-driven content raises several concerns. For instance, advertisements can be exploited for data and information leakage, in addition to running malicious scripts on the user device [84, 6, 128]. Moreover, the lack of censorship raises security concerns regarding the provided services. For instance, in an attempt to reduce the operational cost of the online service, the service providers may relax their security and privacy requirements or may not use them altogether.

Motivated by these concerns, we explore the fundamental and structural differences between free content and premium websites. In doing this analysis, we use the Secure Sockets Layer (SSL) certificate content. The SSL certificate is a digital authentication method that proves the identity of a website and (eventually) provides an encrypted connection between the client and the server.

This work has been published at ACM 1st International Workshop on Cybersecurity and Social Sciences (CySSS '22) held in conjunction with ACM Asia Conference on Computer and Communications Security (ASIACCS), 2022.

SSL certificate is a critical element of a website to secure the users' data and protect them against mischievous phishing and skimmers.

To this end, we investigate the validity of the SSL certificate for both free and premium services. In particular, we focus on understanding the fundamental differences between free and premium content websites in three directions: (i) Errors within the SSL certificate, including unmatched client name, expired certificate, or invalid/vulnerable information and content, (ii) SSL certificate issuer organization analysis, including the most commonly used certificate providers, such as *Cloudflare Inc.*, *Let's Encrypt*, *DigiCert Inc*, and the SSL certification issuer countries distribution analysis (*e.g.* United States, United Kingdom, and Belgium), and (iii) SSL certificate signature algorithm analysis (*e.g. SHA256 with RSA*, *SHA256 with ECDSA*, and *SHA1 with RSA*).

Understanding the different characteristics of the SSL certificate is crucial for user risk exposure analysis. The most common issues within the SSL certificates are (i) **Untrusted SSL Certificate:** The certificate is not signed by a trusted certificate authority. The website, in this case, publishes a certificate self-signed by the server. (ii) **Domain Name Mismatch:** This happens when the website's URL is different from the domain name in the SSL certificate, which indicates either illegal use of the certificate or inconsistent domain change. (iii) **Mixed content warning:** This warning is issued when elements among the website content are unsecured, indicating that either such elements are malicious or can be exploited. (iv) **SSL Certificate Expired:** Expired SSL certificate may result in out-of-date security practices, causing further exploitation.

The aforementioned reasons motivate for understanding the differences between free and premium services and websites. Toward this goal, our analyses uncover that the two categories are indeed distinguishable, each with shared behavior among its services. Our experimental evaluation shows that 35.85% of the free websites' certificates have significant issues, with up to 17% invalid SSL certificates due to unsecure content and 12% with mismatched domain names. Surprisingly, we uncover the usage of the emerging ECDSA encryption algorithm among the free websites, a faster and more secure option in comparison with the more popular option of RSA-2048 used along with

SHA-256 in premium websites.

Summary of Completed Work

Starting with a list of 1,562 free and premium services websites obtained from the top results of Google, DuckDuckGo, and Bing search engines, we extract and analyze the SSL certificates toward assessing potential exploitation and risks, across the following verticals.

1. **SSL Certificate Validity Analysis.** We analyze the SSL certificates, extracting existing issues that expose the user to vulnerabilities. We uncover that, on the fundamental level, the free and premium websites are highly distinguishable, with free websites certificates being labeled as 17% invalid, 7% expired, and 12% with mismatched domain names, a ratio that is much higher than its premium counterpart.

2. **SSL Certificate Issuer Analysis.** We analyze the SSL certificate issuing organizations, unveiling that, at the country-level, the SSL certificate issuing organizations are very similar. However, we observed the heavy usage of "Cloudflare" among the free websites (38.22% of the websites), in comparison with only 15.88% among the premium counterparts.

3. **SSL Certificate Signature Analysis.** We study the utilized signature encryption algorithms of the SSL certificates and uncover that on the data encryption level, the premium websites are using RSA with larger public key sizes, in comparison with the emergence of more secure ECDSA algorithm among the free websites (with the used key parameters).

Dataset Overview

We compiled a list of 1,562 free content (834) and premium (728) websites for our analyses. When selecting the websites, we considered the following factors: (i) selecting the most popular websites, *e.g.* websites that appear in the top results by Google, DuckDuckGo, and Bing search engines, and

Table 4.1: An overview of the collected dataset. The collected URLs are associated with five different categories, and belong to free content and premium websites. Overall, 1,562 websites were analyzed for the purpose of this study.

# URLs	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

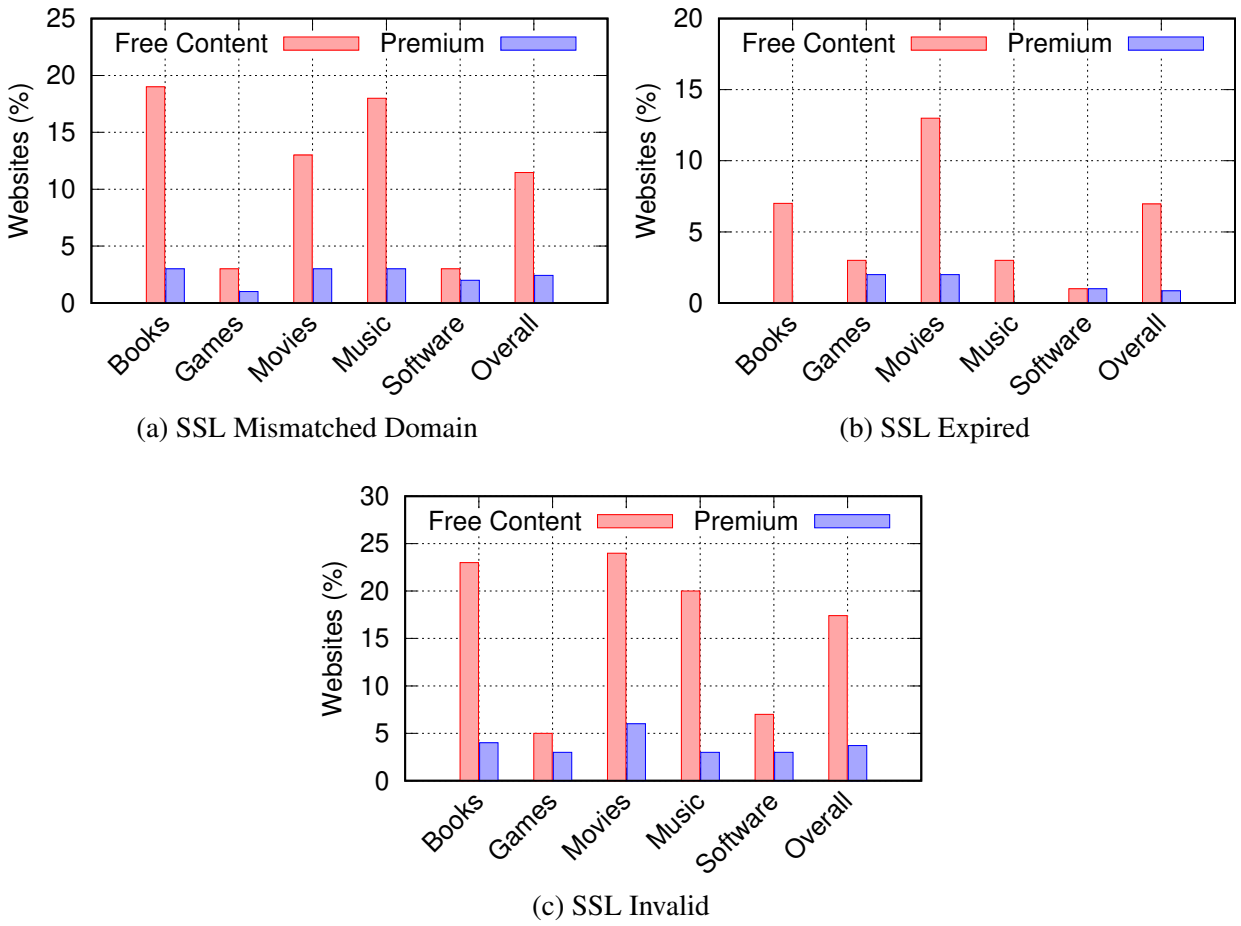
(ii) maintaining a balanced dataset. In addition, we verified and labeled each website manually. The compiled websites are then categorized into five groups based on the provided content: books, games, movies, music, or software. Table 4.1 shows the distribution of the dataset.

Our dataset is then augmented with various SSL certificate attributes, including SSL certificate validation, issuer organization, issuing country, and signature algorithm. Such information was retrieved using APIVoid [19], a framework that provides cyber threat detection and analysis, and OpenSSL [110], a command-line tool to retrieve SSL certificates chains from target websites and parse them to a readable format. We analyze whether the websites' associated SSL certificates are expired, invalid, or unmatched (with respect to the domain name they are used for). In particular, we focus on information that reflects users' exposure to risk, including:

1. Unmatched hostname in the certificate.
2. Expired certificate.
3. Invalid/fabricated certificate.
4. The certificate validity.

SSL Certificate Analysis of Free Content Websites

In this section, we analyze the fundamental characteristics of free content websites' and premium websites' SSL certificates, including their validity analysis, issuer organization and country distri-



(a) SSL Mismatched Domain

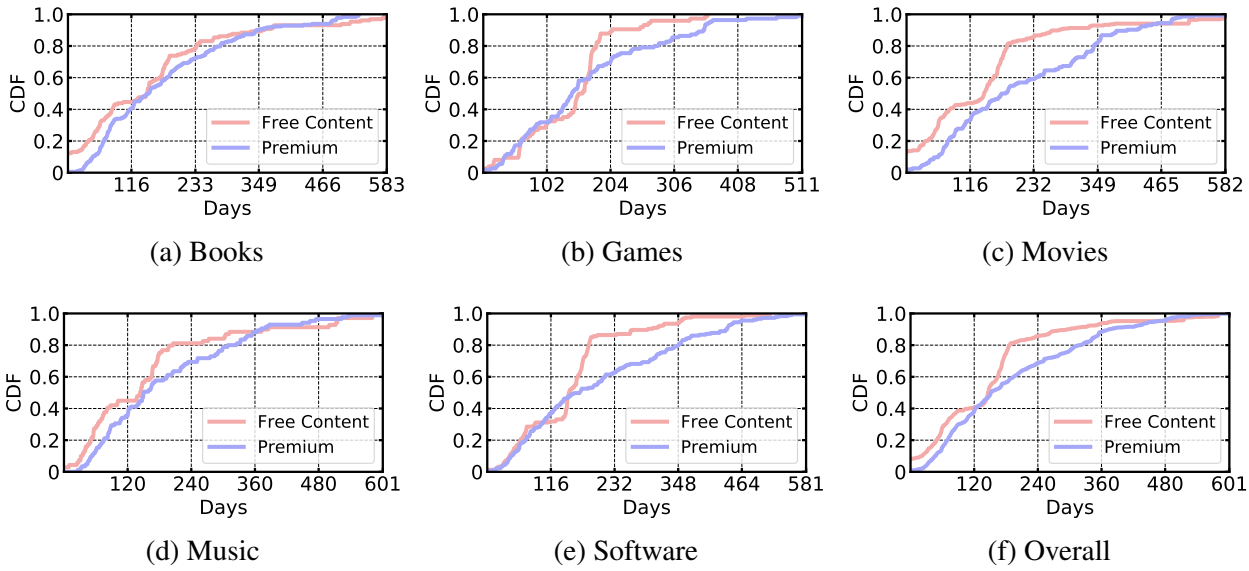
(b) SSL Expired

(c) SSL Invalid

Figure 4.1: The SSL certificate analysis results. We observe that almost 36% of the free content websites have problematic SSL certificates (unmatched, expired, or invalid) compared to 7% in premium websites.

bution, and signature algorithm and public key size (security parameters).

SSL Certificate Validity Analysis. Among the compiled free websites, we notice a significant portion (35.85%) of them have issues with their certificates (*i.e.* 11.47% unmatched name, 6.97% expired, and 17.42% invalid certificate), compared to only 6.99% of the premium websites' SSL certificates. To better understand the implications of such issues, we divided our analysis into four directions as follows.



(a) Books (b) Games (c) Movies

(d) Music (e) Software (f) Overall

Figure 4.2: The CDF of SSL certificate validity days. The premium websites SSL certificates are valid over extended period of time, unlike the free content counterparts, where multiple instances are expired.

1. **SSL Mismatched Domain.** Mismatched domain indicates either (i) impersonation of another website, or (ii) inconsistent website migration and domain transfer, and both indicate a lack of rigorous security practices by the service providers. Fig. 4.1a shows the vast discrepancy between free and premium websites in the context of SSL mismatching. For instance, 19% of the "Books" related free websites have unmatched domain names, in comparison with only 3% of the premium counterpart. Along with the five categories, 12% of the free content websites' SSLs have domain mismatch, in comparison with only 2.3% of the premium websites.

2. **SSL Validity Days.** We analyze the SSL certificate validity (*i.e.* number of days until the SSL certificate becomes invalid if not renewed), in Fig. 4.2. Notice that the validity days for premium websites are noticeably higher than their free websites, particularly for "Movies" and "Software" websites.

3. **SSL Expired.** SSL expiration may not directly affect users' security or usage experience, but is an indication that the used data encryption may be out-of-date, increasing the risk of future

exploitation. In line with our previous observation, we notice that the expired SSL certificates (*i.e.* ≤ 0 validity days) for free content websites are significantly more than for the premium websites—*i.e.* 7% of the free content websites' SSL certificates are expired, in comparison with only 1% of the premium websites—as shown in Fig. 4.1b. This may be attributed to the fact that free content websites operators are not renewing the SSL certificate for being unwilling to increase their operational costs. Nonetheless, this practice leads to potential risks regarding user information and data privacy.

4. **SSL Invalid.** Unlike the previous issues, invalid SSL indicates the usage of vulnerable and insecure elements within the website content. This can highly affect the users' data and client safety with potential data and resources exploitation. Toward understanding the risks associated with free content websites, Fig. 4.1c shows the percentage of websites associated with invalid SSL certificates. In particular, 17% of the free content websites' certificates are invalid, in comparison with only 4% of the premium websites. This gap is even higher for the "Movies" category, with 24.5% and 6% invalid free and premium websites SSL certificates, respectively.

Key Takeaway: On the SSL certificate fundamental level, the free and premium websites are highly distinguishable, where the free websites' certificates are 17% invalid, 7% expired, and 12% with mismatched domain names.

SSL Certificate Issuer Analysis. A certificate authority (CA) is an organization that issues digital certificates by signing with their private key. To understand the characteristic differences between free and premium websites, we analyze the hosting platforms and their country-level distribution.

1. **SSL Certificate Issuer Organization.** Table 4.2 and Table 4.3 show the distribution of the free and premium websites certificates' issuing organizations. We found the free content websites heavily use "Cloudflare" for their SSL certificates (38.22% of websites), in comparison with only 15.88% of the premium websites. Moreover, while "DigiCert" is not commonly used for free websites, with only 5.19% of the websites' SSL certificates associated with the orga-

Table 4.2: A comparison between free content and premium websites (%) in terms of SSL certificate issuer organizations.

Free Content Websites						
Issuer Organization	Books	Games	Movies	Music	Software	Overall
Cloudflare_ Inc.	24.80	59.72	35.29	27.94	48.39	38.22
Let's Encrypt	35.20	27.78	32.94	39.71	30.97	33.04
Sectigo Limited	15.20	4.17	11.76	4.41	5.81	9.48
DigiCert_ Inc.	6.40	2.78	6.27	10.29	1.29	5.19
cPanel_ Inc.	4.00	1.39	3.53	0.00	3.23	2.96
Cisco	0.80	1.39	5.49	0.00	0.00	2.37
GoDaddy.com_ Inc.	4.80	1.39	0.78	4.41	1.94	2.22
Others	8.80	1.39	3.92	13.24	8.39	6.52
Premium Websites						
Issuer Organization	Books	Games	Movies	Music	Software	Overall
DigiCert_ Inc.	22.16	25.23	24.11	20.24	22.47	22.89
Cloudflare_ Inc.	18.92	23.42	7.09	14.29	15.73	15.88
Let's Encrypt	19.46	12.61	13.48	15.48	14.04	15.31
Amazon	14.59	7.21	13.48	13.10	8.99	11.59
GoDaddy.com_ Inc.	7.57	3.60	8.51	10.71	12.92	8.87
Sectigo Limited	7.57	10.81	5.67	3.57	12.36	8.44
GlobalSign nv-sa	5.95	8.11	10.64	11.90	3.37	7.30
Others	3.78	9.01	17.02	10.71	10.11	9.73
All Websites						
Issuer Organization	Books	Games	Movies	Music	Software	Overall
Cloudflare_ Inc.	21.29	37.70	25.25	20.39	30.93	26.86
Let's Encrypt	25.81	18.58	26.01	26.32	21.92	24.02
DigiCert_ Inc.	15.81	16.39	12.63	15.79	12.61	14.19
Sectigo Limited	10.65	8.20	9.60	3.95	9.31	8.95
Amazon	9.35	4.37	5.81	9.21	5.41	6.70
GoDaddy.com_ Inc.	6.45	2.73	3.54	7.89	7.81	5.60
GlobalSign nv-sa	3.87	4.92	4.04	8.55	2.40	4.22
Others	6.77	7.10	13.13	7.89	9.61	9.46

nization, it is commonly used among the premium websites, with 22.89% of their certificates issued by the organization.

2. **SSL Certificate Issuing Country.** Next, we explored the country-level distribution of the SSL certificate issuing organizations, shown in Table 4.4. Notice that, for both categories, the United States dominates the distribution, with 86.88% and 83.31% of the free and premium websites

Table 4.3: The difference between free content and premium websites (%) in terms of SSL certificate issuer organizations.

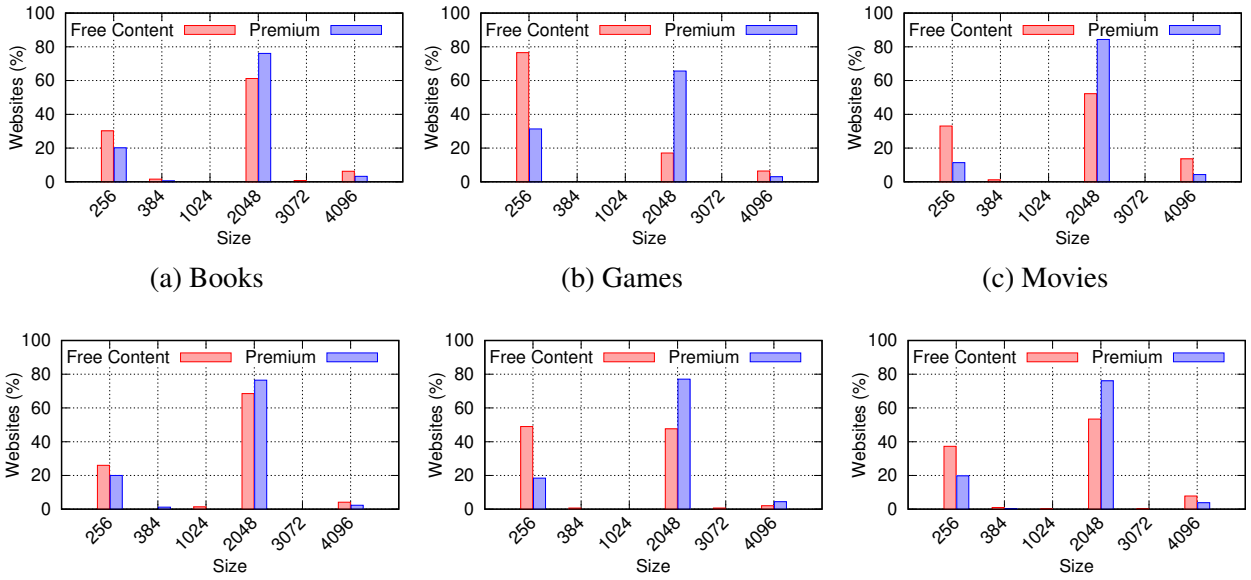
	Free Content		Premium		
Issuer Organization	#	%	#	%	Diff (%)
Cloudflare_ Inc.	258	38.22	111	15.88	+22.34
Let's Encrypt	223	33.04	107	15.31	+17.73
Sectigo Limited	64	9.48	59	8.44	+01.04
DigiCert_ Inc.	35	5.19	160	22.89	-17.70
cPanel_ Inc.	20	2.96	7	1.00	+01.96
Cisco	16	2.37	0	0.00	+02.37
GoDaddy.com_ Inc.	15	2.22	62	8.87	-06.65
Amazon	11	1.63	81	11.59	-09.96
GlobalSign nv-sa	7	1.04	51	7.30	-06.26
Google Trust Services LLC	1	0.15	8	1.14	-01.00
Entrust_ Inc.	0	0.00	12	1.72	-01.72
Others	25	3.70	41	5.87	-02.16

SSL certificate issuing organizations, respectively. Overall, 94.56% of the issuing organizations are located within the United States and the United Kingdom.

Key Takeaway: On country-level, the SSL certificate issuing organizations are very similar. However, we uncover the heavy usage of "Cloudflare" among the free websites (38.22% of the websites), in comparison with only 15.88% among the premium counterparts.

SSL Certificate Signature Analysis. In a website's SSL certificate, the key is split into two pieces. One piece is used to encrypt a message and the other is used to decrypt it. These keys allow exchanging information over unsecured channels. Alternatively, the decryption (private) key is used for signing messages and the encryption (public) key is used for signature verification by the recipient. The strength of the encryption (alternatively, signature) is determined by two factors: (i) the used algorithm and (ii) the used key size.

Signature Algorithms. Table 4.5 and Table 4.6 show the different signature algorithms used by free and premium websites to sign data (*i.e.* payload). We observe that while 60.74% of the free content websites use SHA256 with RSA signature mechanism (hash-then-sign), the majority



(a) Books (b) Games (c) Movies

(d) Music (e) Software (f) Overall

Figure 4.3: The key size analysis results. We observed that while majority of websites uses the key size of 2048, the portion of free content websites using key size of 256 is significantly higher than premium websites, particularly in "Games" and "Software" categories.

(83.26%) of the premium websites use this algorithms combination (∼23% difference). This is mainly attributed to it being the traditional go-to algorithm solution adopted by service providers. On the other hand, 38.37% of the free content websites rely on the newer and faster ECDSA (Elliptic Curve Digital Signature Algorithm) algorithm, which uses shorter keys for the same security level as in RSA with larger keys. In comparison, ECDSA is used by only 16.60% of the premium websites. We note that, while ECDSA is a newer and more efficient algorithm adopted by the newer free websites, recent studies suggest that it is more vulnerable to attacks [126] than the traditional RSA algorithm with post-quantum adversary.

Key Size. The other factor in enhancing the strength of the encryption and signature is the key size. A larger key size exponentially increases the time needed to crack and decrypt the encrypted information (and conversely for the signature forgery). Fig. 4.3 shows the commonly used key sizes among the websites. We note that the "Firefox" Internet browser no longer supports a key

Table 4.4: A comparison between free content and premium websites(%) in terms of SSL certification issuer countries.

Free Content Websites						
Issuer Country	Books	Games	Movies	Music	Software	Overall
United States (US)	82.17	100.00	86.30	86.30	91.28	86.88
United Kingdom (UK)	14.73	0.00	13.70	6.85	4.70	10.08
Belgium (BE)	0.78	0.00	0.00	4.11	1.34	1.12
Austria (AT)	0.00	0.00	0.00	0.00	1.34	0.64
Self-Sign	0.78	0.00	0.00	1.37	1.34	0.64
Australia (AU)	0.78	0.00	0.00	0.00	0.00	0.32
Netherlands (NL)	0.78	0.00	0.00	0.00	0.00	0.16
China (CN)	0.00	0.00	0.00	1.37	0.00	0.16
Premium Websites						
Issuer Country	Books	Games	Movies	Music	Software	Overall
United States (US)	84.78	80.81	81.56	80.00	82.12	83.31
United Kingdom (UK)	7.61	10.10	7.80	3.53	13.41	8.83
Belgium (BE)	5.98	8.08	10.64	11.76	3.35	6.26
France (FR)	0.54	0.00	0.00	3.53	0.00	0.64
China (CN)	1.09	0.00	0.00	0.00	0.56	0.48
Japan (JP)	0.00	0.00	0.00	1.18	0.00	0.16
Italy (IT)	0.00	0.00	0.00	0.00	0.56	0.16
Austria (AT)	0.00	1.01	0.00	0.00	0.00	0.16
All Websites						
Issuer Country	Books	Games	Movies	Music	Software	Overall
United States (US)	83.71	86.99	83.18	82.91	86.28	85.10
United Kingdom (UK)	10.54	6.85	9.81	5.06	9.45	9.46
Belgium (BE)	3.83	5.48	7.01	8.23	2.44	3.69
France (FR)	0.32	0.00	0.00	1.90	0.00	0.32
China (CN)	0.64	0.00	0.00	0.63	0.30	0.32
Japan (JP)	0.00	0.00	0.00	0.63	0.00	0.08
Italy (IT)	0.00	0.00	0.00	0.00	0.30	0.08
Austria (AT)	0.00	0.68	0.00	0.00	0.61	0.40
Australia (AU)	0.32	0.00	0.00	0.00	0.00	0.16
Netherlands (NL)	0.32	0.00	0.00	0.00	0.00	0.08
Self-Sign	0.32	0.00	0.00	0.63	0.61	0.32

size of less than 2048. We observe that the shorter keys (256, for example) are associated with the ECDSA algorithm choice. However, we note that ECDSA-256 has the same security level as RSA-3072. Among the free websites, 38% are using this key size and algorithm choice, making them more secure than the majority of websites utilizing RSA key of 2048 bits (in comparison

Table 4.5: A comparison between free content and premium websites (%) in terms of SSL certificate signature algorithms.

Free Content Websites						
Signature Algorithm	Books	Games	Movies	Music	Software	Overall
SHA256 with RSA	73.60	38.89	64.71	69.12	50.32	60.74
SHA256 with ECDSA	24.80	59.72	35.29	27.94	48.39	38.22
SHA1 with RSA	0.80	1.39	0.00	1.47	0.00	0.44
SHA384 with RSA	0.80	0.00	0.00	1.47	0.65	0.44
SHA384 with ECDSA	0.00	0.00	0.00	0.00	0.65	0.15
Premium Websites						
Signature Algorithm	Books	Games	Movies	Music	Software	Overall
SHA256 with RSA	80.54	75.68	92.20	83.33	83.71	83.26
SHA256 with ECDSA	19.46	23.42	7.80	16.67	16.29	16.60
SHA384 with RSA	0.00	0.90	0.00	0.00	0.00	0.14
SHA1 with RSA	0.00	0.00	0.00	0.00	0.00	0.00
SHA384 with ECDSA	0.00	0.00	0.00	0.00	0.00	0.00

to only 19% of premium websites). Across all categories, the premium websites usage of keys size 2048 is significantly higher, showing that they might not be as secure—although theoretically, since such an insecurity is only possible with a post-quantum adversary.

A plausible explanation of the choice of algorithm and key size is that free content websites are emerging, often with a short life time, making them apt to the utilization of new algorithms, in contrast to well-established premium services deployed for many years where key-rollover and algorithm update are costly.

Key Takeaway: On the data signature algorithms, more premium websites are still using RSA while more free content websites have adopted the more recent ECDSA algorithm, possibly due to the more recent emergence of free content websites, making them easier to adopt new technologies.

Summary & Concluding Remarks

The Internet is the dominant channel for marketing, promotion, and communication, particularly via providing online physical and digital services. Recent years have witnessed the rise of websites

Table 4.6: The difference between free and premium content websites (%) in terms of SSL certificate signature algorithms.

	Free Content		Premium		
Signature Algorithm	#	%	#	%	Diff (%)
SHA256 with RSA	410	60.74	582	83.26	-22.52
SHA256 with ECDSA	258	38.22	116	16.60	+21.63
SHA1 with RSA	3	0.44	0	0.00	+00.44
SHA384 with RSA	3	0.44	1	0.14	+00.30
SHA384 with ECDSA	1	0.15	0	0.00	+00.15

that provide content for free, dubbed free content websites. In this work, we explored the unique SSL certificate characteristics of premium and free content websites to understand their commonalities and differences. Understanding the potential risks associated with invalid SSL certificates, including, but not limited to, untrusted SSL certificates, expired certificates, mixed content warnings, or invalid/vulnerable information and content. Through our analysis, we highlight that 35.85% of the free websites' certificates have significant issues, with 17% invalid, 7% expired, and 12% with mismatched domain names. Moreover, we uncover the usage of the emerging ECDSA encryption algorithm among the free websites with a key size that would seem to provide better (classical) security and performance than that of the algorithm (and associated key size) used in premium websites; with 38% of free content websites using the ECDSA key size of 256, in comparison with only 20% of their premium counterparts. Toward a safe and secure web environment, we highlight that free content websites would highly benefit from consistent monitoring and management, particularly with the increasing trend of invalid SSL certificates among them (although not the only risk). Our observations raise concerns regarding the safety of using such free services, especially when such usage could put users at risk, and call for in-depth analysis of their potential risks, ramifications, and remedies.

CHAPTER 5: A COMPREHENSIVE ANALYSIS AND MEASUREMENTS OF FREE CONTENT WEBSITES' INFRASTRUCTURE AND HTTP HEADERS

The rise of online services' popularity sheds light on their reoccurring data security and privacy risks. For instance, online services that provide free content to the user (*i.e.* free software or music) are driven by advertisements and data collection [132, 64, 61, 30, 138]. Although the free content websites are not necessarily mean illegal websites, they may suffer from various security threats due to their heavy reliance on third-party tools and integrated services without a customized control of security tools. The premium content websites, on the other hand, offer services through fees, *e.g.* subscriptions or pay-as-you-use models. This allows the websites to reduce the reliance on advertisements, ensuring a very high level of quality of service as a result of well-designed websites that are well-maintained through dedicated engineering and operational efforts.

The adopted monetary scheme in free content websites may lead to prioritizing data collection and user tracking over users' safety [84, 83, 96]. Particularly, the extensive usage of third-party advertisements in free content websites results in various exploitations and vulnerabilities. Moreover, the used third-party tools may be exploited for data and information leakage or even executing malicious scripts on the user device [84].

With all available literature on their risks, it is crucial to understand the security configurations and adaptation differences between free and premium content websites, answering the ongoing question of *"are premium websites safer to use than their free counterparts?"*. In this work, we analyze the websites across two vectors: (i) HTTP header attributes, including the network attributes and packets transfer configurations, and most importantly, the security protocols utilization. (ii) Website's domain hosting infrastructure, including the IP address lifespan and changing frequency, alongside the hosting service. In particular, we investigate HTTP attribute-wise differences between free and premium websites, discovering intra-shared patterns among free content websites,

with commonly followed security configurations.

Studying and analyzing the HTTP header is well conducted in the literature. For instance, McGahagan *et al.* [68] applied HTTP header features to the field of malicious website detection, analyzing ≈46,000 websites by extracting 22 HTTP header features. Their study showed that the usage of HTTP header features leads to more accurate detection of malicious websites. Similarly, Laughter *et al.* [75] proposed a malicious HTTP request detection method using a machine learning approach and analyzing the network traffic, including the HTTP header fields. Their study showed that using HTTP response headers enables identifying malicious requests with an accuracy of 93.6%.

While HTTP header attributes were previously utilized for malicious content detection and identification, in this work, we focus on the intra-shared configurations and patterns among free and premium content websites, including security and encoding configurations. Our analysis concludes that premium content websites are more secure in comparison with their free counterparts. This is mainly contributed to that followed configurations, including HTTP alternative service, where a higher percentage of premium websites, 85.71%, do not allow HTTP alternative service, and 42.58% have the content type configuration set to *"nosniff"*. Further, premium websites allow data encoding for secure and fast transfer, whereas free content websites may limit the encoding protocol for performance gains. Our analysis of seven HTTP header attributes uncovers that the security configurations of premium content websites are significantly superior to their free content counterparts. This is more prevalent later in this work, where five out of the most distinguishable attributes are security-related, with premium websites enabling the security configurations. In contrast, free content websites do not adopt secure protocols.

Summary of Completed Work

Starting with a list of 1,562 free and premium services websites, we assess HTTP header and domain infrastructural differences between free and premium content websites. In particular, our

contributions are across the following verticals.

1. **Dataset Collection & Augmentation.** We compiled a dataset of 834 free content websites alongside 728 premium content websites. Then, the HTTP header and domain infrastructure information are extracted for further analysis.

2. **HTTP High-level Analysis.** We conduct an HTTP response-based analysis. Our findings highlight that the reliance of free content websites on cache usage is significantly higher than premium websites, and there is a lack of diverse configurations among free content websites, indicating a potential usage of default configurations.

3. **HTTP Attributes Analysis.** We investigate the HTTP content attributes toward understanding the fundamental differences between free and premium content websites. We highlight that premium content websites are more secure in comparison with their free counterparts, where premium content websites HTTP header configurations highly reduce the attacker adversarial surface, leading to a more secure user experience.

4. **HTTP Encoding Configurations Analysis.** Investigating the encoding protocols, we highlight that premium websites allow encoding of data for secure and fast transfer, whereas free content websites may limit the encoding protocol for performance gains.

5. **HTTP Security Attributes Analysis.** We analyze seven HTTP header attributes, uncovering that the security configurations of premium content websites are significantly superior to their free content counterparts. This includes securing the communication channel between the client and the server by preventing data sniffing and man-in-the-middle attacks, among others.

6. **Domain Infrastructure Analysis.** We conduct infrastructure-level analysis of free and premium content websites. Our findings conclude the heavy usage of Amazon, Akamai, and Google services among premium websites in contrast to Cloudflare services among the free content websites. Further, premium websites provide higher security against DDoS attacks, changing the IP address of the websites and associated resources multiple times per day.

Table 5.1: An overview of the collected dataset. The collected URLs are associated with five different categories, and belong to free content and premium websites. Overall, 1,562 websites are crawled for the purpose of this study.

	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

7. **Websites Attributes Correlation.** We extract the most important feature for online services categorization. Our analysis shows that among the top-10 features used to distinguish between free and premium content websites, five are related to communication and security configurations, raising alarming concerns regarding free content websites.

Data Collection

For this work, we compiled a dataset of 834 free content websites, alongside 728 premium websites. The process of dataset collection and processing is divided into three steps: (i) URLs collection, including retrieving and categorizing the free and premium website URLs. (ii) HTTP header-related attributes extraction, where each website's URL is queried to obtain the HTTP response information for further analysis. (iii) Infrastructure-related features, augmenting the dataset with other essential features to study the infrastructural differences between free and premium websites.

URLs Collection. In this work, we compiled a list of 1,562 free content and premium websites. When selecting the websites, we consider three factors: (i) the most popular websites, (ii) websites that appear in the top results by Google, DuckDuckGo, and Bing search engines, and (iii) websites that would contribute to maintain a balanced dataset. Each website was manually examined and labeled as either premium or free content. The websites are then categorized manually into five groups based on the content they provide: books, games, movies, music, or software. The distribution of the collected URLs is shown in Table 5.1.

HTTP Response Headers Extraction. After retrieving the URLs, various HTTP header attributes are extracted. The HTTP response headers contain information exchanged between the server and the client during the communication session. In order to extract those features, we use the Python library "Request" [123], which allows the user to send HTTP requests to a server using the URL and receive the response using the *"Get"* function. At first, the collected HTTP header data consisted of 760 different features. However, most of those features do not contain valuable information for website analysis. Therefore, those features are taken out of this study as they provide no valuable insights. Out of the 760 features, and after removing unrelated/non-valuable ones, we compiled a total of 27 features for our evaluation.

Infrastructure-level Features. We further investigate the infrastructural information, including the domain name server information, hosting infrastructure, and IP address change frequency. Toward this goal, we utilized several tools to augment the dataset with attributes that can help in differentiating free content websites from their premium counterparts in terms of their infrastructure preference. For instance, we use *SecurityTrails* [133] to retrieve the websites' historical CDN and hosting servers (i.e., DNS records) for the past 13 years. Moreover, we track the IP address changes and lifetime within online content websites.

HTTP Response Header Analysis

HTTP response header contains the exchanged information between the server and the client. In this section, we provide four HTTP response header attributes analysis, including: (i) HTTP High-level Analysis, analyzing the general attributes of the HTTP response headers, including *HTTP Status Code*, *HTTP Connection Type*, *HTTP Servers*, and *HTTP Cache-Control*. (ii) HTTP Attributes Analysis, including analyzing the *HTTP Alternative Service*, *HTTP Content-Length*, and *HTTP X-Content-Type-Options*. (iii) HTTP Encoding Configurations Analysis, including all features of HTTP response headers related to encoding, such as *HTTP Encoding Type*, and *HTTP Content-Encoding (Gzip)*. (iv) HTTP Security Attributes Analysis, analyzing the security-related

features such as *HTTP X-Frame-Options*, and *HTTP X-XSS-Protection*.

HTTP High-level Analysis. In this section, we investigate several general features of HTTP response headers to understand the main similarities and differences between free and premium websites. These features include *HTTP Status Code*, *Connection Type*, *Servers*, and *Cache-Control*.

HTTP Status Code. Upon sending an HTTP request to the server, the server responds with a status code indicating whether the request has been completed or not. For instance, the code "200" indicates that the HTTP request made by the client has been successfully completed. Table 5.2 shows that the "200" HTTP status code is the most popular among the free content and premium websites with a percentage of 74.11% and 97.20%, respectively. However, we find a high percentage (23.72%) of the free content websites having the HTTP status code of "403" compared to only 1.96% in the premium counterparts. The "403" code indicates unauthorized access to the server, which is an indication that the client has no right to perform an HTTP request to the server "Access Forbidden". The aforementioned status codes are the most common with the performed HTTP requests. However, there are other status codes that make a total of less than 3%, such as "Unauthorized", "Service unavailable", "Connection timed out", and "Origin is unreachable".

HTTP Connection Type. Connection type is another feature of the HTTP response headers, indicating whether the connection between the client and server stays open or will close after completing the current request. The majority of the websites, 90.31% in free content and 85.99% of the premium websites, keep the default value for their connections which is the "keep-alive", as shown in Table 5.3. The "keep-alive" flag allows the subsequent requests arriving on the server to be completed without opening a new session. On the other hand, only 5.23% of the free content and 3.22% of the premium websites use the "close" connection type, which indicates that the server requires closing the connection after completing each HTTP request, where a new session is needed for further communications.

HTTP Servers. This feature is related to the server that handles the HTTP request and generates the appropriate response to the client's request. Our analysis uncovers that most collected web-

Table 5.2: A comparison between the different categories of websites (%) in terms of HTTP status code. The "Others" group includes the aggregate of websites with status codes: 401, 404, 500, 503, 521, 522, or 523.

Free content Websites						
Code	Books	Games	Movies	Music	Software	Overall
200	71.23	56.58	73.14	72.15	86.78	74.11
403	28.08	39.47	23.62	26.58	12.07	23.72
Others	0.68	3.95	3.24	1.27	1.15	2.17
Premium Websites						
Code	Books	Games	Movies	Music	Software	Overall
200	97.35	97.30	98.01	96.47	96.63	97.20
403	1.59	2.70	1.32	2.35	2.25	1.96
Others	1.06	0.00	0.66	1.18	1.12	0.84
All Websites						
Code	Books	Games	Movies	Music	Software	Overall
200	85.97	80.75	81.30	84.76	91.76	85.11
403	13.13	17.65	16.30	14.02	7.10	13.35
Others	0.90	1.60	2.39	1.22	1.14	1.54

sites rely on external CDNs to host their contents. Table 5.4 and Table 5.5 show the statistics of the server type information that is embedded in HTTP response headers of the free and premium websites. We found that both groups heavily use "Cloudflare", with 28.57% of free content websites and 22.69% of the premium websites. We notice that among the free content websites, "Cisco Umbrella" is the second highest used server, with 21.56% compared to 0% among the premium counterparts. The "Apache" server, on the other hand, is the second mostly used server with 19.47%.

Another significant factor that differentiates the free content websites from premium websites is the total number of private servers (i.e., privately owned by the website's owner), as shown in Table 5.4 and Table 5.5. The "Others" group contains only 11 private servers in the free content websites compared to 52 private servers in the premium counterparts. This finding indicates that premium websites are more diverse in using private servers in comparison with free content websites, which rely more on the CDN servers.

Table 5.3: A comparison between the different categories of websites (%) in terms of HTTP connection type. The "N/A" group includes websites with unavailable connection type.

Free Content Websites						
Connection	Books	Games	Movies	Music	Software	Overall
Keep-alive	89.73	92.11	87.06	91.14	95.40	90.31
Close	4.11	2.63	7.77	5.06	2.87	5.23
N/A	6.16	5.26	5.18	3.80	1.72	4.46
Premium Websites						
Connection	Books	Games	Movies	Music	Software	Overall
Keep-alive	87.83	92.79	83.44	80.00	84.83	85.99
Close	5.29	1.80	3.31	1.18	2.81	3.22
N/A	6.88	5.41	13.25	18.82	12.36	10.78
All Websites						
Connection	Books	Games	Movies	Music	Software	Overall
Keep-alive	88.66	92.51	85.87	85.37	90.06	88.25
Close	4.78	2.14	6.30	3.05	2.84	4.27
N/A	6.57	5.35	7.83	11.59	7.10	7.48

HTTP Cache-Control. HTTP caching is used to improve a website's performance and reduce network latency by reusing the previously fetched content in the cache memory or disk. In the HTTP response header, the server can specify several instructions on controlling the caching on the client's browser. Table 5.6 shows the usage of "Must-revalidate" in 20.66% of the free content websites and 27.87% of the premium websites within the HTTP cache-control instructions. The "Must-revalidate' means that the HTTP response can be reused from the cache until the expiration of the communication time; once it becomes expired, the origin server must re-validate the response to be reused by the client. From our cache control analysis, we find that premium websites take a much lower risk when dealing with HTTP connections for the following reasons; (i) 37.54% of the premium websites, compared to 19.52% in free content websites, use the "No-cache" configuration, which indicates that the response cannot be reused from the cache unless the origin server validates it. (ii) 28.85% of the premium websites, compared to 17.09% in free content websites, use the "No-store" configuration, indicating that the response is not stored in the cache. (iii) 17.51% of premium websites, compared to 8.04% in free content websites, use the "Private

Table 5.4: A comparison between the different categories of websites (%) in terms of HTTP server. The "Others" group includes the aggregate of 11 servers for free content websites and 52 servers for premium websites.

Free Content Websites						
Server	Books	Games	Movies	Music	Software	Overall
Cloudflare	12.33	38.16	26.86	12.66	48.28	28.57
Cisco Umbrella	27.40	38.16	20.71	24.05	9.77	21.56
Nginx	28.77	5.26	24.92	30.38	9.77	20.92
Apache	17.81	9.21	16.18	12.66	18.39	15.94
Openresty	2.74	0.00	7.12	5.06	1.15	4.08
LiteSpeed	0.00	3.95	0.97	6.33	7.47	3.06
Others	10.96	5.26	3.24	8.86	5.17	5.87
Premium Websites						
Server	Books	Games	Movies	Music	Software	Overall
Cloudflare	27.51	33.33	13.25	16.47	21.91	22.69
Apache	20.11	14.41	13.91	18.82	26.97	19.47
Nginx	15.34	18.02	14.57	22.35	19.66	17.51
Microsoft	5.29	1.80	3.97	4.71	5.06	4.34
Openresty	2.65	2.70	3.31	1.18	1.12	2.24
AmazonS3	0.53	0.90	3.97	3.53	1.12	1.82
Others	28.57	28.83	47.02	32.94	24.16	31.93
All Websites						
Server	Books	Games	Movies	Music	Software	Overall
Cloudflare	20.90	35.29	22.39	14.63	34.94	25.77
Nginx	21.19	12.83	21.52	26.22	14.77	19.29
Apache	19.10	12.30	15.43	15.85	22.73	17.62
Cisco Umbrella	11.94	15.51	13.91	11.59	4.83	11.28
Openresty	2.69	1.60	5.87	3.05	1.14	3.20
Microsoft	4.18	1.60	1.52	3.05	2.84	2.60
Others	20.00	20.86	19.35	25.61	18.75	20.23

cache" settings indicating that the response can be stored only in a private cache.

Key Takeaway: The reliance of free content websites on cache usage is significantly higher than premium ones. Moreover, our analysis shows that premium website configurations are more diverse in comparison with free content websites. This, conservatively said, may indicate that free content websites are not customizing the configurations to meet the website's needs but follow standard or default configurations.

Table 5.5: The difference in (%) between free content and premium websites in terms of HTTP server.

	Free Content		Premium		
Server	#	%	#	%	Diff (%)
Cisco Umbrella	169	21.56	0	0.00	+21.56
Cloudflare	224	28.57	162	22.69	+5.88
Apache	125	15.94	139	19.47	-3.52
Nginx	164	20.92	125	17.51	+3.41
Microsoft	8	1.02	31	4.34	-3.32
LiteSpeed	24	3.06	8	1.12	+1.94
Openresty	32	4.08	16	2.24	+1.84
AmazonS3	0	0.00	13	1.82	-1.82
eBay	0	0.00	9	1.26	-1.26
Tengine	1	0.13	5	0.70	-0.57
Others	37	4.72	206	28.85	-24.13

HTTP Attributes Analysis. In this section, we discuss the analysis of HTTP response header attributes related to the HTTP content. These features, presented in Fig. 5.1, include; *HTTP Alternative Service*, *HTTP Content-Length*, and *HTTP X-Content-Type-Options*.

HTTP Alternative Service. The HTTP Alternative Service header (Alt-Svc) allows a server to specify a particular content to be loaded from another server different from the origin. Note that the content still appears to the client as if it was downloaded from the same server. However, there are several security concerns about using *"Alt-Svc"*. For instance, Nottingham *et al.* [107, 108] presented several security considerations of using "Alt-Svc" such as downgrade attacks and tracking clients over time. Moreover, they stated that attackers could exploit *"Alt-Svc"* to initiate malicious port scans from the server that they control. Our analysis in Fig. 5.1a shows that, overall, 24.87% of free content websites are using the Alt-Svc, in comparison with only 14.29% of the premium websites. Most of those websites are among websites providing "Software" and "Games" content with 43.10% and 31.58% free content websites and 15.17% and 21.62% premium websites, respectively.

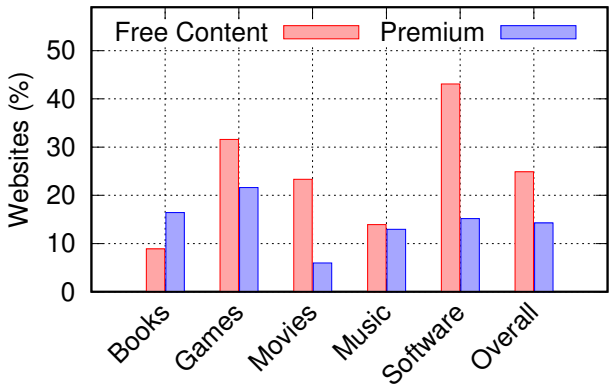
HTTP Content Length. The *Content-Length* attribute denotes the message body size that is being

Table 5.6: A comparison between the different categories of websites (%) in terms of HTTP Cache-Control. The "N/A" group includes websites with unavailable Cache-Control attributes.

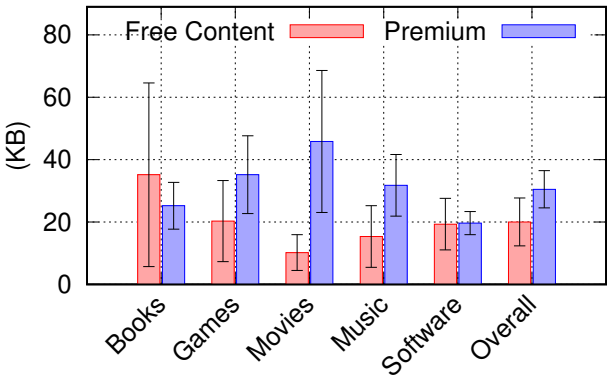
Free Content Websites						
Cache-Control	Books	Games	Movies	Music	Software	Overall
Must-revalidate	21.92	26.32	24.92	17.72	10.92	20.66
No-cache	21.92	19.74	23.62	16.46	11.49	19.52
No-store	19.18	18.42	22.01	13.92	7.47	17.09
Private cache	9.59	5.26	8.74	5.06	8.05	8.04
Public cache	2.74	5.26	3.56	6.33	10.92	5.48
Premium Websites						
Cache-Control	Books	Games	Movies	Music	Software	Overall
No-cache	35.98	44.14	41.06	31.76	34.83	37.54
No-store	26.46	31.53	30.46	30.59	27.53	28.85
Must-revalidate	30.69	28.83	28.48	25.88	24.72	27.87
Private cache	25.93	18.02	13.91	17.65	11.24	17.51
Public cache	8.47	7.21	9.93	9.41	11.24	9.38
All Websites						
Cache-Control	Books	Games	Movies	Music	Software	Overall
Must-revalidate	26.87	27.81	26.09	21.95	17.90	24.10
No-cache	29.85	34.22	29.35	24.39	23.30	28.10
No-store	23.28	26.20	24.78	22.56	17.61	22.70
Private cache	18.81	12.83	10.43	11.59	9.66	12.55
Public cache	5.97	6.42	5.65	7.93	11.08	7.34

sent within the HTTP response header from the server to the client. Fig. 5.1b shows that, except for the "Books" group, the overall average HTTP content-length in the premium website (30.49 KB) is higher than the average of the free content websites (20.01 KB). Moreover, the highest average length with the highest difference is in the "Movies" group, where it reached (45.83 KB) on premium websites compared to only (10.20 KB) on free content websites. Notice that the HTTP content length significant statistical differences are overlapping except for "Music" and "Movies". This indicates that using this attribute is infeasible for distinguishing between free and premium content websites.

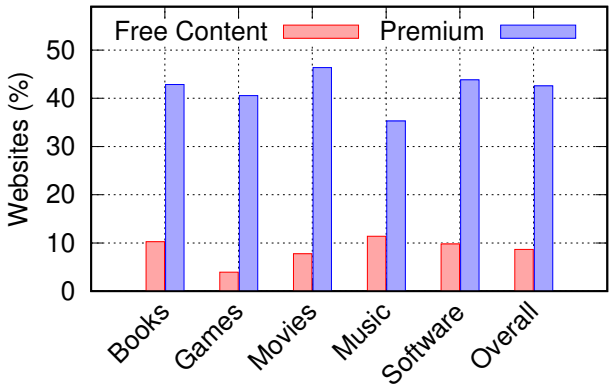
HTTP X-Content-Type-Options. This attribute is a part of the HTTP response header used by the server to inform the client's browser to follow the Mail Extensions (MIME) content types. The



(a) HTTP Alternative Service (Available).



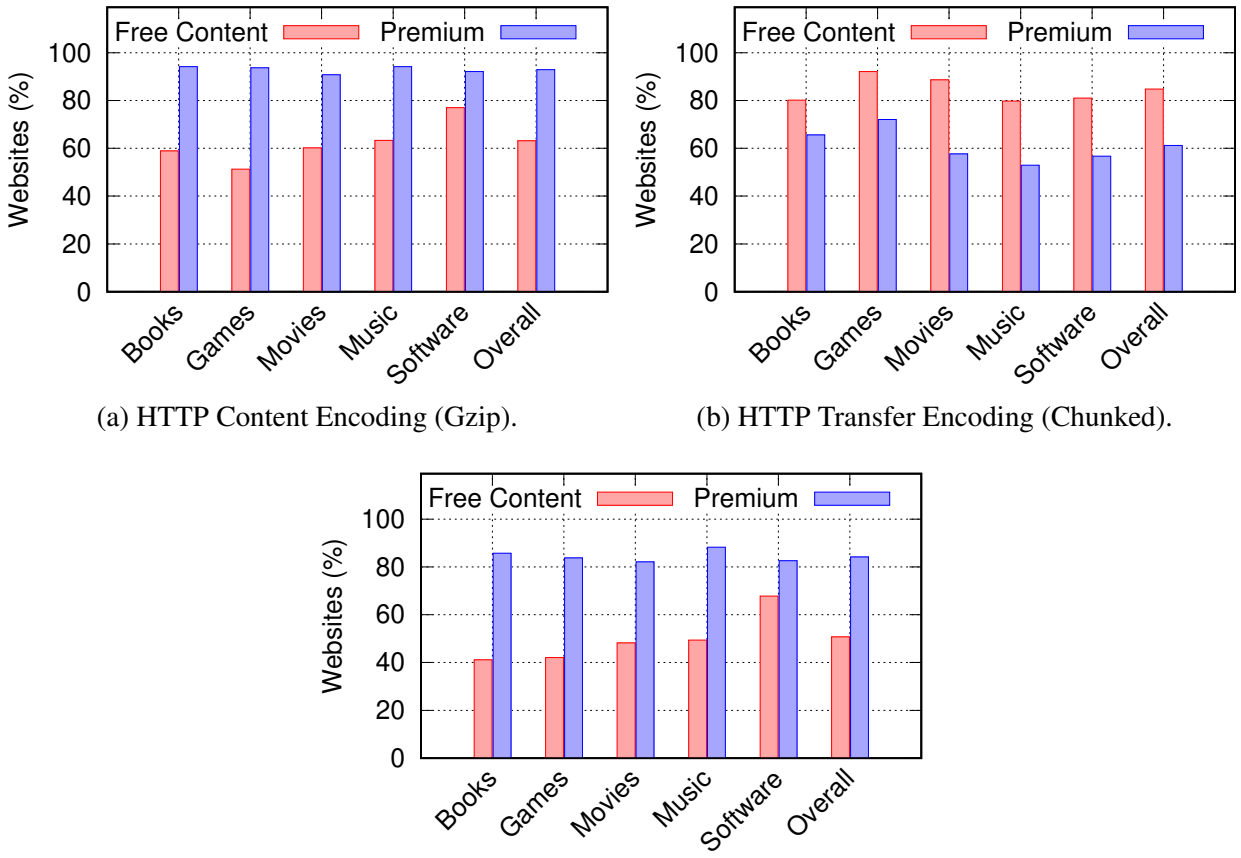
(b) HTTP Content Length (Average).



(c) X-Content-Type-Options (Nosniff).

Figure 5.1: Comparing different HTTP content features between free and premium websites.

"nosniff" assigned value indicates that the file type must be the same as the one declared by the server without any changes. Ignoring this feature in the HTTP header and not setting the value to *"nosniff"* can cause multiple security issues as discussed in the literature: (i) A browser can be misled by an attacker to execute a resource that was not intended for the web application to execute [77]. (ii) A browser can perform MIME-sniffing to cause an interpretation of the content and display it as a different content type [119]. (iii) An attacker can conduct a session hijacking attack by uploading malicious content to the server while the server expects another type of content [93]. Therefore, to prevent all previous security concerns, the website's server should set the response header value of *X-Content-Type-Options* to *"nosniff"*. Our analysis uncovers that only 8.67% of the free content websites send their HTTP response with the *X-Content-Type-Options* header value



(a) HTTP Content Encoding (Gzip).

(b) HTTP Transfer Encoding (Chunked).

(c) HTTP Accept Encoding.

Figure 5.2: Comparing different HTTP Encoding features between free content and premium websites.

set to *"nosniff"* compared to 42.58% in the premium websites.

Key Takeaway: Premium content websites are more secure in comparison with their free counterparts. This is mainly contributed to the higher percentage of premium websites that do not allow HTTP alternative services and have the content type configuration set to *"nosniff"*. These configurations reduce the attacker's adversarial surface, leading to a more secure user communication.

HTTP Encoding Analysis. In the following, we analyze the HTTP response headers attributes related to HTTP encoding protocols. These features include HTTP encoding type, Content-Encoding (Gzip), Transfer-Encoding (Chunked), and Accept-Encoding, and shown in Fig. 5.2 and Table 5.7.

Table 5.7: A comparison between the different categories of websites (%) in terms of HTTP encoding type. The "Others" group includes the aggregate of websites with encoding: WINDOWS-1251, ISO-639-2, gb2312, GBK, us-ascii, or iso-8859-15.

Free Content Websites						
Encoding	Books	Games	Movies	Music	Software	Overall
UTF-8	55.48	53.95	71.20	58.23	83.33	67.98
ISO-8859-1	41.78	43.42	28.48	41.77	16.67	31.12
Others	2.74	2.63	0.32	0.00	0.00	0.89
Premium Websites						
Encoding	Books	Games	Movies	Music	Software	Overall
UTF-8	86.77	87.39	80.79	78.82	83.71	83.89
ISO-8859-1	11.11	12.61	17.88	20.00	14.61	14.71
Others	2.12	0.00	1.32	1.18	1.69	1.40
All Websites						
Encoding	Books	Games	Movies	Music	Software	Overall
UTF-8	73.13	73.80	74.35	68.90	83.52	75.57
ISO-8859-1	24.48	25.13	25.00	30.49	15.63	23.30
Others	2.39	1.07	0.65	0.61	0.85	1.13

HTTP Encoding Type. The type of HTTP encoding is correlated to a website's performance by either consuming or saving network bandwidth and client resources. Therefore, studying the differences between free content and premium websites related to using a specific encoding protocol is essential. From our analysis, Table 5.7 shows that the majority of the websites within both categories are using the UTF-8 for HTTP response header encoding, with 67.98% in free content websites and 83.89% in the premium category. This is in line with the literature, where Faghani *et al.* [49, 151] showed that UTF-8 was the most commonly used encoding protocol within and up to 97.70 of the websites using UTF-8 encoding. Notice that the usage of "ISO-8859-1" encoding within the free content website is up to 31.12%, in comparison with only 14.71% in the premium websites.

HTTP Content-Encoding (Gzip). The HTTP Content-Encoding header states how the server encodes the content of the messages to allow the accurate client decoding of the received messages. The main usage of HTTP content encoding is to compress the messages for faster communication

between the server and the client. Note that *"Gzip"* is one of the most popular compression methods used by HTTP response headers for compatible encoding. Fig. 5.2a shows that *"Gzip"* is used in most of the free content and premium websites with a percentage of 63.14% and 92.86% respectively.

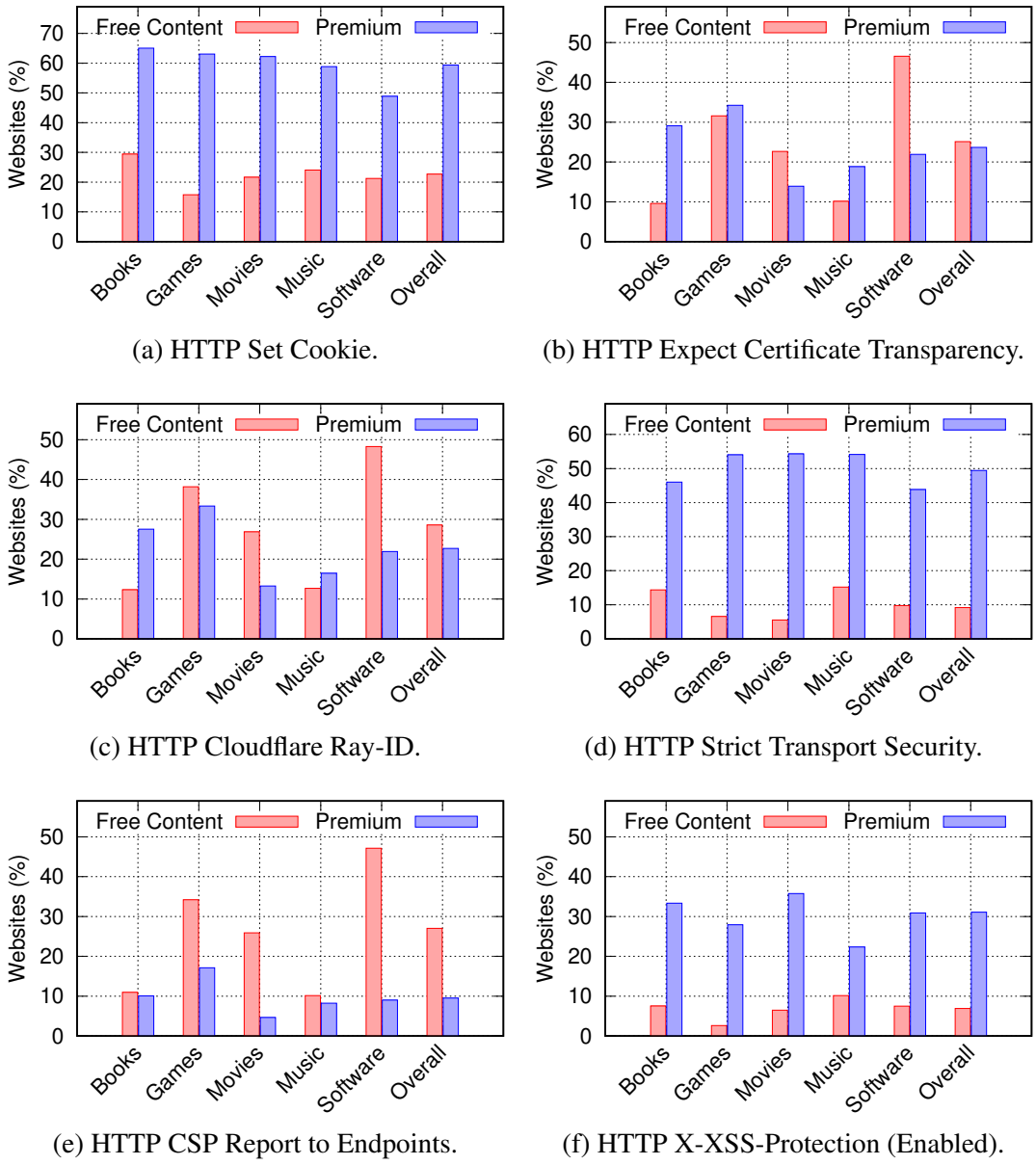
HTTP Transfer-Encoding (Chunked). Transfer encoding is the form in which the data is transferred from and to the server and client. Most websites, 84.82% of the free content and 61.20% of the premium websites, apply the "Chunked" encoding form by dividing the data into smaller units.

Accept-Encoding. The server may not accept encoding the message body even if it has the same encoding method as the client. For example, this may occur when the message body is already encoded, and there is no need to perform additional encoding, or when the server does not have enough resources to complete the encoding. However, our analysis shows that 84.17% of the premium websites accept HTTP header encoding, compared to 50.77% of the free content as shown in Fig. 5.2c.

Key Takeaway: Our findings highlight that, on encoding mechanisms and attributes, the free and premium websites are distinguishable. Premium websites, on the one hand, allow encoding of data for secure and fast transfer, whereas free content websites may limit the encoding configurations for performance gains.

HTTP Security Attributes Analysis. Toward understanding the security configurations of free and premium content websites, in this section, we study the security-related HTTP response headers, including the features presented in Table 5.8 and Fig. 5.3. These features include HTTP X-Frame-Options, Set-Cookie, Expect Certificate Transparency, Cloudflare Ray-ID, Strict Transport Security, and X-XSS-Protection.

HTTP X-Frame-Options. The main use of the HTTP *X-Frame-Options* header is to specify whether the client is allowed to render a page in a different HTML frame completely or only within the same origin server. Enabling third-party server rendering may lead to various security



(a) HTTP Set Cookie.

(b) HTTP Expect Certificate Transparency.

(c) HTTP Cloudflare Ray-ID.

(d) HTTP Strict Transport Security.

(e) HTTP CSP Report to Endpoints.

(f) HTTP X-XSS-Protection (Enabled).

Figure 5.3: Comparing different HTTP security features between free content and premium websites.

issues, such as the click-jacking attacks, where the attacker can ”hijack” the clicks by tricking the client into clicking on a button or link different from what they are trying to click [80, 127]. One prevention technique is to accurately configure the *X-Frame-Options* header to avoid click-jacking attacks. There are three options for the header's value (i) *SAMEORIGIN*: where the client's

Table 5.8: A comparison between the different categories of websites (%) in terms of HTTP X-Frame-Options. The "Undefined" group includes websites with unavailable X-Frame-Options.

Free Content Websites						
X-Frame-Options	Books	Games	Movies	Music	Software	Overall
SAMEORIGIN	9.59	9.21	11.97	7.59	9.77	10.33
DENY	3.42	0.00	0.65	2.53	0.00	1.15
ALLOW-FROM	0.00	0.00	0.32	0.00	0.00	0.13
Undefined	86.99	90.79	87.06	89.87	90.23	88.39
Premium Websites						
X-Frame-Options	Books	Games	Movies	Music	Software	Overall
SAMEORIGIN	42.33	42.34	47.02	37.65	41.01	42.44
DENY	7.94	9.01	8.61	8.24	5.62	7.70
ALLOW-FROM	0.53	0.00	0.00	0.00	1.69	0.56
Undefined	49.21	48.65	44.37	54.12	51.69	49.30
All Websites						
X-Frame-Options	Books	Games	Movies	Music	Software	Overall
SAMEORIGIN	28.06	28.88	23.48	23.17	25.57	25.63
DENY	5.97	5.35	3.26	5.49	2.84	4.27
ALLOW-FROM	0.30	0.00	0.22	0.00	0.85	0.33
Undefined	65.67	65.78	73.04	71.34	70.74	69.76

browser is allowed to display the web page in a frame that shares the same origin as the original page. (ii) *DENY*: to prevent the client's browser from displaying the web page in a frame. (iii) *ALLOW-FROM*: where the server can specify an origin that can be used for a page to be displayed in its frame. Our analysis in Table 5.8 shows that 11.61% of the free content websites deploy the X-Frame-Options headers (10.33% "SAMEORIGIN", 1.15% "DENY", and 0.13% ALLOW-FROM), compared to 50.70% in the premium websites (42.44%, 7.70%, and 0.56%) respectively.

HTTP Set Cookie. The Set-Cookie HTTP attribute is used to exchange a small piece of information, "a cookie" between the server and the client. A cookie can contain different types of information about: (i) session management, (ii) user profile data that the server needs to save, such as logins, items in the shopping carts, or game scores, (iii) User profile preferences and themes, (iv) User tracking, where the server can record a user behavior to be analyzed later for different purposes, such as targeted advertising [53]. Our analysis (shown in Fig. 5.3a) uncovers that the

Table 5.9: The percentage of total records for the top-10 infrastructure providers between (2008 and 2021).

Free Content Websites		Premium Websites		All Websites	
Organization	Records(%)	Organization	Records(%)	Organization	Records(%)
Amazon.com, Inc.	13.57	Amazon.com, Inc.	42.46	Amazon.com, Inc.	33.18
Netherlands	9.82	Akamai Technologies, Inc.	15.41	Akamai Technologies, Inc.	10.60
Cloudflare, Inc.	9.46	Akamai International B.V.	10.91	Akamai International B.V.	7.42
Leaseweb USA, Inc.	5.80	Google LLC	6.68	Google LLC	4.57
Confluence Networks Inc	4.90	SoftLayer Technologies Inc.	1.57	Cloudflare, Inc.	3.57
GoDaddy.com, LLC	4.37	Hurricane Electric LLC	1.51	Netherlands	3.18
OVH SAS	4.02	Nintendo Of America inc.	1.39	GoDaddy.com, LLC	2.24
root SA	3.29	Wal-Mart Stores Inc.	1.34	Confluence Networks Inc	1.94
Cloudie Limited	2.85	GoDaddy.com, LLC	1.23	SoftLayer Technologies Inc.	1.94
SoftLayer Technologies Inc.	2.71	AT&T Services, Inc.	1.05	Leaseweb USA, Inc.	1.87
Total	60.79	Total	83.56	Total	70.51

HTTP Set-Cookie feature is used more among the premium websites with a percentage of 59.38%, compared to only 22.70% of the free content website.

HTTP Expect Certificate Transparency (Expect-CT). The Expect-CT header allows websites to report the certificate transparency requirements, which enables discovering any misuse of the website's certificates. Our analysis in Fig. 5.3b shows that the overall averages of free content and premium websites are 25.13% and 23.67%, respectively.

HTTP Cloudflare Ray-ID. The HTTP CF-RAY header contains a hashed value from Cloudflare used for information encoding to prove the source of the served resource. Fig. 5.3c shows an unnoticeable difference, except for the "Software" group, between the free content and premium websites with a percentage of (26.37%).

HTTP Strict Transport Security. The HTTP Strict-Transport-Security (HSTS) response header states that the server can only be accessed using the secure HTTPS connection instead of HTTP. Some websites allow the client to access the server through HTTP and redirect the connection to HTTPS. However, this causes several security concerns of exposing the user to a non-encrypted communication channel. Therefore, applying HSTS results in protecting the website against various man-in-the-middle attacks, such as protocol downgrade attacks, cookie hijacking, or directing a client to a malicious site [52]. Fig. 5.3d highlights a large difference in using the HSTS in the

HTTP response header by the free content and premium websites. The overall usage of the premium website is 49.44%, with three groups ("Games", "Movies", "Music") exceeding 54%. On the other hand, the overall percentage of HSTS usage in free content type is only 9.18%, with the "Music" group being the highest with 15.19%.

HTTP Content-Security-Policy (Report-To). The Content-Security-Policy (CSP) "Report-To" is an HTTP attribute that enables the client to report information such as violations, deprecation, interventions, and network errors to allocated web servers ("endpoints") [51]. Fig. 5.3e shows that the feature of reporting to endpoints applied more in the free content websites with a percentage of 27.04% (with the "Software" group being the highest with 47.13%, followed by the "Games" group with 34.21%). On the other hand, only 9.52% of the premium websites apply this feature.

HTTP X-XSS-Protection. The HTTP X-XSS-Protection response header is a feature that stops pages from loading when cross-site scripting (XSS) attacks are detected [77]. In this attack, a harmful code is injected into the website to be run by the client's browser. From our analysis in Fig. 5.3f, we uncover that 31.09% of the premium websites have enabled the X-XSS-Protection in their HTTP response header, in comparison to only 6.89% of the free content websites.

Key Takeaway: In general, our analysis of seven HTTP header attributes indicates that premium content websites' security configurations are significantly superior to their free content counterparts. This includes securing the communication channel and preventing data sniffing and man-in-the-middle attacks.

Domain Infrastructure Analysis

The website's infrastructure is the control engine that manages and organizes its structure. The infrastructure consists of a combination of a domain name system (DNS), proxy service, cloud service, CDN, and IP hosting service. In general, websites tend to deploy their servers or copies of their contents to infrastructures that protect against various types of attacks, such as distribu-

Table 5.10: The percentage of the current records for the top-10 infrastructure providers.

Free Content Websites		Premium Websites		All Websites	
Organization	Records(%)	Organization	Records(%)	Organization	Records(%)
Cloudflare, Inc.	38.00	Amazon.com, Inc.	23.12	Cloudflare, Inc.	28.58
Liquid Web, L.L.C	4.95	Cloudflare, Inc.	18.06	Amazon.com, Inc.	13.11
Team Internet AG	4.83	Google LLC	4.51	Liquid Web, L.L.C	3.01
Amazon.com, Inc.	4.34	Fastly	3.83	Google LLC	2.75
Trellian Pty. Limited	2.65	Akamai Technologies, Inc.	3.28	Team Internet AG	2.56
GoDaddy.com, LLC	1.93	Microsoft Corporation	2.19	Fastly	2.05
IP Volume inc	1.69	Akamai International B.V.	1.64	Akamai Technologies, Inc.	1.66
Netherlands	1.57	OVH SAS	1.37	GoDaddy.com, LLC	1.47
OVH SAS	1.57	GoDaddy.com, LLC	1.23	OVH SAS	1.47
Hetzner Online GmbH	1.45	Hangzhou Alibaba Advertising Co.	1.23	Trellian Pty. Limited	1.41
Total	62.97	Total	60.47	Total	58.06

tive denial-of-service attacks (DDoS). In addition, they use outsourced infrastructures to hide the websites' real identities and IP addresses.

Moreover, infrastructure organizations that provide CDN services enhance the visiting experience of a website's visitors. In this situation, the visitor would access the website by requesting the content from the nearest CDN servers rather than the actual web server. Therefore, the websites' infrastructures can affect the performance and security features. For this study, we collect the historical DNS records from 2008 to March 2021 via securityTrails [133]. The data contains the DNS records of over 1,500 websites(Premium & Free Content) in the past 13 years. Then we analyze the websites' infrastructures by providing an overview of the IP address lifetime and the infrastructure preference of the free content and premium websites.

Infrastructure Preference. Many websites rely on network infrastructures to deliver their contents to clients in a reliable, timely matter. For instance, "movies" websites necessitate pushing the video streaming to clients with higher resolution and fewer fluctuations. For better service quality and user experience, website owners often dynamically manage infrastructure resources for website scalability and geographical coverage.

Historical DNS Records. Conducting a measurement study on *Historical DNS Records* is helpful to understand the popularity of infrastructure providers. Table 5.9 shows the infrastructures that have the top-10 highest proportions of the total counts of historical DNS records since 2008.

Premium websites and free content websites show different organization preferences. Amazon, for example, contributed 13.57% of the market share in the free content websites, compared to 42.46% in the premium websites, resulting in it being the most popular infrastructure in both groups. In the rest of the infrastructures, premium websites rely on Akamai (26.32%), whereas, in the free content websites, Akamai is not among the top-10 list. On the other hand, Netherlands (9.82%) and Cloudflare (9.46%) are among the top-three infrastructure organizations among free content websites.

Current use of DNS Records. More recently, as shown in Table 5.10, Cloudflare becomes the most popular infrastructure organization among the free content websites (38.00%). On the other hand, Cloudflare is the second most popular provider within the premium group, with 18.06% of the websites associated with its infrastructures. Meanwhile, Amazon is the highest with 23.12%. Moreover, we discover that the premium and free content websites show controversial organizational preferences, particularly for the rest of the top-10 popular infrastructures. For example, if an organization is popular on free content websites (e.g., Liquid Web), we observe that it will be handling less market share within the premium websites. This observation can help in differentiating between free content and premium websites.

IP Address Stability. Based on the previous analysis, we discover that the free content and premium websites show different infrastructure preference. Since different infrastructures have different statistical characteristics, each website has its own IP address changing protocols. In practice, websites utilize a group of virtual IP addresses due to cloud hosting, load balancing, DNS proxy, and CDN usage. This causes receiving different IP addresses when a query is repeated via IP-info [67]. Our analysis using securityTrails [133] shows that a website may own multiple IP addresses at the same time, and a website may change the IP addresses several times. According to Fig. 5.4, this behavior is more prevalent in premium websites, changing their IP addresses more frequently compared to free content websites on a yearly basis. Overall, both groups of websites changed their IP addresses the most between 2018 and 2019. We note that the frequency of chang-

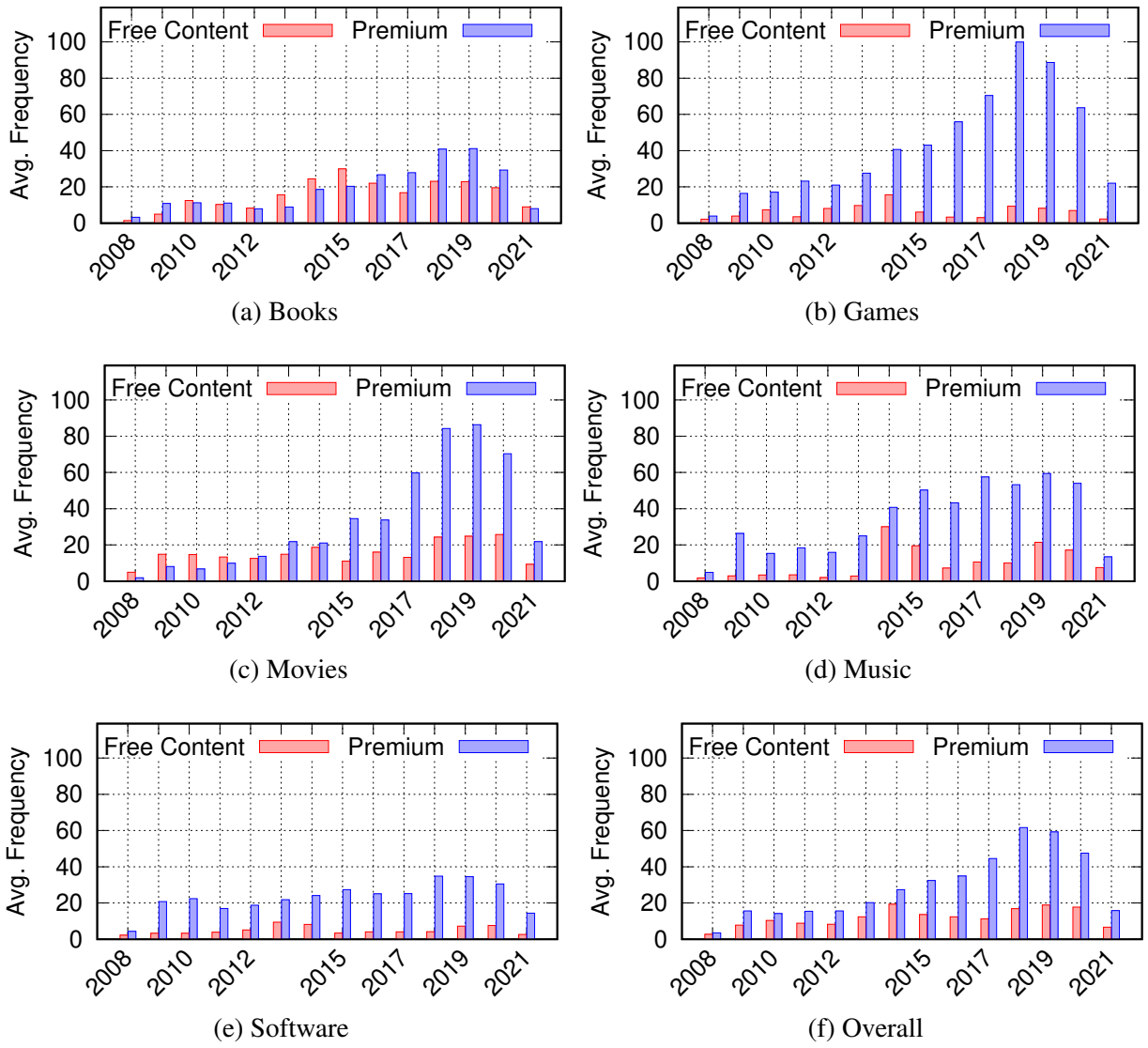


Figure 5.4: The average frequency of changing IP address within free and premium websites between 2008 and 2021.

ing IP addresses is relatively lower in 2021 since this work considers the data and information retrieved as of March 2021. Notice that the "Games" category has the most IP changing frequency for the premium websites, followed by "Movies" and "Music". Whereas the "Games" category has the lowest IP address changing frequency along with the "Software" category in the free content websites.

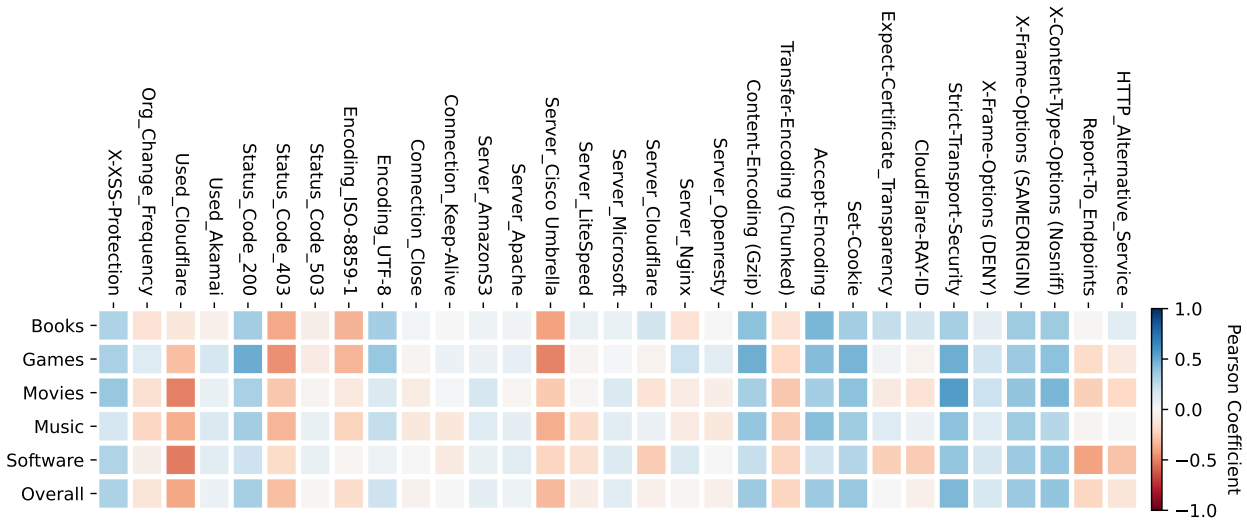


Figure 5.5: Correlation between Features and websites being free content or premium.

Key Takeaway: The infrastructural patterns among free and premium content websites are distinguishable. While premium websites tend to use Amazon, Akamai, and Google services, the free content websites heavily rely on Cloudflare. Moreover, premium websites provide higher security against DDoS attacks, by frequently changing their IP addresses.

HTTP Inter-Behavioral Patterns

In the previous analysis, we uncover that free and premium websites are indeed different, with premium websites adopting more strict security configurations than free content websites. In this section, we highlight the attributes and their associated behaviors among free and premium content websites. This allows for a better understanding of the main differences between the different website categories.

Features Correlation. First, we assume that each attribute within the HTTP header response is independent. Then, using Pearson correlation coefficient [25], we investigate whether there is a proportional relationship between each investigated attribute and whether the website is free or premium. We note that a value close to "-1" indicates that the attribute flag is more likely to be

"set" (*i.e.* enabled) in free content websites. In contrast, a value close to "1" indicates that the attribute flag is more likely to be *"set"* (*i.e.* enabled) in premium content websites.

Fig. 5.5 shows the correlation coefficient of 31 HTTP header response attributes. Among other observations, we notice that free content websites are more likely to use Cloudflare and Cisco Umbrella hosting servers, and use ISO 88591 encoding. On the other hand, premium content websites are more likely to have X-XSS-Protection enabled, with content encoding for performance optimization, alongside enabling several security-related configurations, such as strict transport layer and setting X-Frame option to *"SAMEORIGIN"*.

Features Importance. Understanding the distinguishable characteristics of free and premium websites should take into account the intra-relationships among the extracted features and attribute, an assumption not taken into consideration in the correlation study. To address this and provide a better understanding of associating certain behavior and attributes to free and premium websites, we use a gradient boosting model to extract the most distinguishable patterns among free and premium content websites.

We use 31 attributes explored in this work as features of the model. Fig. 5.6 shows the attributes that are directly correlated with the free content and premium websites. As shown, among the top-10 most important features, five features are related to security and communication protocols, which indicates that the differences between free and premium content websites are not only within the monetary scheme but also the followed security practices.

Key Takeaway: Premium content websites adopt finer-grained and diverse security configurations to reduce the adversarial threats for securing user communications. Our analysis shows that among the top-10 features used to distinguish between free and premium content websites, five are related to communication and security configurations, raising alarming concerns regarding free content websites.

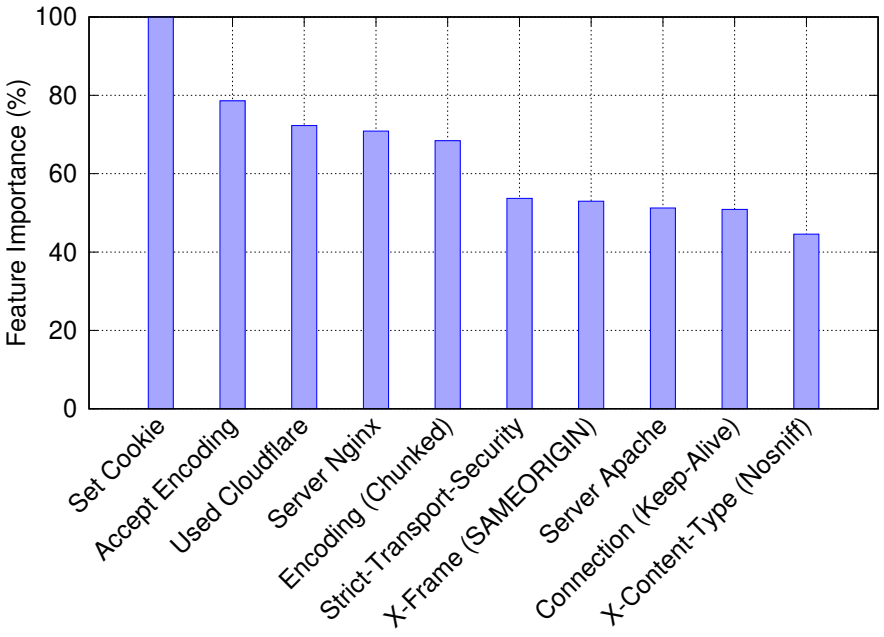


Figure 5.6: The most important features (with importance score) that can be used to differentiate between free content and premium websites.

Summary & Concluding Remarks

In this work, we have explored the potential risks associated with free content websites, resulted from following default HTTP header configurations. To do so, we have studied the structural and fundamental differences between free and premium websites by analyzing their HTTP response headers attributes and flags, in addition to the hosting domain infrastructural behavior. Our analysis shows that the most distinguishable features among free and premium content websites are the security and encoding-related HTTP header protocols. In addition, we uncover that premium websites often employ latest security configurations, reducing the potential attacks and threat surfaces. On the other hand, free content websites follow practices that can be vastly exploited, including the usage of third-party *iframes* and *redirections*, allowing HTTP communication channels, and transferring data and packets without secure encryption.

Our observations and findings confirm that free content websites are vulnerable to a variety of attacks. We raise several concerns regarding using the free content websites as we unveil that free

content websites are more relaxed in their security configurations and protocols adaptation, which directly translates to expanded attack surface in comparison with premium content websites.

CHAPTER 6: MEASURING THE PRIVACY DIMENSION OF FREE CONTENT WEBSITES THROUGH AUTOMATED PRIVACY POLICY ANALYSIS AND ANNOTATION

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. 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 [50, 28, 74, 12, 27, 106, 13, 11, 14], attracting significant traffic towards them and are placed high on websites ranking.

With the free content websites raising in popularity, as a result of being more accessible to the user, it is important to understand how 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 statements are legal statements that inform Internet users about websites and businesses data collection and information usage practices. With a semi-universal enforcement of the General Data Protection Regulation (GDPR), privacy policies became more and more elaborate

---

The work in this chapter has been published at ACM 20th Workshop on Privacy in the Electronic Society (WPES '21) held in conjunction with The ACM Conference on Computer and Communications Security (CCS) 2021, and the 10th International Workshop on Natural Language Processing for Social Media (SocialNLP '22); Companion Proceedings of the Web Conference (WWW), 2022.

and technical about how Personally Identifiable Information (PII) is collected, stored, handled, and distributed [87]. Those policies are long and complex, and it is argued that the ordinary user may not thoroughly understand the context of the policy, nor the website's actual practices [97].

Although service providers are continuously improving the readability and comprehensiveness of their policies and the disclosure of their practices, privacy policies remain difficult to understand [31, 38, 59, 97, 109]. It has been estimated that it would take the average user 201 hours to read the privacy policies encountered per year [97]. Moreover, it is unclear whether these policies are even sufficient to address the security and privacy aspects of services usage. As such, privacy and data practices of the service providers may be hidden within long, vague, and ambiguous policies, which may not be clearly disclosed to users [120, 159, 158].

A key challenge in this space is the lack of a standard format in privacy policies, which leads to ambiguity. Especially, users would be overwhelmed by both the breadth (i.e., number of policies) and depth (i.e., individual policy complexity) of those policies. While several attempts have been made towards making privacy policies easier to read, by introducing the Privacy Preference Project (P3P) [44] in 2002, privacy policies still lack a standard format as of the moment of writing this work. Motivated by that, recent studies [17, 42, 43, 166, 18, 162, 63, 158, 90] have worked on annotating privacy policies, manually and automated, to summarize and present the critical security and privacy aspects included in the privacy policy. Using state-of-the-art natural language processing and deep learning techniques, recent works [17, 42, 43, 166, 158, 63, 90, 1] effectively annotate the privacy policies content, on both sentence-level and segment-level, with high performance.

In this work, we investigate several annotation techniques for a practical automation of policy annotation, and to reduce the time and efforts required for such a task. The goal of our annotation is to provide users with easy-to-interpret high-level annotations on whether various privacy policies they encounter in their daily life meet certain requirements with respect to a broad set of privacy and security expectations. Our pipeline, called TLDR, employs advances in deep representation and machine learning.

In particular, we built an ensemble of classifiers using six word representation techniques; word mapping [90], count vectorizer [137], TF-IDF [62], Doc2Vec [78], Universal Sentence Encoder (USE) [32], and WordPiece [160], and learning algorithms; Logistic Regression (LR) [152], Support Vector Machine (SVM) [41], Random Forest (RF) [65], Convolutional Neural Networks (CNN) [165], Deep Neural Networks (DNN) [131], and Bidirectional Encoder Representations from Transformers (BERT) [47], for automating privacy policy annotation.

TLDR operates at the paragraph (segment) level, and is trained on nine categories highlighting different uses typically found in the privacy policies.

The ensemble outputs a binary decision for each category, positive and negative.

A positive outcome indicates that a segment contains information on the privacy policy category, while a negative outcome indicates that a segment does not contain such information.

Through experiments on a widely used dataset, TLDR achieves high performance in categorizing privacy policy practices, with an average F_1 score of 91%, and can highlight important segments within a privacy policy.

Then, we utilized TLDR to uncover 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.

Summary of Completed Work

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 Annotation.** We propose TLDR, a pipeline that employs various deep privacy policy representation techniques and an automated ensemble of privacy policy classifiers, leveraging advances in machine learning and natural language processing (NLP) for policy representation and classification. TLDR achieves a state-of-the-art average F-1 score of 91%.

2. **Privacy Policy Reporting.** 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.

3. **Privacy Embedded Information.** 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.

4. **Generic Privacy Policies.** 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.

Privacy Policy Annotation Pipeline

Privacy policies are diverse with no standard format. This may, in many cases, result in vague information reporting, where information is embedded within multiple sentences or even paragraphs. Extracting information regarding the privacy policy and associated practices, in most cases, is not

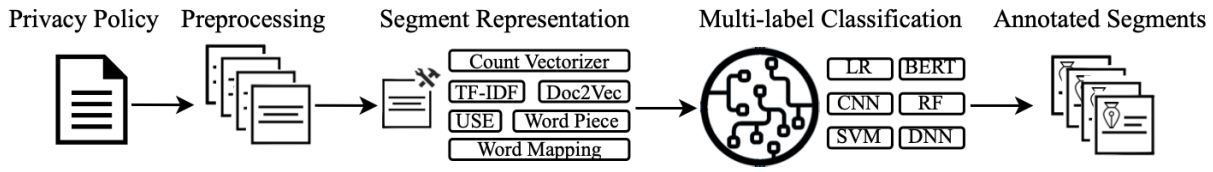


Figure 6.1: The data preprocessing and ensemble prediction pipeline. The processed segments are represented using different feature representation techniques, and then fed to the corresponding category classifier for multi-label classification.

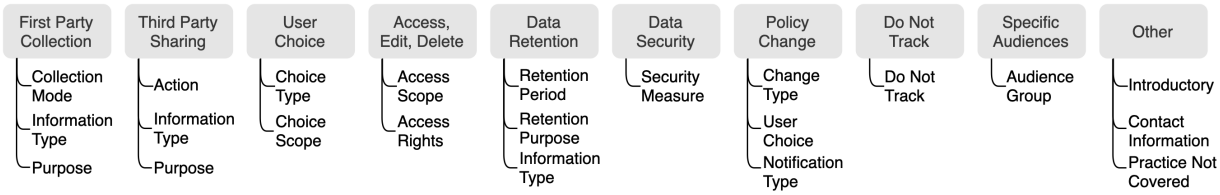


Figure 6.2: The taxonomy used by Wilson *et al.* [158] in categorizing the privacy policy practices and labeling each segment. We consider the high level nine categories in the process of building the ensemble classifier.

a straightforward task, and requires various data and text representations depending on the dimensionality in which the patterns of the policy may exist. As such, we study utilizing various text representation and pattern extraction techniques for the classification of privacy policy practices.

In particular, we developed TLDR, an ensemble of automated machine and deep learning models to extract privacy and data collection practices from the policies. The TLDR pipeline is shown in Fig. 6.1. In our pipeline, the segments are first preprocessed, then various text representation techniques are applied to extract deep representative features. Afterward, an ensemble of classifiers 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.

Ground Truth and Key Terminology. It is challenging to build a ground truth dataset for policy annotation, as that requires manual labor and domain expertise. As a baseline for privacy policy an-

Table 6.1: 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).

notation, we used the Online Privacy Policies (OPP-115) dataset, proposed by Wilson *et al.* [158]. The dataset consists of privacy policies collected from 115 websites, manually annotated by *ten law school students*. Each policy is split into paragraphs, referred to as "segments". Each segment was labeled by *three annotators*, selecting phrases associated with each privacy policy practices category. Among the 115 policies, there are 3,792 segments, averaging 33 segments per policy. All annotators worked independently and needed 72 minutes to annotate each policy on average, associating sentences with one or more category.

Fig. 6.2 shows the taxonomy used by Wilson *et al.* [158] for segment labeling. Privacy policies are categorized into high-level and low-level categories, including critical information regarding the privacy policy practices, such as "first-party data collection", "third party information sharing", and "user tracking" practices. We used the high-level categories (excluding "Others") for our automated annotator. A brief description of each category is in Table 6.1.

We note that the annotation within a sentence of a segment can be generalized to the entire segment. For instance, the segment is assigned a binary label (positive or negative) for the presence or absence of each privacy policy category. In this work, we used the segment-level labels produced by *the majority vote*: Once two annotators agree that a segment contains a privacy policy practice in a given category, we associate the segment with that category.

Privacy Policy Preprocessing. The privacy policies in OPP-115 are stored as Hypertext Markup Language (HTML) files, with *segments* contained in each privacy policy identified by the separator ("———") defined by Wilson *et al.* [158]. Each segment consists of several *sentences*, and typically discusses one or more aspects of the privacy practices of the service provider. For each segment, the *stopwords* are removed using the Gensim [122], an open-source library used for processing extensive text collections. Stopwords are common words that do not add meaning to a sentence. Words that fall in this category may include prepositions and pronouns, and can be removed without sacrificing the meaning of the sentence.

The Natural Language Tool Kit (NLTK) [91] WordNet Lemmatizer is used for *lemmatization* and *stemming*. Word lemmatization is done by grouping the different forms of a word together so that they can be analyzed as a single term. Similarly, word-stemming is done by reducing the inflection in words to their root forms, such as mapping a group of words to the same stem, even though the stem itself may not be a valid word in the language. The segment preprocessing removes the generic words and words/sub-words that do not contribute to the meaning or context of the segment, making the learning process more efficient and accurate.

Segment Representation. To find a suitable highly-discriminative representation for each category, various text representation techniques are used in TLDR: word mapping [90], count vectorizer [137], TF-IDF [62], Doc2Vec [78], Universal Sentence Encoder (USE) [32], and WordPiece [160]. In the following, we provide a brief description of those techniques as used in TLDR.

Word Mapping. Originally proposed by Liu *et al.* [90], word mapping consists of a predefined set of terms, T, which are used to represent the segment. Terms that are most frequent within sentences as labeled by the manual annotators, and indicate the presence of a privacy practice category, are considered in T. Given a segment $s \in S$, the word mapping representation is defined as:

$$V_s = \begin{cases} 1 & t \in S, \forall t \in T \\ 0 & t \notin S, \forall t \in T \end{cases} \tag{6.1}$$

Table 6.2: A predefined set of the most frequent terms from the manual annotation process, and used in the word mapping approach as the vocabulary of interest in representing each segment.

Category	Terms
First Party Collection/Use	use, collect, demographic, address, survey, service, information, require, identify
Third Party Sharing/Collection	party, share, sell, disclose, company, advertiser, provide, partner, transfer, sell, report
User Choice/Control	opt, unsubscribe, disable, choose, consent, agree, withdraw, refuse, permit
User Access, Edit, & Deletion	delete, profile, correct, account, change, update, modify, affiliate, track
Data Retention	retain, store, delete, deletion, database, participate, maintain, ensure, reserve, hold
Data Security	secure, security, seal, safeguard, protect, ensure, confidentiality, authorization, protect, practices
Policy Change	change, change privacy, policy time, current, policy agreement, time, current agreement
Do Not Track	signal, track, track request, respond, browser, advertising, content, visit, cookie, service, respond
Specific Audiences	child, California resident, European, age, parent

where V_s is the vector representing the segment s. Table 6.2 shows the most frequent words per category, considered as the vocabulary baseline of this presentation. The presence or absence of each word will be represented by "1" or "0" in a vector of size $1 \times |T|$.

Count Vectorizer. Similar to word mapping, this approach is used to convert a segment $s \in S$ to a vector V_s of terms counts. The terms are considered n-grams of the n number of sequential words, with a sliding window of one. The vocabulary of the count vectorizer contains all unique terms in the segments set S. The feature vector V_s is of size $1 \times |T|$, where V_{s_i} is calculated as follows:

$$V_{s_i} = \begin{cases} V_{s_i} + 1, \ s_i = t \\ V_{s_i}, \ s_i \neq t \end{cases}, \quad \forall s_i \in s \ \forall t \in T, \tag{6.2}$$

where s_i is the i^{th} term in the segment s. In simple terms, the generated vector V_s represents the count of each term t in the vocabulary T in the segment s.

Term Frequency–Inverse Document Frequency (TF-IDF). In the TF-IDF approach, a term t in a segment $s \in S$ of a vocabulary T is assigned a weight using the following representation model:

$$\mathsf{TF\text{-}IDF}(t, s, S) = \mathsf{TF}(t, s) + \mathsf{IDF}(t, S), \tag{6.3}$$

where $\mathsf{TF}(t, s)$ is the term frequency of term t in segment s and $\mathsf{IDF}(t, S)$ is defined as follows:

$$\mathsf{IDF}(t, S) = \log(|S|/DF(t, S)) + 1, \tag{6.4}$$

where $\mathsf{DF}(t, S)$ is the number of segments that contain the term t. The core concept of the TF-IDF is to find the inverse document frequency of each n-gram term. Using TF-IDF, a widely used technique for text representation, we explore statistics of the occurrence of n-consecutive words or terms across the segments. The probability of sequence of words, using the chain rule, is calculated as follows:

$$p(w_1, \ldots, w_T) = p(w_1) \prod_{i=2}^{T} p(w_i | w_1, \ldots, w_{i-1}). \tag{6.5}$$

Given the arbitrary length of segments, (6.5) may become infeasible to compute. To this end, we employ a commonly used approximation, where only the previous n terms are considered in computing the representation. For example, for the conditional probability of a word w_i given n-words can be described as:

$$p(w_i | w_1, \ldots, w_{i-1}) = p(w_i | w_{i-n+1}, \ldots, w_{i-1}). \tag{6.6}$$

The model in (6.6) is a probabilistic approximation realized computationally for n-grams by counting their relative occurrences in all segments using the maximum likelihood estimation:

$$\hat{p}\left(w_a\right) = \frac{c\left(w_a\right)}{N}, \hat{p}\left(w_b \mid w_a\right) = \frac{c\left(w_a, w_b\right)}{\sum_{w_b} c\left(w_a, w_b\right)} \approx \frac{c\left(w_a, w_b\right)}{c\left(w_a\right)} \tag{6.7}$$

where N is the length of the corpus (number of terms) and $c(\cdot)$ is the count terms in the corpus.

Doc2Vec. Originally proposed by Le *et al.* [78], Doc2Vec is a pre-trained model that creates a numeric representation of a document, regardless of the document length. This approach is an adaptation of the original Word2Vec approach [101], extended for documents (segments). In a nutshell, Doc2Vec finds the best numerical representation of a segment s according to the words in

that segment, using Word2Vec text representations and the continuous bag-of-words as underlying techniques for words, and by merging the words numerical representations into a single feature vector V_s.

Universal Sentence Encoder (USE). USE [32] encodes text into high-dimensional vectors that can be used in several applications, e.g., text classification. The generated embeddings are discriminative and unique, and have been shown to be effective in conducting various NLP tasks, including text classification and semantic similarity. Technically, USE takes a variable-length English text as an input and outputs a 1×512 dimensional vector representing the text. As a pre-trained model, USE is originally trained on sentences, phrases, or short paragraphs, and optimized on "greater-than word length texts". In the literature, several datasets are curated and used for the training, including the Stanford Natural Language Inference (SNLI) corpus [29], and Wikipedia [157] English articles. In essence, USE is a deep learning transformer [146] trained with a deep averaging network (DAN) encoder on billions of articles.

WordPiece. We use WordPiece [160] to represent segments for BERT [47]. For this purpose only, the original segment is preprocessed with WordPiece. WordPiece creates a vocabulary of a fixed number of words, subwords, and characters. Such a variable granularity solves the out-of-vocabulary problem by splitting unrecognized words into subwords. When no subword matches in the predefined dictionary, the candidate word is split further into characters, and mapped to the corresponding embedding.

Preprocessing and feature representation are critical in TLDR implementation to unveil the hidden patterns within each segment without the need for human labor or manual annotation.

Learning Algorithms. TLDR trains an ensemble of learning algorithms for associating segments with their corresponding privacy policy categories. Doing so reveals the abstract content of the privacy policy without the need for reading such content. This allows us to provide an overview of the content of the privacy policy, and highlight any aspects that the policy is missing. We leverage six machine and deep learning algorithms for privacy policy detection, evaluating their

effectiveness in detecting various segment-level categories. The following is a brief description of each learning algorithm.

Logistic Regression (LR). In a simple notation, LR is a statistical model that uses a logistic function to model a binary dependent variable, known as binary classification ("0" or "1"). Given an input training set (X, Y), LR learns to distinguish between positive ("1") and negative ("0") segments for each category by drawing a boundary line (assuming a linear relationship). In the higher domain, LR estimates the boundary between the positive and negative classes and optimizes the boundary by minimizing the following:

$$\mathsf{Loss}(f(X), Y) = \begin{cases} -\log(f(X)), \ Y = 1 \\ -\log(1 - f(X)), \ Y \neq 1 \end{cases}, \tag{6.8}$$

where $f(X)$ is the prediction and Y is the ground truth label.

Support Vector Machine (SVM). The SVM algorithm operates by finding a hyperplane in an N-dimensional space, where N is the length of a feature vector that distinctly classifies the segments. Toward that, SVM finds a plane that has the maximum distance between segments of both classes (positive and negative). To do so, SVM calculates the loss of each segment, defined as follows:

$$\mathsf{Loss}(X, Y, F(X)) = \begin{cases} 0, \ y \times f(x) \ \geq \ 1 \\ 1 - y \times f(x), \ y \times f(x) < 1 \end{cases}. \tag{6.9}$$

Random Forest (RF). Typically used with non-linear classification tasks, RF consists of N decision trees, each of which is trained on random features selected for individual trees. Such a technique allows for variance reduction in the output of the individual trees and mitigates the effect of noise on the training process. For RF with N decision trees, the final prediction is calculated by either a majority vote on the predictions of the decision trees, or by outputting the average prediction of all the trees, calculated as $f_{RF} = \frac{1}{N} \sum_{n=1}^{N} f_n \left(X_s'\right)$, where f_n is the prediction of the n^{th}

tree, X'_s is the vector representation of segment s for a randomly selected feature set ($X' \subset X_.$).

Convolutional Neural Networks (CNN). The CNN model is a powerful deep learning algorithm, typically used in image classification and pattern recognition. The basic unit of the CNN model is a convolution layer, which consists of several filters convolving over the input to generate feature maps. Once a feature vector is fed into a convolutional layer, it becomes abstracted to a feature map, with the shape of (feature map height) $\times$ (feature map width) $\times$ (feature map depth), with two attributes: (1) convolutional kernels defined by a width and a height (hyper-parameters), (2) the depth of the convolution filter, which is equal to the depth of the segment vector representation the feature map. In general, CNN provides excellent results in extracting patterns in higher dimensionality when the pattern location is irrelevant (in the feature space).

Deep Neural Networks (DNN). DNN consists of multiple consecutive fully connected layers, extracting deep encoded patterns. For a single layer l, the model configures the layer parameters in the learning stage. Each layer is denoted by $h^{(l)} = a(W^{(l)} \times X + b^{(l)})$, where $a(.)$ is an activation function of layer l, $W^{(l)}$ is the weights of the features from layer $l-1$ to layer l, and $b^{(l)}$ is the bias of layer l.

Bidirectional Encoder Representations from Transformers (BERT). BERT [47] is a transformer based language model that benefits from the attention mechanism provided by the transformer architecture [148]. In essence, BERT has two six-layers of encoders and decoders, each of which has the ability to learn contextual relations between words in a given context. To utilize the BERT model, each segment is preprocessed using WordPiece, where the word, subword, and character-level terms matching mechanisms are used. BERT fits a wide variety of NLP language tasks, including text classification, question answering, and named entity recognition.

Learning Algorithm Selection. The selection of the learning algorithm for each category is essential to achieve a highly accurate privacy policy annotator. For instance, learning algorithms vary in the level of pattern extraction, with CNN, DNN, SVM, and BERT extracting patterns in high dimensionality, whereas LR and RF are used for extracting statistical and low dimensionality

patterns within the segments. Moreover, both machine and deep learning are used, with techniques such as convolutional filters (CNN), information attention (BERT), and random decision trees (RF).

Acknowledging that privacy policy categories are unique, TLDR leverages the best performing data representation for category classification. The wide range of explored representations and learning algorithms is due to the diversity of categories, where treating them indiscriminately results in a reduced performance.

Evaluation and Discussion

The evaluation of TLDR is threefold. First, we assess the performance of TLDR using classification evaluation metrics on the OPP-115 for both baseline performance and to contrast TLDR's classification capabilities with the literature.

Annotation Results of TLDR. Experimental Setup. Training and Validation. For this evaluation, we use the OPP-115 dataset. We split the OPP-115 dataset into *training* and *validation* sets. In TLDR, we follow two splitting techniques: *segment-based splitting*, where segments are randomly split into 80-20 training and validation sets regardless of the associated documents, and (2) *document-based splitting*, where 80% of the documents are used for training the ensemble while the remaining 20% of the documents are used for validation.

The segments are represented using five-word representation techniques: word mapping, count vectorizer, TF-IDF, Doc2Vec, and USE. WordPiece word representation is only used with BERT.

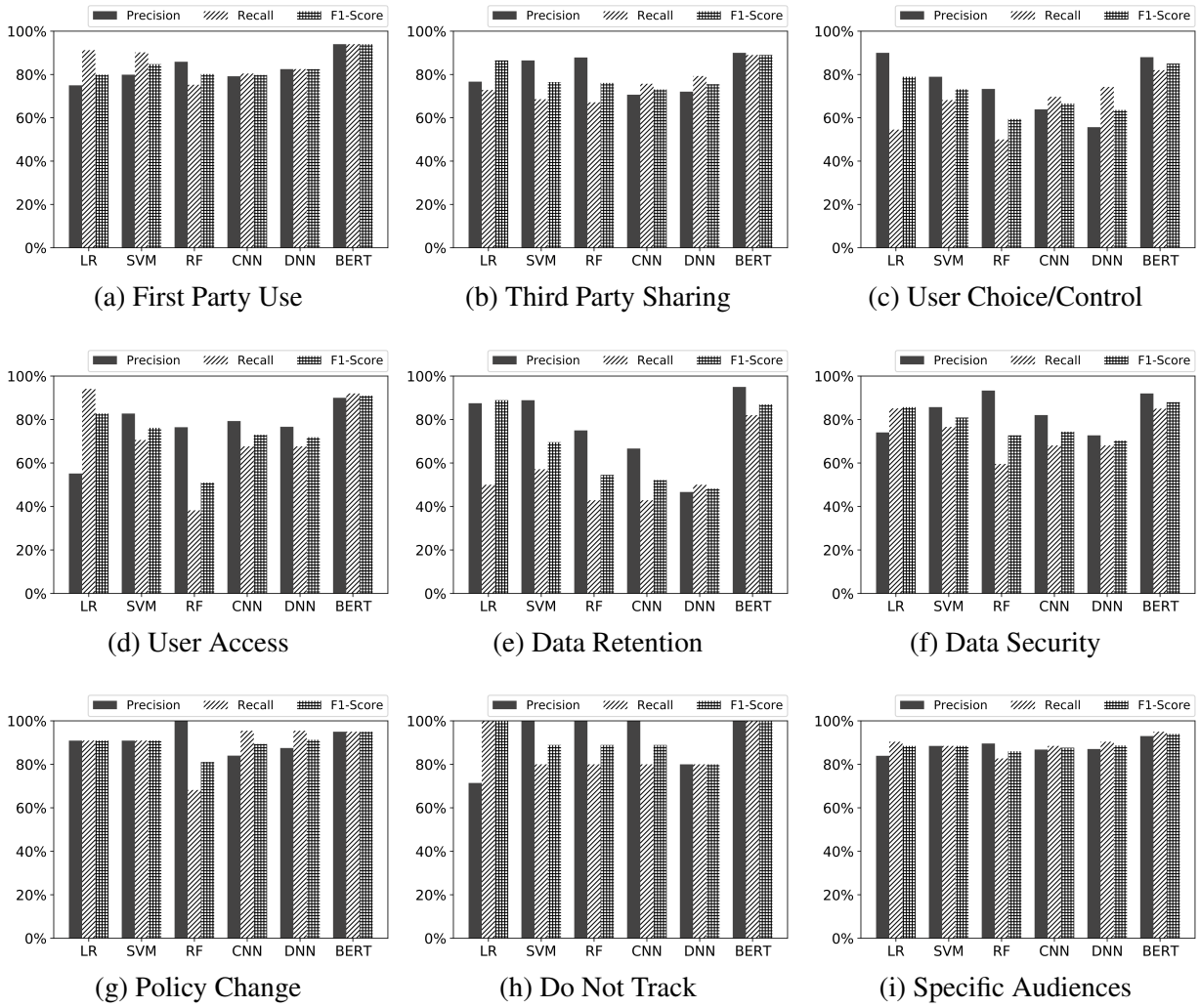
Hyper-Parameter Selection. We obtained the best parameters per word using brute force and grid search. Namely, the most frequent candidate words are vectorizer and TF-IDF are set in the range $[1,000, 5,000]$, with an increment of 1,000. We set the Doc2Vec encoding vector size to $1\times64, 128, 256, 512$, respectively, selecting the best performing vector size. The USE pre-trained model does not require any parameters, and outputs a vector of size 1×512 for each segment.

Moreover, we used the default LR and SVM configuration in the learning stage, provided by the SKLearn libarary [113]. The configuration of the LR model includes using the L_2 distance as the penalty metric, 100 maximum iterations, and balanced class weights. Similarly, the SVM model is configured with *"rbf"* kernel, and balanced class weights. For RF, we set the number of decision trees to 100, with a majority vote prediction output.

We adopt the architecture by Harkous *et al.* [63] for our CNN model, and replace the convolutional layers with fully connected layers to build the DNN model. We configure the BERT model with a 512 maximum number of words, with the number of features in the range of [1,000, 5,000] with an increment of 1,000. The BERT model is trained with learning rates of $[5 \times e^{-5}, 3 \times e^{-5}, 2 \times e^{-5}]$, and 10 training epochs. The best performing BERT model is obtained with 1,000 features and $2 \times e^{-5}$ learning rate.

Evaluation Metrics. The learning algorithms are evaluated using precision, recall, and F_1 score. The precision metric answers the question of "*How many segments labeled as positive are correct?*", and can be mathematically calculated as $P = TP/(TP + FP)$ where TP is the true positives, referring to the positive segements that were correctly classified, and FP is the false positives, referring to the negative segments that were incorrectly classified as positive by the learning model. The recall metric answers the question of "*How many positive segments were classified correctly?*", and is defined as $R = TP/(TP + FN)$ where FN is the false negatives, referring to the positive segments incorrectly classified as negative. The F_1 score is a measure that combines the precision and recall, thus called their harmonic mean, and is calculated as $F_1 = 2 \times (P \times R)/(P + R)$.

Annotation Results. Fig. 6.3 shows the evaluation of each category using the six learning algorithms of TLDR and Table 6.3 shows the best performing architecture for each category used for analysis. The BERT-based model is shown to outperform all other techniques in all categories, and therefore is used as a baseline model to build an ensemble of learning algorithms for automated privacy policy annotation.



(a) First Party Use (b) Third Party Sharing (c) User Choice/Control

(d) User Access (e) Data Retention (f) Data Security

(g) Policy Change (h) Do Not Track (i) Specific Audiences

Figure 6.3: The performance of the learning algorithms on each privacy policy category on OPP-115 dataset. The best performing learning algorithm is then used in the ensemble classifier.

Similarly, we report the best performing evaluation results of Wilson *et al.* [158] and Liu *et al.* [90] on the OPP-115 dataset using the F_1 score. As shown, except for "Specific Audiences" and "Do Not Track" categories, TLDR outperforms its counterparts by a large margin, particularly for "Data Retention".

The main intuition of this work is to provide a highly accurate privacy annotation system that is capable of highlighting segments that include information of interest to users, as recent studies showed that reading privacy policy is a time-consuming task. High F_1 score indicates that the

Table 6.3: TLDR's performance (F_1) using the best performing word representations and learning algorithms on OPP-115.

Category	**TLDR**	**Wilson [158]**	**Harkous [63]**	**Liu [90]**
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 6.4: 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

segments returned as positive by the ensemble are most likely to cover the information of the category of interest in a particular policy.

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 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 6.4.

Table 6.5: 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

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 [134], 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 6.5, 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.

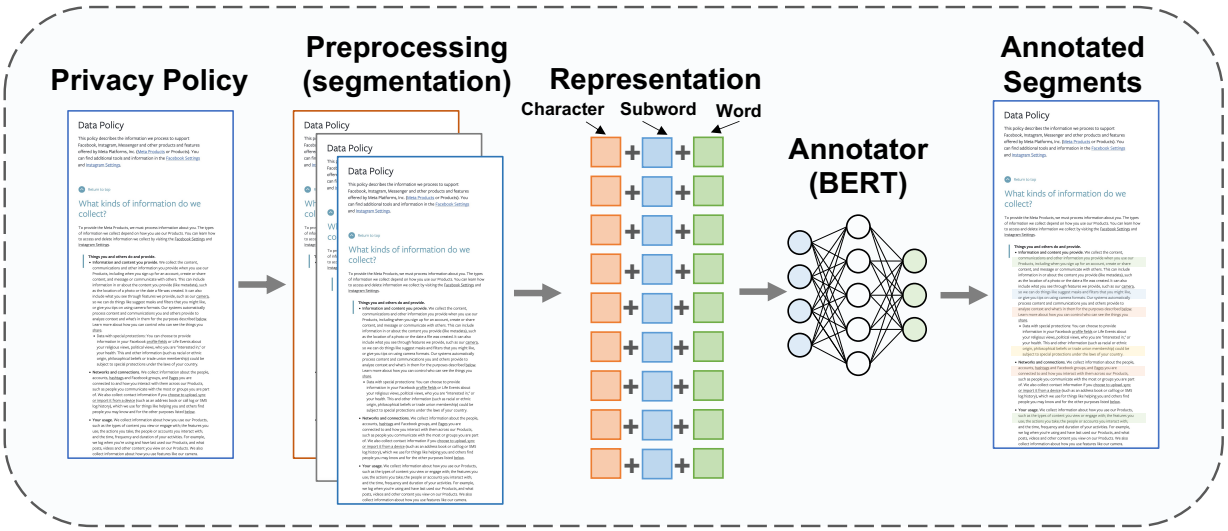


Figure 6.4: An overview of TLDR's pipeline. The processed segments are represented using different feature representation techniques, and then fed to the corresponding category classifier for multi-label classification.

For our analysis, a python library called BEAUTIFULSOUP [81] 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. Fig. 6.4 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 pre-

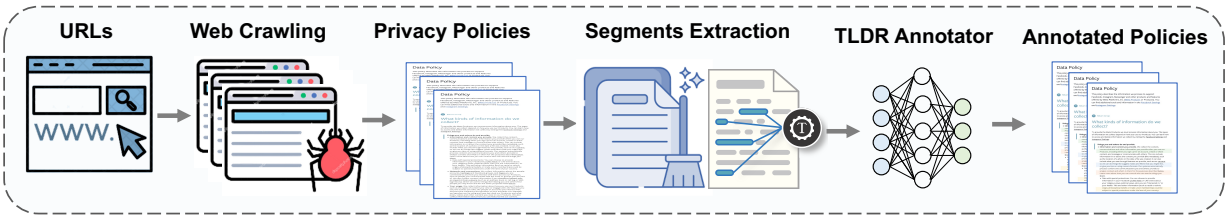


Figure 6.5: Our data collection and segment extraction pipeline, including crawling the website structure and searching for the privacy policy. Once found, paragraphs are extracted and preprocessed to extract the policy segments.

processed and represented in a manner similar to the OPP-115 segment preprocessing. Fig. 6.5 depicts the data preprocessing and representation processes in more detail.

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.

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 6.6 shows the percentage of the websites containing various privacy policy categories for free 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.

Table 6.6: The percentage of **websites** with positive segments per category for free content 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 6.7: The percentage of highlighted **segments** from free content 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 6.8: 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

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 6.6). 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.

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 6.7, 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

Table 6.9: 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

(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 6.8.

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.

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.

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. Under-

Table 6.10: Percentage of websites with positive segments per category to compare Alexa top-10,000 websites with free content and premium websites

Category	Alexa-Top10k	Free Content	Premium
First Party Use	94.64	86.90	95.73
Third Party Sharing	90.37	84.52	89.96
User Choice	74.56	52.38	79.27
User Access	62.06	50.00	65.81
Data Retention	59.34	30.95	57.26
Data Security	75.71	67.86	75.00
Policy Change	74.46	71.43	72.22
Do Not Track	19.81	12.70	21.58
Specific Audiences	72.33	67.86	74.15

standing 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 [116], 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 6.9 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).

Case Study: Alexa Top-10,000 Websites

The work we have presented so far is illuminating in the sense that it shows the privacy differential between free and premium content websites, which is an additional dimension of the cost of using the free content websites. However, a natural question would be how such a behavior differs from that of the general privacy policy content of other websites. To this end, in this section we investigate the commonly used data collection and privacy practices and how the general practices compare to their free content and premium websites. To do so, we leverage the implemented model (TLDR) towards analyzing Alexa [16] top-10,000 websites privacy policies. The Alexa top-10,000 websites represent the most visited websites by users worldwide and reflect the general data and privacy practices on the web. Analyzing such websites would uncover the common practices of popular websites and their service providers, targeting a large portion of Internet users. In particular, we highlight the privacy policy reporting practices differences between Alexa top-1000 websites compared to free content and premium websites. Our analysis focuses on how the most popular websites' privacy policies report their key information and attributes and how they introduce the important areas, including data security and user tracking. This allow for a better understanding of the trade-off between the provided services and the privacy of the users. In the following, a description of the dataset collection process and evaluation results.

Dataset Collection and Processing. Privacy Policy Extraction. We start by obtaining the privacy policies of the websites among the Alexa top-10,000 websites list. This is done using Selenium [134], an automated browser testing framework that provides extensions to emulate user interaction with browsers.

Among the top-10,000 websites, we successfully extracted the privacy policies of 5,598 websites. Table 6.11 compares the number of retrieved and validated privacy policies and the average number of segments and words per policy of Alexa top-10,000 websites, free content websites, and premium websites. Then, the privacy policies are extracted by searching the webpages within a website for terms such as *privacy policy*, and *privacy*. Once found, the associated HTML with the

Table 6.11: An overview of the number of retrieved and validated privacy policies and the average number of segments and words per policy to compare Alexa top-10,000 websites with free content and premium websites. TP=Total Policies, VP=Valid Policies, AS=Avg. Segments, AW=Avg. Words.

Category	#URLs	TP (#)	TP (%)	VP (#)	VP (%)	AS (#)	AW (#)
Free Content	834	485	58.15	252	30.22	65.89	2,806.06
Premium	728	625	85.85	468	64.29	58.93	2,784.17
Alexa-Top10k	10,000	7,345	73.45	5,598	55.98	61.79	2,764.35

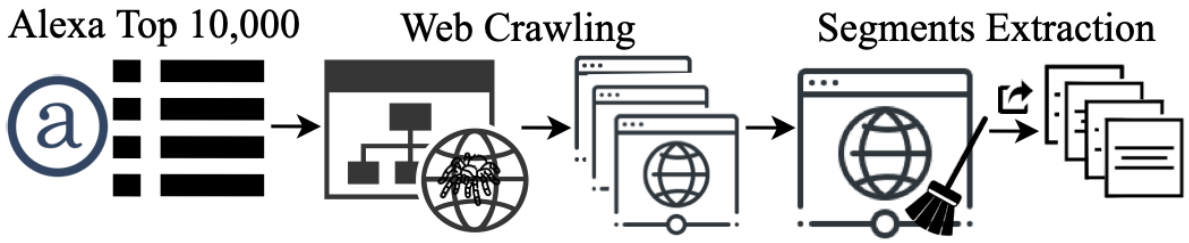


Figure 6.6: Our data collection and segment extraction pipeline, including crawling the website structure and searching for the privacy policy. Once found, paragraphs are extracted and preprocessed to extract the policy segments.

privacy policy is saved for processing. The remaining websites are either non-English or do not directly link their privacy policy in the website structure. Using the HTML paragraph tag ($< p >$), we extract all paragraphs using Beautiful Soup [81], a python library for parsing HTML and XML documents. The extracted paragraphs are considered as our potential segments. We removed all segments with less than ten words, as in most cases they are introductory sentences and do not contribute towards the privacy practices, nor contain privacy and data collection practices. The remaining segments are then associated with the extracted privacy policy for category analysis. The process of website crawling and cleaning is illustrated in Fig. 6.6.

Validation. To evaluate the correctness of the extracted policies, we manually inspect 1,000 extracted policies, and verified that 95.8% of them are correctly extracted. As such, we proceed with the extracted policies under the assumption that they are correctly extracted.

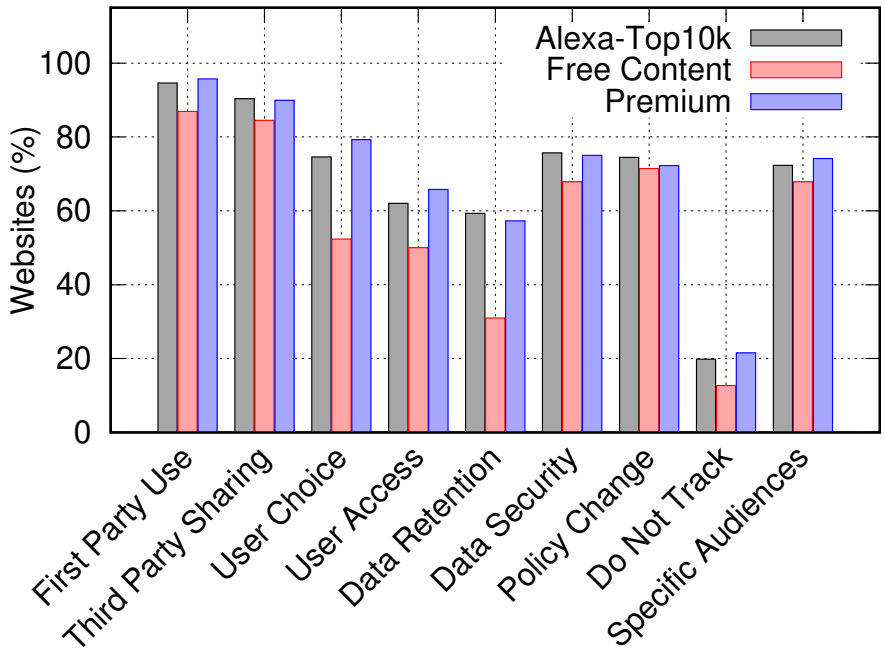


Figure 6.7: Percentage of websites with positive segments per category to compare Alexa top-10,000 websites with free content and premium websites.

Data Preprocessing & Representation. The extracted segments (345,920) are then preprocessed and represented in a way similar to free content and premium websites' segment preprocessing. We removed the stop-words, then the words in the segment were lemmatized and stemmed. We limit the configurations of the hyper-parameters to the best performing within the feature representations and learning algorithms referred to in Table 6.3. The data preprocessing and representation tasks are illustrated in Fig. 6.4.

Evaluation & Discussion. ***Privacy Practices Reporting.*** We used TLDR to extract existing privacy practices within Alexa top-10,000 websites. Fig. 6.7 shows the percentage of websites containing information regarding the policy categories to compare Alexa top-10,000 websites with free content and premium websites.

Overall, the "first-party use" and "third party sharing" categories are the most common within the privacy policies, with 95.73% of the premium websites and 94.64% of the Alexa top-10,000 websites containing first-party use information compared to 86.90% in the free content websites.

Table 6.12: The percentage of segments and words highlighted by TLDR and associated with each category to compare Alexa top-10,000 with free content and premium websites.

	Segments			Words		
Category	Alexa-Top10k	Free Content	Premium	Alexa-Top10k	Free Content	Premium
First Party Use	25.73	25.76	32.91	31.48	40.45	31.41
Third Party Sharing	16.72	16.00	15.77	20.74	19.15	21.04
User Choice	5.36	5.70	6.12	6.03	6.29	6.42
User Access	3.75	3.23	3.14	4.75	3.56	3.96
Data Retention	3.08	2.43	1.89	4.07	2.39	3.40
Data Security	3.36	3.39	2.62	4.51	2.72	4.29
Policy Change	2.32	2.22	2.65	2.92	3.07	2.84
Do Not Track	1.46	0.45	0.31	1.70	0.24	0.49
Specific Audiences	7.24	7.34	8.37	9.60	8.44	10.71
All categories	56.24	58.96	64.33	67.12	69.37	71.01

Moreover, 90.37% of the Alexa websites and 89.96% of premium websites include information regarding third-party sharing compared to 84.52% in the free content websites. On the other hand, the "do not track" category is the least common within the privacy policies, with only 19.81% of the Alexa websites and 21.58% of premium websites reporting information associated with it compared to 12.70% in the free content websites. Given that the ensemble achieves an F_1 score of 100% in this category, the results are of high confidence. Overall, free content websites have the lowest percentage of websites with positive segments per category in all categories. Whereas premium websites have the highest percentage in five categories, and Alexa top-10,000 websites are the highest in the other four categories.

Missing Information. By examining the ensemble results, we found that a large number of websites' privacy policies miss key information and attributes, by not covering important areas including data security and user tracking. This comes as a surprise, given that the extracted privacy policies are from the top-visited websites, which are potentially the subject of great interest, and their policies are the subject of great scrutiny.

Privacy Policies Embedded Information. Table 6.12 shows the percentage of segments and words highlighted by TLDR for a category of interest to compare Alexa top-10,000 with free content and

premium websites. The ensemble selects the segments that contain information regarding each category. Our findings for "All categories" show that privacy practices are covered widely in premium websites compared to Alexa top-10,000 and free content websites.

Summary & Concluding Remarks

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. In this work, we revisit the automated privacy policy annotation problem by exploring and improving the accuracy of annotation through various learning and representation techniques. In particular, we propose TLDR, a pipeline employing deep representation techniques and an ensemble of machine and deep learning-based model to automatically and accurately categorize each segment (paragraph) in the privacy policy to its corresponding high-level content category, achieving an accuracy of more than 90% in various categories. We then 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 $\approx$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 and remedies.

CHAPTER 7: CONCLUSION

Online services are categorized, based on their monetization options, into free content and premium websites. Analyzing such websites uncovers their behavioral characteristics and how they might affect the users' security or usage experience. In this dissertation, we investigate and quantify, through measurements, the potential vulnerability of such free content websites. For this purpose, we curated a dataset of 834 free content websites offering books, games, movies, music, and software. For comparison purposes, we also sampled a comparable number of premium content websites, where users need to pay for using the service for the same type of content. Through our analysis, we contribute domain-, content-, and risk-level analyses, examining and contrasting the websites' domain names, creation times, SSL certificates, HTTP requests, page size, average load time, and content type. For risk analysis, we consider and examine the maliciousness of these websites at the website- and files-level. For our modality of analysis, we explore SSL certificates' structural and fundamental differences between free and premium content websites, unveiling that 36% of the free websites' certificates have major issues. Further, to better understand the data collection and sharing practices, we propose TLDR, a pipeline that leverages advances in machine learning and natural language processing for policy representation and classification. TLDR achieves a state-of-the-art F-1 score of 91% in annotating paragraphs in the privacy policy. We utilize TLDR to 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. Encouraged by the clear differences between the two types of websites, we explore the automation and generalization of the risk modeling of the free content risky websites, showing that a simple machine learning-based technique can produce 86.81% accuracy in identifying them. Our observations raise concerns regarding the safety of using such free services from a transport standpoint and call for in-depth analysis of their risks.

APPENDIX A: PUBLICATIONS COPYRIGHT

ACM Copyright and Audio/Video Release

Title of the Work: TLDR: Deep Learning-Based Automated Privacy Policy Annotation with Key Policy Highlights
Submission ID:wpes38

Author/Presenter(s): Abdulrahman Alabduljabbar,Ahmed Abusnaina,Ulku Meteriz,David Mohaisen

Type of material:full paper

Publication and/or Conference Name: WPES '21: 20th Workshop on Privacy in the Electronic Society Proceedings

I. Copyright Transfer, Reserved Rights and Permitted Uses

* Your Copyright Transfer is conditional upon you agreeing to the terms set out below.

Copyright to the Work and to any supplemental files integral to the Work which are submitted with it for review and publication such as an extended proof, a PowerPoint outline, or appendices that may exceed a printed page limit, (including without limitation, the right to publish the Work in whole or in part in any and all forms of media, now or hereafter known) is hereby transferred to the ACM (for Government work, to the extent transferable) effective as of the date of this agreement, on the understanding that the Work has been accepted for publication by ACM.

Reserved Rights and Permitted Uses

(a) All rights and permissions the author has not granted to ACM are reserved to the Owner, including all other proprietary rights such as patent or trademark rights.

(b) Furthermore, notwithstanding the exclusive rights the Owner has granted to ACM, Owner shall have the right to do the following:

(i) Reuse any portion of the Work, without fee, in any future works written or edited by the Author, including books, lectures and presentations in any and all media.

(ii) Create a "Major Revision" which is wholly owned by the author

(iii) Post the Accepted Version of the Work on (1) the Author's home page, (2) the Owner's institutional repository, (3) any repository legally mandated by an agency funding the research on which the Work is based, and (4) any non-commercial repository or aggregation that does not duplicate ACM tables of contents, i.e., whose patterns of links do not substantially duplicate an ACM-copyrighted volume or issue. Non-commercial repositories are here understood as repositories owned by non-profit organizations that do not charge a fee for accessing deposited articles and that do not sell advertising or otherwise profit from serving articles.

(iv) Post an "Author-Izer" link enabling free downloads of the Version of Record in the ACM Digital Library on (1) the Author's home page or (2) the Owner's institutional repository;

(v) Prior to commencement of the ACM peer review process, post the version of the Work as submitted to ACM ("Submitted Version" or any earlier versions) to non-peer reviewed servers;

(vi) Make free distributions of the final published Version of Record internally to the Owner's employees, if applicable;

(vii) Make free distributions of the published Version of Record for Classroom and Personal Use;

(viii) Bundle the Work in any of Owner's software distributions; and

(ix) Use any Auxiliary Material independent from the Work. (x) If your paper is withdrawn before it is published in the ACM Digital Library, the rights revert back to the author(s).

When preparing your paper for submission using the ACM TeX templates, the rights and permissions information and the bibliographic strip must appear on the lower left hand portion of the first page.

The new ACM Consolidated TeX template Version 1.3 and above automatically creates and positions these text blocks for you based on the code snippet which is system-generated based on your rights management choice and this particular conference.

NOTE: For authors using the ACM Microsoft Word Master Article Template and Publication Workflow, The ACM Publishing System (TAPS) will add the rights statement to your papers for you. Please check with your conference contact for information regarding submitting your source file(s) for processing.

Please put the following LaTeX commands in the preamble of your document - i.e., before \begin{document}:

```
\copyrightyear{2021}
\acmYear{2021}
\setcopyright{acmcopyright}\acmConference[WPES '21]{Proceedings of the 20th Workshop on Privacy in the Electronic Society}{November 15, 2021}{Virtual Event, Republic of Korea}
\acmBooktitle{Proceedings of the 20th Workshop on Privacy in the Electronic Society (WPES '21), November 15, 2021, Virtual Event, Republic of Korea}
\acmPrice{15.00}
\acmDOI{10.1145/3463676.3485608}
\acmISBN{978-1-4503-8527-5/21/11}
```

NOTE: For authors using the ACM Microsoft Word Master Article Template and Publication Workflow, The ACM Publishing System (TAPS) will add the rights statement to your papers for you. Please check with your conference contact for information regarding submitting your source file(s) for processing.

If you are using the ACM Interim Microsoft Word template, or still using or older versions of the ACM SIGCHI template, you must copy and paste the following text block into your document as per the instructions provided with the templates you are using:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific

permission and/or a fee. Request permissions from Permissions@acm.org.

WPES '21, November 15, 2021, Virtual Event, Republic of Korea

ACM ISBN 978-1-4503-8527-5/21/11...$15.00
https://doi.org/10.1145/3463676.3485608

NOTE: Make sure to include your article's DOI as part of the bibstrip data; DOIs will be registered and become active shortly after publication in the ACM Digital Library. Once you have your camera ready copy ready, please send your source files and PDF to your event contact for processing.

☑ A. Assent to Assignment. I hereby represent and warrant that I am the sole owner (or authorized agent of the copyright owner(s)), with the exception of third party materials detailed in section III below. I have obtained permission for any third-party material included in the Work.

☐ B. Declaration for Government Work. I am an employee of the National Government of my country/region and my Government claims rights to this work, or it is not copyrightable (Government work is classified as Public Domain in U.S. only)

II. Permission For Conference Recording and Distribution

* Your Audio/Video Release is conditional upon you agreeing to the terms set out below.

I hereby grant permission for ACM to include my name, likeness, presentation and comments in any and all forms, for the Conference and/or Publication.

I further grant permission for ACM to record and/or transcribe and reproduce my presentation as part of the ACM Digital Library, and to distribute the same for sale in complete or partial form as part of an ACM product on CD-ROM, DVD, webcast, USB device, streaming video or any other media format now or hereafter known.

I understand that my presentation will not be sold separately as a stand-alone product without my direct consent. Accordingly, I give ACM the right to use my image, voice, pronouncements, likeness, and my name, and any biographical material submitted by me, in connection with the Conference and/or Publication, whether used in excerpts or in full, for distribution described above and for any associated advertising or exhibition.

Do you agree to the above Audio/Video Release? ◉ Yes ○ No

III. Auxiliary Material

Do you have any Auxiliary Materials? ○ Yes ◉ No

IV. Third Party Materials

In the event that any materials used in my presentation or Auxiliary Materials contain the work of third-party individuals or organizations (including copyrighted music or movie excerpts or anything not owned by me), I understand that it is my responsibility to secure any necessary permissions and/or licenses for print and/or digital publication, and cite or attach them below.

◉ We/I have not used third-party material.
○ We/I have used third-party materials and have necessary permissions.

V. Artistic Images

If your paper includes images that were created for any purpose other than this paper and to which you or your employer claim copyright, you must complete Part V and be sure to include a notice of copyright with each such image in the paper.

◉ We/I do not have any artistic images.
○ We/I have any artistic images.

VI. Representations, Warranties and Covenants

The undersigned hereby represents, warrants and covenants as follows:

(a) Owner is the sole owner or authorized agent of Owner(s) of the Work;

(b) The undersigned is authorized to enter into this Agreement and grant the rights included in this license to ACM;

(c) The Work is original and does not infringe the rights of any third party; all permissions for use of third-party materials consistent in scope and duration with the rights granted to ACM have been obtained, copies of such permissions have been provided to ACM, and the Work as submitted to ACM clearly and accurately indicates the credit to the proprietors of any such third-party materials (including any applicable copyright notice), or will be revised to indicate such credit;

(d) The Work has not been published except for informal postings on non-peer reviewed servers, and Owner covenants to use best efforts to place ACM DOI pointers on any such prior postings;

(e) The Auxiliary Materials, if any, contain no malicious code, virus, trojan horse or other software routines or hardware components designed to permit unauthorized access or to disable, erase or otherwise harm any computer systems or software; and

(f) The Artistic Images, if any, are clearly and accurately noted as such (including any applicable copyright notice) in the Submitted Version.

☑ I agree to the Representations, Warranties and Covenants

Funding Agents

1. National Research Foundation of Korea award number(s): NRF-2016K1A1A2912757

DATE: **09/08/2021** sent to jabbar@knights.ucf.edu at **17:09:03**

ACM Permission and Release Form

Title of non-ACM work: Automated Privacy Policy Annotation with Information Highlighting Made Practical Using Deep Representations Submission ID: **pp003**

Author(s): Abdulrahman Alabduljabbar:University of Central Florida;Ahmed Abusnaina:University of Central Florida;Ülkü Meteriz Yýldýran:University of Central Florida;David Mohaisen:University of Central Florida

Type of material: **poster; supplemental material(s)**

TITLE OF ACM PUBLICATION: CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security Proceedings

Grant Permission

As the owner or authorized agent of the copyright owner(s) I hereby grant non-exclusive permission for ACM to include the above-named material (the *Material*) in any and all forms, in the above-named publication.

I further grant permission for ACM to distribute or sell this submission as part of the above-named publication in electronic form, and as part of the ACM Digital Library, compilation media (CD, DVD, USB) or broadcast, cablecast, laserdisc, multimedia or any other media format now or hereafter known. (*Not all forms of media will be utilized.*) If your paper is withdrawn before it is published in the ACM Digital Library, the rights revert back to the author(s).

☑ Yes, I grant permission as stated above.

Multiple Author Submission Options
◉ I am submitting this permission and release form on behalf of all co-authors
○ I cannot submit this permission and release form on behalf of all co-authors

The following notice of publication and ownership will be displayed with the Material in all publication formats:

> The new ACM Consolidated TeX template Version 1.3 and above automatically creates and positions these text blocks for you based on the code snippet which is system-generated based on your rights management choice and this particular conference.
>
> *Please put the following LaTeX commands in the preamble of your document - i.e., before \begin{document}:*
>
> \copyrightyear{2021}
> \acmYear{2021}
> \setcopyright{rightsretained}
> \acmConference[CCS '21]{Proceedings of the 2021 ACM SIGSAC

Conference on Computer and Communications Security}{November 15--19, 2021}{Virtual Event, Republic of Korea}
\acmBooktitle{Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security (CCS '21), November 15--19, 2021, Virtual Event, Republic of Korea}\acmDOI{10.1145/3460120.3485335}
\acmISBN{978-1-4503-8454-4/21/11}
NOTE: For authors using the ACM Microsoft Word Master Article Template and Publication Workflow, The ACM Publishing System (TAPS) will add the rights statement to your papers for you. Please check with your conference contact for information regarding submitting your source file(s) for processing.

If you are using the ACM Interim Microsoft Word template, or still using or older versions of the ACM SIGCHI template, you must copy and paste the following text block into your document as per the instructions provided with the templates you are using:

CCS '21, November 15–19, 2021, Virtual Event, Republic of Korea

ACM ISBN 978-1-4503-8454-4/21/11.
https://doi.org/10.1145/3460120.3485335

Audio/Video Release
* Your Audio/Video Release is conditional upon you agreeing to the terms set out below.

I further grant permission for ACM to record and/or transcribe and reproduce my presentation and likeness in the conference publication and as part of the ACM Digital Library and to distribute the same for sale in complete or partial form as part of an ACM product on CD-ROM, DVD, webcast, USB device, streaming video or any other media format now or hereafter known.

I understand that my presentation will not be sold separately as a stand-alone product without my direct consent. Accordingly, I further grant permission for ACM to include my name, likeness, presentation and comments and any biographical material submitted by me in connection with the conference and/or publication, whether used in excerpts or in full, for distribution described above and for any associated advertising or exhibition.

Do you agree to the recording, transcription and distribution? ◉ Yes ○ No

Auxiliary Materials, not integral to the Work

Do you have any Auxiliary Materials? ○ Yes ◉ No

* Your Auxiliary Materials Release is conditional upon you agreeing to the terms set out below.

[Defined as additional files, video or software and executables that are not submitted for review and publication as an integral part of the Work but are supplied by the author as useful resources.] I hereby grant ACM permission to serve files containing my Auxiliary Material from the ACM Digital Library. I hereby represent and warrant that my Auxiliary (software) does not knowingly and surreptitiously incorporate malicious code, virus, trojan horse or other software routines or hardware components designed to permit unauthorized access or to disable, erase or otherwise harm any computer systems or software.

☑ I agree to the above Auxiliary Materials permission statement.

☐ This software is knowingly designed to illustrate technique(s) intended to defeat a system's security. The code has been explicitly documented to state this fact.

Third Party Materials * http://www.acm.org/publications/third-party-material

In the event that any materials used in my submission or Auxiliary Materials contain the work of third-party individuals or organizations (including copyrighted music or movie excerpts or anything not owned by me), I understand that it is my responsibility to secure any necessary permissions and/or licenses for print and/or digital publication, and cite or attach them below. Third-party copyright must be clearly stated in the caption(s) or images or in the text narrative near the object(s) in the Work and in any presentation of it and in Auxiliary Materials as applicable.

ACM offers Fair Use Guidelines at
http://www.acm.org/publications/guidance-for-authors-on-fair-use

* Small-performing rights licenses must be secured for the public performance of any copyrighted musical composition. Synchronization licenses must be secured to include any copyrighted musical composition in film or video presentations.

◉ I have not used third-party material.

○ I have used third-party materials and have necessary permissions.

Representations, Warranties and Covenants

The undersigned hereby represents, warrants and covenants as follows:

(a) Owner is the sole owner or authorized agent of Owner(s) of the Work;

(b) The undersigned is authorized to enter into this Agreement and grant the rights included in this license to ACM;

(c) The Work is original and does not infringe the rights of any third party; all permissions for use of third-party materials consistent in scope and duration with the rights granted to ACM have been obtained, copies of such permissions have been provided to ACM, and the Work as submitted to ACM clearly and accurately indicates the credit to the proprietors of any such third-party materials (including any applicable copyright notice), or will be revised to indicate such credit.

(d) The Work has not been published except for informal postings on non-peer reviewed servers, and Owner covenants to use best efforts to place ACM DOI pointers on any such prior postings;

(e) The Auxiliary Materials, if any, contain no malicious code, virus, trojan horse or other software routines or hardware components designed to permit unauthorized access or to disable, erase or otherwise harm any computer systems or software; and

(f) The Artistic Images, if any, are clearly and accurately noted as such (including any applicable copyright notice) in the Submitted Version.

Additionally, please reference the following representations that must be agreed to prior to submission and acceptance of your paper.

https://www.acm.org/publications/policies/roles-and-responsibilities#author representations

☑ I agree to the Representations, Warranties and Covenants.

Funding Agents

1. National Research Foundation of Korea award number(s):
NRF-2016K1A1A2912757

DATE: **09/08/2021** sent to mohaisen@ucf.edu at **17:09:34**

ACM Copyright and Audio/Video Release

Title of the Work: Understanding the Security of Free Content Websites by Analyzing their SSL Certificates: A Comparative Study
Submission ID:cysss09

Author/Presenter(s): Abdulrahman Alabduljabbar:University of Central Florida;Runyu Ma:George Mason University;Soohyeon Choi:University of Central Florida;Rhongho Jang:Wayne State University;Songqing Chen:George Mason University;David Mohaisen:University of Central Florida

Type of material:full paper

Publication and/or Conference Name: CySSS '22: The 1st Workshop on Cybersecurity and Social Sciences Proceedings

I. Copyright Transfer, Reserved Rights and Permitted Uses

* Your Copyright Transfer is conditional upon you agreeing to the terms set out below.

(vi) Make free distributions of the final published Version of Record internally to the Owner's employees, if applicable;

(vii) Make free distributions of the published Version of Record for Classroom and Personal Use;

(viii) Bundle the Work in any of Owner's software distributions; and

(ix) Use any Auxiliary Material independent from the Work. (x) If your paper is withdrawn before it is published in the ACM Digital Library, the rights revert back to the author(s).

When preparing your paper for submission using the ACM TeX templates, the rights and permissions information and the bibliographic strip must appear on the lower left hand portion of the first page.

The new ACM Consolidated TeX template Version 1.3 and above automatically creates and positions these text blocks for you based on the code snippet which is system-generated based on your rights management choice and this particular conference.

NOTE: For authors using the ACM Microsoft Word Master Article Template and Publication Workflow, The ACM Publishing System (TAPS) will add the rights statement to your papers for you. Please check with your conference contact for information regarding submitting your source file(s) for processing.

Please put the following LaTeX commands in the preamble of your document - i.e., before \begin{document}:

```
\copyrightyear{2022}
\acmYear{2022}
\setcopyright{acmcopyright}\acmConference[CySSS '22]{Proceedings of the 1st Workshop on Cybersecurity and Social Sciences }{May 30, 2022}{Nagasaki, Japan}
\acmBooktitle{Proceedings of the 1st Workshop on Cybersecurity and Social Sciences (CySSS '22), May 30, 2022, Nagasaki, Japan}
\acmPrice{15.00}
\acmDOI{10.1145/3494108.3522769}
\acmISBN{978-1-4503-9177-1/22/05}
```

NOTE: For authors using the ACM Microsoft Word Master Article Template and Publication Workflow, The ACM Publishing System (TAPS) will add the rights statement to your papers for you. Please check with your conference contact for information regarding submitting your source file(s) for processing.

If you are using the ACM Interim Microsoft Word template, or still using or older versions of the ACM SIGCHI template, you must copy and paste the following text block into your document as per the instructions provided with the templates you are using:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise,

or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

CySSS '22, May 30, 2022, Nagasaki, Japan

ACM ISBN 978-1-4503-9177-1/22/05…$15.00
https://doi.org/10.1145/3494108.3522769

NOTE: Make sure to include your article's DOI as part of the bibstrip data; DOIs will be registered and become active shortly after publication in the ACM Digital Library. Once you have your camera ready copy ready, please send your source files and PDF to your event contact for processing.

☑ A. Assent to Assignment. I hereby represent and warrant that I am the sole owner (or authorized agent of the copyright owner(s)), with the exception of third party materials detailed in section III below. I have obtained permission for any third-party material included in the Work.

☐ B. Declaration for Government Work. I am an employee of the National Government of my country/region and my Government claims rights to this work, or it is not copyrightable (Government work is classified as Public Domain in U.S. only)

Are any of the co-authors, employees or contractors of a National Government? ○ Yes ◉ No

---

II. Permission For Conference Recording and Distribution

* Your Audio/Video Release is conditional upon you agreeing to the terms set out below.

I hereby grant permission for ACM to include my name, likeness, presentation and comments in any and all forms, for the Conference and/or Publication.

I further grant permission for ACM to record and/or transcribe and reproduce my presentation as part of the ACM Digital Library, and to distribute the same for sale in complete or partial form as part of an ACM product on CD-ROM, DVD, webcast, USB device, streaming video or any other media format now or hereafter known.

I understand that my presentation will not be sold separately as a stand-alone product without my direct consent. Accordingly, I give ACM the right to use my image, voice, pronouncements, likeness, and my name, and any biographical material submitted by me, in connection with the Conference and/or Publication, whether used in excerpts or in full, for distribution described above and for any associated advertising or exhibition.

Do you agree to the above Audio/Video Release? ◉ Yes ○ No

III. Auxiliary Material

Do you have any Auxiliary Materials? ○ Yes ◉ No

IV. Third Party Materials

In the event that any materials used in my presentation or Auxiliary Materials contain the work of third-party individuals or organizations (including copyrighted music or movie excerpts or anything not owned by me), I understand that it is my responsibility to secure

any necessary permissions and/or licenses for print and/or digital publication, and cite or attach them below.

◉ We/I have not used third-party material.
○ We/I have used third-party materials and have necessary permissions.

V. Artistic Images
If your paper includes images that were created for any purpose other than this paper and to which you or your employer claim copyright, you must complete Part V and be sure to include a notice of copyright with each such image in the paper.
◉ We/I do not have any artistic images.
○ We/I have any artistic images.

VI. Representations, Warranties and Covenants

The undersigned hereby represents, warrants and covenants as follows:

(a) Owner is the sole owner or authorized agent of Owner(s) of the Work;

(b) The undersigned is authorized to enter into this Agreement and grant the rights included in this license to ACM;

(c) The Work is original and does not infringe the rights of any third party; all permissions for use of third-party materials consistent in scope and duration with the rights granted to ACM have been obtained, copies of such permissions have been provided to ACM, and the Work as submitted to ACM clearly and accurately indicates the credit to the proprietors of any such third-party materials (including any applicable copyright notice), or will be revised to indicate such credit;

(d) The Work has not been published except for informal postings on non-peer reviewed servers, and Owner covenants to use best efforts to place ACM DOI pointers on any such prior postings;

(e) The Auxiliary Materials, if any, contain no malicious code, virus, trojan horse or other software routines or hardware components designed to permit unauthorized access or to disable, erase or otherwise harm any computer systems or software; and

(f) The Artistic Images, if any, are clearly and accurately noted as such (including any applicable copyright notice) in the Submitted Version.

☑ I agree to the Representations, Warranties and Covenants

Funding Agents

1. National Research Foundation of Korea award number(s): NRF-2016K1A1A2912757

DATE: **03/03/2022** sent to amohaisen@gmail.com at **17:03:11**

ACM Copyright and Audio/Video Release

Title of the Work: Measuring the Privacy Dimension of Free Content Websites through Automated Privacy Policy Analysis and Annotation
Submission ID:w15c04

Author/Presenter(s): Abdulrahman Alabduljabbar:University of Central Florida;David Mohaisen:University of Central Florida

Type of material:full paper

Publication and/or Conference Name: WWW '22 Companion: The Web Conference 2022 Proceedings

I. Copyright Transfer, Reserved Rights and Permitted Uses

* Your Copyright Transfer is conditional upon you agreeing to the terms set out below.

Copyright to the Work and to any supplemental files integral to the Work which are submitted with it for review and publication such as an extended proof, a PowerPoint outline, or appendices that may exceed a printed page limit, (including without limitation, the right to publish the Work in whole or in part in any and all forms of media, now or hereafter known) is hereby transferred to the ACM (for Government work, to the extent transferable) effective as of the date of this agreement, on the understanding that the Work has been accepted for publication by ACM.

Reserved Rights and Permitted Uses

(a) All rights and permissions the author has not granted to ACM are reserved to the Owner, including all other proprietary rights such as patent or trademark rights.

(b) Furthermore, notwithstanding the exclusive rights the Owner has granted to ACM, Owner shall have the right to do the following:

(i) Reuse any portion of the Work, without fee, in any future works written or edited by the Author, including books, lectures and presentations in any and all media.

(ii) Create a "Major Revision" which is wholly owned by the author

(iii) Post the Accepted Version of the Work on (1) the Author's home page, (2) the Owner's institutional repository, (3) any repository legally mandated by an agency funding the research on which the Work is based, and (4) any non-commercial repository or aggregation that does not duplicate ACM tables of contents, i.e., whose patterns of links do not substantially duplicate an ACM-copyrighted volume or issue. Non-commercial repositories are here understood as repositories owned by non-profit organizations that do not charge a fee for accessing deposited articles and that do not sell advertising or otherwise profit from serving articles.

(iv) Post an "Author-Izer" link enabling free downloads of the Version of Record in the ACM Digital Library on (1) the Author's home page or (2) the Owner's institutional repository;

(v) Prior to commencement of the ACM peer review process, post the version of the Work as submitted to ACM ("Submitted Version" or any earlier versions) to non-peer reviewed servers;

(vi) Make free distributions of the final published Version of Record internally to the Owner's

employees, if applicable;

(vii) Make free distributions of the published Version of Record for Classroom and Personal Use;

(viii) Bundle the Work in any of Owner's software distributions; and

(ix) Use any Auxiliary Material independent from the Work. (x) If your paper is withdrawn before it is published in the ACM Digital Library, the rights revert back to the author(s).

When preparing your paper for submission using the ACM TeX templates, the rights and permissions information and the bibliographic strip must appear on the lower left hand portion of the first page.

The new ACM Consolidated TeX template Version 1.3 and above automatically creates and positions these text blocks for you based on the code snippet which is system-generated based on your rights management choice and this particular conference.

NOTE: For authors using the ACM Microsoft Word Master Article Template and Publication Workflow, The ACM Publishing System (TAPS) will add the rights statement to your papers for you. Please check with your conference contact for information regarding submitting your source file(s) for processing.

Please put the following LaTeX commands in the preamble of your document - i.e., before \begin{document}:

```
\copyrightyear{2022}
\acmYear{2022}
\setcopyright{acmcopyright}\acmConference[WWW '22 Companion]{Companion Proceedings of the Web Conference 2022}{April 25--29, 2022}{Virtual Event, Lyon, France}
\acmBooktitle{Companion Proceedings of the Web Conference 2022 (WWW '22 Companion), April 25--29, 2022, Virtual Event, Lyon, France}
\acmPrice{15.00}
\acmDOI{10.1145/3487553.3524663}
\acmISBN{978-1-4503-9130-6/22/04}
```

NOTE: For authors using the ACM Microsoft Word Master Article Template and Publication Workflow, The ACM Publishing System (TAPS) will add the rights statement to your papers for you. Please check with your conference contact for information regarding submitting your source file(s) for processing.

If you are using the ACM Interim Microsoft Word template, or still using or older versions of the ACM SIGCHI template, you must copy and paste the following text block into your document as per the instructions provided with the templates you are using:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise,

or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

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

NOTE: Make sure to include your article's DOI as part of the bibstrip data; DOIs will be registered and become active shortly after publication in the ACM Digital Library. Once you have your camera ready copy ready, please send your source files and PDF to your event contact for processing.

☑ A. Assent to Assignment. I hereby represent and warrant that I am the sole owner (or authorized agent of the copyright owner(s)), with the exception of third party materials detailed in section III below. I have obtained permission for any third-party material included in the Work.

☐ B. Declaration for Government Work. I am an employee of the National Government of my country/region and my Government claims rights to this work, or it is not copyrightable (Government work is classified as Public Domain in U.S. only)

Are any of the co-authors, employees or contractors of a National Government? ○ Yes ◉ No

---

II. Permission For Conference Recording and Distribution

* Your Audio/Video Release is conditional upon you agreeing to the terms set out below.

I hereby grant permission for ACM to include my name, likeness, presentation and comments in any and all forms, for the Conference and/or Publication.

I further grant permission for ACM to record and/or transcribe and reproduce my presentation as part of the ACM Digital Library, and to distribute the same for sale in complete or partial form as part of an ACM product on CD-ROM, DVD, webcast, USB device, streaming video or any other media format now or hereafter known.

I understand that my presentation will not be sold separately as a stand-alone product without my direct consent. Accordingly, I give ACM the right to use my image, voice, pronouncements, likeness, and my name, and any biographical material submitted by me, in connection with the Conference and/or Publication, whether used in excerpts or in full, for distribution described above and for any associated advertising or exhibition.

Do you agree to the above Audio/Video Release? ◉ Yes ○ No

III. Auxiliary Material

Do you have any Auxiliary Materials? ○ Yes ◉ No

IV. Third Party Materials

In the event that any materials used in my presentation or Auxiliary Materials contain the work of third-party individuals or organizations (including copyrighted music or movie excerpts or anything not owned by me), I understand that it is my responsibility to secure

any necessary permissions and/or licenses for print and/or digital publication, and cite or attach them below.

◉ We/I have not used third-party material.
○ We/I have used third-party materials and have necessary permissions.

V. Artistic Images
If your paper includes images that were created for any purpose other than this paper and to which you or your employer claim copyright, you must complete Part V and be sure to include a notice of copyright with each such image in the paper.
◉ We/I do not have any artistic images.
○ We/I have any artistic images.

VI. Representations, Warranties and Covenants

The undersigned hereby represents, warrants and covenants as follows:

(a) Owner is the sole owner or authorized agent of Owner(s) of the Work;

(b) The undersigned is authorized to enter into this Agreement and grant the rights included in this license to ACM;

(c) The Work is original and does not infringe the rights of any third party; all permissions for use of third-party materials consistent in scope and duration with the rights granted to ACM have been obtained, copies of such permissions have been provided to ACM, and the Work as submitted to ACM clearly and accurately indicates the credit to the proprietors of any such third-party materials (including any applicable copyright notice), or will be revised to indicate such credit;

(d) The Work has not been published except for informal postings on non-peer reviewed servers, and Owner covenants to use best efforts to place ACM DOI pointers on any such prior postings;

(e) The Auxiliary Materials, if any, contain no malicious code, virus, trojan horse or other software routines or hardware components designed to permit unauthorized access or to disable, erase or otherwise harm any computer systems or software; and

(f) The Artistic Images, if any, are clearly and accurately noted as such (including any applicable copyright notice) in the Submitted Version.

☑ I agree to the Representations, Warranties and Covenants

Funding Agents

1. National Research Foundation of Korea award number(s): 2016K1A1A2912757

DATE: **03/10/2022** sent to amohaisen@gmail.com at **11:03:48**

LIST OF REFERENCES

[1] M. Abuhamad, T. AbuHmed, A. Mohaisen, and D. Nyang. Large-scale and language-oblivious code authorship identification. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS 2018, Toronto, ON, Canada, October 15-19, 2018*, pages 101–114. ACM, 2018.

[2] U. G. S. Administration. Rules and policies - protecting pii - privacy act, May 2022.

[3] A. Alabduljabbar, A. Abusnaina, Ü. Meteriz-Yildiran, and D. Mohaisen. Automated privacy policy annotation with information highlighting made practical using deep representations. In *CCS '21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021*, pages 2378–2380. ACM, 2021.

[4] A. Alabduljabbar, A. Abusnaina, Ü. Meteriz-Yildiran, and D. Mohaisen. TLDR: deep learning-based automated privacy policy annotation with key policy highlights. In *WPES '21: Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society, Virtual Event, Korea, 15 November 2021*, pages 103–118. ACM, 2021.

[5] A. Alabduljabbar, R. Ma, A. Abusnaina, S. Alshamrani, R. Jang, S. Chen, and D. Mohaisen. No free lunch: Measuring and modeling the free content websites in the wild. Technical report, University of Central Florida, 2022.

[6] A. Alabduljabbar, R. Ma, S. Alshamrani, R. Jang, S. Chen, and D. Mohaisen. 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, 2022.

[7] A. Alabduljabbar, R. Ma, S. Choi, R. Jang, S. Chen, and D. Mohaisen. 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, 2022.

[8] A. Alabduljabbar and D. Mohaisen. Measuring the privacy dimension of free content websites through automated privacy policy analysis and annotation. In *Companion Proceedings of the Web Conference 2022*, 2022.

[9] H. Alasmary, A. Khormali, A. Anwar, J. Park, J. Choi, A. Abusnaina, A. Awad, D. Nyang, and A. Mohaisen. Analyzing and detecting emerging internet of things malware: A graph-based approach. *IEEE Internet Things J.*, 6(5):8977–8988, 2019.

[10] O. Alrawi and A. Mohaisen. Chains of distrust: Towards understanding certificates used for signing malicious applications. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11-15, 2016, Companion Volume*, pages 451–456. ACM, 2016.

[11] S. Alshamrani, M. Abuhamad, A. Abusnaina, and D. Mohaisen. 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*, pages 1–6, 2020.

[12] S. Alshamrani, A. Abusnaina, M. Abuhamad, A. Lee, D. Nyang, and D. A. Mohaisen. An analysis of users engagement on twitter during the COVID-19 pandemic: Topical trends and sentiments. In *Computational Data and Social Networks - 9th International Conference, CSoNet 2020, Dallas, TX, USA, December 11-13, 2020, Proceedings*, volume 12575 of *Lecture Notes in Computer Science*, pages 73–86. Springer, 2020.

[13] S. Alshamrani, A. Abusnaina, M. Abuhamad, D. Nyang, and D. Mohaisen. Hate, obscenity, and insults: Measuring the exposure of children to inappropriate comments in youtube. *arXiv preprint arXiv:2103.09050*, 2021.

[14] S. Alshamrani, A. Abusnaina, and D. Mohaisen. 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*, pages 1–6, 2020.

[15] I. Alsmadi and F. Mira. Website security analysis: variation of detection methods and decisions. In *2018 21st Saudi Computer Society National Computer Conference (NCC)*, pages 1–5, 2018.

[16] Amazon. Alexa top websites, May 2022.

[17] W. Ammar, S. Wilson, N. Sadeh, and N. A. Smith. Automatic categorization of privacy policies: A pilot study. *School of Computer Science, Language Technology Institute, Technical Report CMU-LTI-12-019*, 2012.

[18] B. Andow, S. Y. Mahmud, W. Wang, J. Whitaker, W. Enck, B. Reaves, K. Singh, and T. Xie. Policylint: Investigating internal privacy policy contradictions on google play. In *28th USENIX Security Symposium, USENIX*, pages 585–602, 2019.

[19] APIVoid. A framework provides JSON APIs useful for cyber threat analysis, threat detection and prevention, May 2022.

[20] H. I. Archive. Top 1,000,000: Page weight report, May 2022.

[21] S. Baek, Y. Jung, A. Mohaisen, S. Lee, and D. Nyang. Ssd-insider: Internal defense of solid-state drive against ransomware with perfect data recovery. In *38th IEEE International Conference on Distributed Computing Systems, ICDCS 2018, Vienna, Austria, July 2-6, 2018*, pages 875–884. IEEE Computer Society, 2018.

[22] P. A. Barraclough, G. Fehringer, and J. Woodward. Intelligent cyber-phishing detection for online. *Comput. Secur.*, 104:102123, 2021.

[23] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat. A comprehensive survey of ai-enabled phishing attacks detection techniques. *Telecommun. Syst.*, 76(1):139–154, 2021.

[24] A. Bates, J. Pletcher, T. Nichols, B. Hollembaek, D. Tian, K. R. B. Butler, and A. Alkhelaifi. Securing SSL certificate verification through dynamic linking. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, pages 394–405. ACM, 2014.

[25] J. Benesty, J. Chen, Y. Huang, and I. Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer, 2009.

[26] J. Berkowsky and T. Hayajneh. Security issues with certificate authorities. In *8th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2017, New York City, NY, USA, October 19-21, 2017*, pages 449–455. IEEE, 2017.

[27] A. Bhatti, H. Akram, H. M. Basit, A. U. Khan, S. M. Raza, and M. B. Naqvi. E-commerce trends during covid-19 pandemic. *International Journal of Future Generation Communication and Networking*, 13(2):1449–1452, 2020.

[28] T. Böttger, G. Ibrahim, and B. Vallis. How the internet reacted to covid-19: A perspective from facebook's edge network. In *IMC '20: ACM Internet Measurement Conference, Virtual Event, USA, October 27-29, 2020*, pages 34–41. ACM, 2020.

[29] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015.

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

[31] F. H. Cate. The limits of notice and choice. *IEEE Secur. Priv.*, 8(2):59–62, 2010.

[32] D. Cer, Y. Yang, S. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, Y. Sung, B. Strope, and R. Kurzweil. Universal sentence encoder. *CoRR*, abs/1803.11175, 2018.

[33] R. Chan, May 2022.

[34] Y. Chen and Z. Su. Guided differential testing of certificate validation in SSL/TLS implementations. In *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering, ESEC/FSE 2015, Bergamo, Italy, August 30 - September 4, 2015*, pages 793–804. ACM, 2015.

[35] T. Chung, Y. Liu, D. R. Choffnes, D. Levin, B. M. Maggs, A. Mislove, and C. Wilson. Measuring and applying invalid SSL certificates: The silent majority. In *Proceedings of the 2016 ACM on Internet Measurement Conference, IMC 2016, Santa Monica, CA, USA, November 14-16, 2016*, pages 527–541. ACM, 2016.

[36] J. Clark and P. C. van Oorschot. Sok: SSL and HTTPS: revisiting past challenges and evaluating certificate trust model enhancements. In *2013 IEEE Symposium on Security and Privacy, SP 2013, Berkeley, CA, USA, May 19-22, 2013*, pages 511–525. IEEE Computer Society, 2013.

[37] A. Cohen, N. Nissim, and Y. Elovici. Maljpeg: Machine learning based solution for the detection of malicious JPEG images. *IEEE Access*, 8:19997–20011, 2020.

[38] F. T. Commission. Protecting consumer privacy in an era of rapid change - A proposed framework for businesses and policymakers (preliminary FTC staff report). *J. Priv. Confidentiality*, 3(1), 2011.

[39] N. Confessore. Cambridge analytica and facebook: The scandal and the fallout so far, May 2022.

[40] W. W. W. Consortium et al. Platform for internet content selection (pics). *http://www. w3c. org/PICS/*, 2003.

[41] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[42] E. Costante, J. den Hartog, and M. Petkovic. What websites know about you. In *Data Privacy Management and Autonomous Spontaneous Security*, pages 146–159, 2012.

[43] E. Costante, Y. Sun, M. Petkovic, and J. den Hartog. A machine learning solution to assess privacy policy completeness: (short paper). In *Proceedings of the 11th annual ACM Workshop on Privacy in the Electronic Society, WPES*, pages 91–96, 2012.

[44] L. F. Cranor. *Web privacy with P3P - the platform for privacy preferences*. O'Reilly, 2002.

[45] R. De', N. Pandey, and A. Pal. Impact of digital surge during covid-19 pandemic: A viewpoint on research and practice. *Int. J. Inf. Manag.*, 55:102171, 2020.

[46] A. Desai, J. Jatakia, R. Naik, and N. Raul. Malicious web content detection using machine leaning. In *2017 2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 1432–1436. IEEE, 2017.

[47] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4171–4186, 2019.

[48] M. Elhadi, A. Alsoufi, A. Msherghi, E. Alshareea, A. Ashini, T. Nagib, N. Abuzid, S. Abodabos, H. Alrifai, E. Gresea, et al. Psychological health, sleep quality, behavior, and internet use among people during the covid-19 pandemic: a cross-sectional study. *Frontiers in psychiatry*, 12, 2021.

[49] S. Faghani, A. Hadian, and B. Minaei-Bidgoli. Charset encoding detection of html documents. In *AIRS*, pages 215–226. Springer, 2015.

[50] A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, C. Dietzel, D. Wagner, M. Wichtlhuber, J. Tapiador, N. Vallina-Rodriguez, O. Hohlfeld, and G. Smaragdakis. The lockdown

effect: Implications of the COVID-19 pandemic on internet traffic. In *IMC '20: ACM Internet Measurement Conference, Virtual Event, USA, October 27-29, 2020*, pages 1–18. ACM, 2020.

[51] M. Foundation. Content security policy (csp), May 2022.

[52] M. Foundation. Strict-transport-security, May 2022.

[53] M. Foundation. Using http cookies, May 2022.

[54] P. Gadient, O. Nierstrasz, and M. Ghafari. Security header fields in HTTP clients. In *21st IEEE International Conference on Software Quality, Reliability and Security, QRS 2021, Hainan, China, December 6-10, 2021*, pages 93–101. IEEE, 2021.

[55] B. Galhotra and A. Dewan. Impact of covid-19 on digital platforms and change in e-commerce shopping trends. In *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)*, pages 861–866. IEEE, 2020.

[56] A. Geniola, M. Antikainen, and T. Aura. A large-scale analysis of download portals and freeware installers. In *Secure IT Systems - 22nd Nordic Conference, NordSec 2017, Tartu, Estonia, November 8-10, 2017, Proceedings*, volume 10674 of *Lecture Notes in Computer Science*, pages 209–225. Springer, 2017.

[57] I. Ghafir, V. Prenosil, M. Hammoudeh, L. Han, and U. Raza. Malicious SSL certificate detection: A step towards advanced persistent threat defence. In *Proceedings of the International Conference on Future Networks and Distributed Systems, ICFNDS 2017, Cambridge, United Kingdom, July 19-20, 2017*, page 27. ACM, 2017.

[58] D. Gillman, Y. Lin, B. Maggs, and R. K. Sitaraman. Protecting websites from attack with secure delivery networks. *Computer*, 48(4):26–34, 2015.

[59] J. Gluck, F. Schaub, A. Friedman, H. Habib, N. M. Sadeh, L. F. Cranor, and Y. Agarwal. How short is too short? implications of length and framing on the effectiveness of privacy

notices. In *Twelfth Symposium on Usable Privacy and Security, SOUPS*, pages 321–340. USENIX Association, 2016.

[60] Google. User report-based google safe browsing for chrome and firebox, May 2022.

[61] K. Greenhill and C. Wiebrands. No library required: the free and easy backwaters of online content sharing. *VALA 2012: eM-powering eFutures*, 2012.

[62] D. J. Hand and N. M. Adams. Data mining. *Wiley StatsRef: Statistics Reference Online*, pages 1–7, 2014.

[63] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. G. Shin, and K. Aberer. Polisis: Automated analysis and presentation of privacy policies using deep learning. In *27th USENIX Security Symposium (USENIX Security 18)*, pages 531–548, 2018.

[64] F. Hecker. Setting up shop: The business of open-source software. *IEEE Softw.*, 16(1):45–51, 1999.

[65] T. K. Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282, 1995.

[66] L. Huang, A. Rice, E. Ellingsen, and C. Jackson. Analyzing forged SSL certificates in the wild. In *2014 IEEE Symposium on Security and Privacy, SP 2014, Berkeley, CA, USA, May 18-21, 2014*, pages 83–97. IEEE Computer Society, 2014.

[67] IPinfo. Comprehensive ip address data, ip geolocation api, May 2022.

[68] J. M. IV, D. Bhansali, M. Gratian, and M. Cukier. A comprehensive evaluation of HTTP header features for detecting malicious websites. In *15th European Dependable Computing Conference, EDCC 2019, Naples, Italy, September 17-20, 2019*, pages 75–82. IEEE, 2019.

[69] C. Johnson, B. Khadka, R. B. Basnet, and T. Doleck. Towards detecting and classifying malicious urls using deep learning. *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl.*, 11(4):31–48, 2020.

[70] D. Jung, S. Lee, and I. Euom. Imagedetox: Method for the neutralization of malicious code hidden in image files. *Symmetry*, 12(10):1621, 2020.

[71] G. Kates. Facebook, for the first time, acknowledges election manipulation, May 2022.

[72] R. Kiruthiga and D. Akila. Phishing websites detection using machine learning. *International Journal of Recent Technology and Engineering*, 8(2):111–114, 2019.

[73] C. Kliman-Silver, A. Hannak, D. Lazer, C. Wilson, and A. Mislove. Location, location, location: The impact of geolocation on web search personalization. In *Proceedings of the 2015 internet measurement conference*, pages 121–127, 2015.

[74] J. Koch, B. Frommeyer, and G. Schewe. Online shopping motives during the covid-19 pandemic—lessons from the crisis. *Sustainability*, 12(24):10247, 2020.

[75] A. Laughter, S. Omari, P. Szczurek, and J. Perry. Detection of malicious http requests using header and url features. In *Proceedings of the Future Technologies Conference*, pages 449–468. Springer, 2020.

[76] A. Lavrenovs and F. J. R. Melon. HTTP security headers analysis of top one million websites. In *10th International Conference on Cyber Conflict, CyCon 2018, Tallinn, Estonia, May 29 - June 1, 2018*, pages 345–370. IEEE, 2018.

[77] A. Lavrenovs and F. J. R. Melón. Http security headers analysis of top one million websites. In *2018 10th International Conference on Cyber Conflict (CyCon)*, pages 345–370. IEEE, 2018.

[78] Q. V. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1188–1196, 2014.

[79] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Neural Information Processing Systems*, pages 556–562. MIT Press, 2000.

[80] S. Lekies and M. Heiderich. On the fragility and limitations of current browser-provided clickjacking protection schemes. In *6th USENIX Workshop on Offensive Technologies, WOOT'12, August 6-7, 2012, Bellevue, WA, USA, Proceedings*, pages 53–63. USENIX Association, 2012.

[81] Leonard. Beautiful soup, May 2022.

[82] N. Leontiadis, T. Moore, and N. Christin. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 930–941, 2014.

[83] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang. Knowing your enemy: understanding and detecting malicious web advertising. In *the ACM Conference on Computer and Communications Security, CCS'12, Raleigh, NC, USA, October 16-18, 2012*, pages 674–686. ACM, 2012.

[84] Z. Li, K. Zhang, Y. Xie, F. Yu, and X. Wang. Knowing your enemy: understanding and detecting malicious web advertising. In *Proceedings of the 2012 ACM conference on Computer and communications security*, pages 674–686, 2012.

[85] J. Liang, J. Jiang, H. Duan, K. Li, T. Wan, and J. Wu. When https meets cdn: A case of authentication in delegated service. In *2014 IEEE Symposium on Security and Privacy*, pages 67–82. IEEE, 2014.

[86] T. Libert. Exposing the hidden web: An analysis of third-party HTTP requests on 1 million websites. *CoRR*, abs/1511.00619, 2015.

[87] T. Linden, H. Harkous, and K. Fawaz. The privacy policy landscape after the gdpr. *Proceedings on Privacy Enhancing Technologies*, 2020:47 – 64, 2020.

[88] D. Liu and J. Lee. CNN based malicious website detection by invalidating multiple web spams. *IEEE Access*, 8:97258–97266, 2020.

[89] F. Liu, S. Wilson, F. Schaub, and N. Sadeh. Analyzing vocabulary intersections of expert annotations and topic models for data practices in privacy policies. In *AAAI Fall Symposia*, 2016.

[90] F. Liu, S. Wilson, P. Story, S. Zimmeck, and N. Sadeh. Towards automatic classification of privacy policy text. *School of Computer Science Carnegie Mellon University*, 2018.

[91] E. Loper and S. Bird. Nltk: the natural language toolkit. *arXiv preprint cs/0205028*, 2002.

[92] L. Lu, R. Perdisci, and W. Lee. Surf: detecting and measuring search poisoning. In *Proceedings of the 18th ACM conference on Computer and communications security*, pages 467–476, 2011.

[93] M. Luo, P. Laperdrix, N. Honarmand, and N. Nikiforakis. Time does not heal all wounds: A longitudinal analysis of security-mechanism support in mobile browsers. In *Proceedings of the 26th Network and Distributed System Security Symposium (NDSS)*, 2019.

[94] M. Mangili, F. Martignon, and A. Capone. Performance analysis of content-centric and content-delivery networks with evolving object popularity. *Comput. Networks*, 94:80–98, 2016.

[95] A. S. Manjeri, R. Kaushik, M. Ajay, and P. C. Nair. A machine learning approach for detecting malicious websites using url features. In *2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 555–561, 2019.

[96] R. Masri and M. Aldwairi. Automated malicious advertisement detection using virustotal, urlvoid, and trendmicro. In *2017 8th International Conference on Information and Communication Systems (ICICS)*, pages 336–341, 2017.

[97] A. M. McDonald and L. F. Cranor. The cost of reading privacy policies. *Isjlp*, 4:543, 2008.

[98] A. Mendoza, P. Chinprutthiwong, and G. Gu. Uncovering HTTP header inconsistencies and the impact on desktop/mobile websites. In *Proceedings of the 2018 World Wide Web*

Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018, pages 247–256. ACM, 2018.

[99] S. Meredith. Facebook-cambridge analytica: A timeline of the data hijacking scandal, May 2022.

[100] U. Meyer and V. Drury. Certified phishing: Taking a look at public key certificates of phishing websites. In *Fifteenth Symposium on Usable Privacy and Security, SOUPS 2019, Santa Clara, CA, USA, August 11-13, 2019*. USENIX Association, 2019.

[101] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR*, 2013.

[102] M. A. Mishari, E. D. Cristofaro, K. M. E. Defrawy, and G. Tsudik. Harvesting SSL certificate data to identify web-fraud. *Int. J. Netw. Secur.*, 14(6):324–338, 2012.

[103] A. Mohaisen and O. Alrawi. Unveiling zeus: automated classification of malware samples. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013, Companion Volume*, pages 829–832. International World Wide Web Conferences Steering Committee / ACM, 2013.

[104] A. Mohaisen, O. Alrawi, and M. Mohaisen. AMAL: high-fidelity, behavior-based automated malware analysis and classification. *Comput. Secur.*, 52:251–266, 2015.

[105] S. P. Mulligan, W. C. Freeman, and C. D. Linebaugh, May 2022.

[106] G. Nimrod. Changes in internet use when coping with stress: older adults during the covid-19 pandemic. *The American journal of geriatric psychiatry*, 28(10):1020–1024, 2020.

[107] M. Nottingham, P. McManus, and J. Reschke. Http alternative services. *RFC 7838*, 2016.

[108] M. Nottingham and M. Thomson. Opportunistic security for HTTP/2. *RFC*, 8164:1–10, 2017.

[109] P. C. of Advisors on Science and Technology. Big data and privacy: A technological perspective. report to the president, executive office of the president. *PCAST Big Data and Privacy*, May 2014.

[110] OpenSSL. A robust, commercial-grade, full-featured Open Source Toolkit for the Secure Sockets Layer (SSL) protocol, May 2022.

[111] S. P. Privacy policies are mandatory by law, May 2022.

[112] K. Patil. Isolating malicious content scripts of browser extensions. *Int. J. Inf. Priv. Secur. Integr.*, 3(1):18–37, 2017.

[113] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[114] P. Peng, L. Yang, L. Song, and G. Wang. Opening the blackbox of virustotal: Analyzing online phishing scan engines. In *Proceedings of the Internet Measurement Conference, IMC 2019, Amsterdam, The Netherlands, October 21-23, 2019*, pages 478–485. ACM, 2019.

[115] Pingdom. Website Performance and Availability Monitoring, May 2022.

[116] Pysimilar. Computing the similarity between two string/text, May 2022.

[117] PyWebCopy. Pywebcopy: Tool for scraping and saving webpages and websites with python, May 2022.

[118] M. Raj, A. Sundararajan, and C. You. Covid-19 and digital resilience: Evidence from uber eats. *arXiv preprint arXiv:2006.07204*, 2020.

[119] M. Ramljak. Security analysis of open home automation bus system. In *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1245–1250. IEEE, 2017.

[120] A. Rao, F. Schaub, N. M. Sadeh, A. Acquisti, and R. Kang. Expecting the unexpected: Understanding mismatched privacy expectations online. In *Twelfth Symposium on Usable Privacy and Security, SOUPS*, pages 77–96. USENIX Association, 2016.

[121] R. S. Rao and A. R. Pais. Detection of phishing websites using an efficient feature-based machine learning framework. *Neural Comput. Appl.*, 31(8):3851–3873, 2019.

[122] R. Rehrek and P. Sojka. Gensim—statistical semantics in python. *Retrieved from genism. org*, 2011.

[123] K. Reitz, I. Cordasco, and N. Prewitt. Requests: HTTP for Humans, May 2022.

[124] C. E. Research. 2016 presidential campaign hacking fast facts, May 2022.

[125] R. Rivera, P. Kotzias, A. Sudhodanan, and J. Caballero. Costly freeware: a systematic analysis of abuse in download portals. *IET Inf. Secur.*, 13(1):27–35, 2019.

[126] M. Roetteler, M. Naehrig, K. M. Svore, and K. Lauter. Quantum resource estimates for computing elliptic curve discrete logarithms. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 241–270. Springer, 2017.

[127] D. Ross and T. Gondrom. HTTP header field x-frame-options. *RFC*, 7034:1–14, 2013.

[128] M. Saad, A. Khormali, and A. Mohaisen. Dine and dash: Static, dynamic, and economic analysis of in-browser cryptojacking. In *2019 APWG Symposium on Electronic Crime Research, eCrime 2019, Pittsburgh, PA, USA, November 13-15, 2019*, pages 1–12. IEEE, 2019.

[129] A. Saverimoutou, B. Mathieu, and S. Vaton. Influence of internet protocols and CDN on web browsing. In *10th IFIP International Conference on New Technologies, Mobility and Security, NTMS 2019, Canary Islands, Spain, June 24-26, 2019*, pages 1–5. IEEE, 2019.

[130] Q. Scheitle, O. Hohlfeld, J. Gamba, J. Jelten, T. Zimmermann, S. D. Strowes, and N. Vallina-Rodriguez. A long way to the top: Significance, structure, and stability of internet top lists.

In *Proceedings of the Internet Measurement Conference 2018, IMC 2018, Boston, MA, USA, October 31 - November 02, 2018*, pages 478–493. ACM, 2018.

[131] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[132] J. H. Schumann, F. von Wangenheim, and N. Groene. Targeted online advertising: Using reciprocity appeals to increase acceptance among users of free web services. *Journal of Marketing*, 78(1):59–75, 2014.

[133] SecurityTrails. Explore complete current and historical data for any internet assets. IP & DNS history, May 2022.

[134] Selenium. Seleniumhq browser automation, May 2022.

[135] V. R. L. Shen, C. Wei, and T. T. Juang. Javascript malware detection using A high-level fuzzy petri net. In *2018 International Conference on Machine Learning and Cybernetics, ICMLC 2018, Chengdu, China, July 15-18, 2018*, pages 511–514. IEEE, 2018.

[136] A. K. Singh and N. Goyal. A comparison of machine learning attributes for detecting malicious websites. In *11th International Conference on Communication Systems & Networks, COMSNETS 2019, Bengaluru, India, January 7-11, 2019*, pages 352–358. IEEE, 2019.

[137] SKLearn. SKLearn: Count Vectorizer, May 2022.

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

[139] R. Sobers. How privacy policies have changed since gdpr, May 2022.

[140] D. Solove. A brief history of information privacy law. *GW Law Faculty Publications & Other Works*, 07 2006.

[141] L. South, D. Saffo, and M. A. Borkin. Detecting and defending against seizure-inducing gifs in social media. In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pages 273:1–273:17. ACM, 2021.

[142] D. state. Domain Tools, Stats, News, Forum and Directory, May 2022.

[143] Sucuri. website security check & malware scanner, May 2022.

[144] G. Tan, P. Zhang, Q. Liu, X. Liu, C. Zhu, and F. Dou. Adaptive malicious URL detection: Learning in the presence of concept drifts. In *17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications / 12th IEEE International Conference On Big Data Science And Engineering, TrustCom/BigDataSE 2018, New York, NY, USA, August 1-3, 2018*, pages 737–743. IEEE, 2018.

[145] I. Tyagi, J. Shad, S. Sharma, S. Gaur, and G. Kaur. A novel machine learning approach to detect phishing websites. In *2018 5th international conference on signal processing and integrated networks (SPIN)*, pages 425–430. IEEE, 2018.

[146] J. Uszkoreit. Transformer: A novel neural network architecture for language understanding. *Google AI Blog*, 31, 2017.

[147] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[148] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems.*, pages 5998–6008, 2017.

[149] VirusTotal. Analyze suspicious files and URLs to detect types of malware, automatically, May 2022.

[150] G. Vrbancic, I. F. Jr., and V. Podgorelec. Swarm intelligence approaches for parameter setting of deep learning neural network: Case study on phishing websites classification. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, WIMS 2018, Novi Sad, Serbia, June 25-27, 2018*, pages 9:1–9:8. ACM, 2018.

[151] W3Techs. Usage statistics of character encodings for websites, May 2022.

[152] S. H. Walker and D. B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika*, 54(1-2):167–179, 1967.

[153] D. Y. Wang, M. Der, M. Karami, L. Saul, D. McCoy, S. Savage, and G. M. Voelker. Search+ seizure: The effectiveness of interventions on seo campaigns. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 359–372, 2014.

[154] R. Wang, Y. Zhu, J. Tan, and B. Zhou. Detection of malicious web pages based on hybrid analysis. *J. Inf. Secur. Appl.*, 35:68–74, 2017.

[155] Y. Wang, G. Xu, X. Liu, W. Mao, C. Si, W. Pedrycz, and W. Wang. Identifying vulnerabilities of SSL/TLS certificate verification in android apps with static and dynamic analysis. *J. Syst. Softw.*, 167:110609, 2020.

[156] C. Warzel and A. Ngu. Google's 4,000-word privacy policy is a secret history of the internet, May 2022.

[157] Wikipedia. Wikipedia, May 2022.

[158] S. Wilson, F. Schaub, et al. 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)*, Aug. 2016.

[159] S. Wilson, F. Schaub, R. Ramanath, N. Sadeh, F. Liu, N. A. Smith, and F. Liu. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International Conference on World Wide Web, WWW*, pages 133–143. ACM, 2016.

[160] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation, May 2022.

[161] W. Yost and C. Jaiswal. Malfire: Malware firewall for malicious content detection and protection. In *8th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference, UEMCON 2017, New York City, NY, USA, October 19-21, 2017*, pages 428–433. IEEE, 2017.

[162] R. N. Zaeem, R. L. German, and K. S. Barber. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Trans. Internet Techn.*, 18(4):53:1–53:18, 2018.

[163] L. Zhang, D. R. Choffnes, T. Dumitras, D. Levin, A. Mislove, A. Schulman, and C. Wilson. Analysis of SSL certificate reissues and revocations in the wake of heartbleed. *Commun. ACM*, 61(3):109–116, 2018.

[164] L. Zhang, D. R. Choffnes, D. Levin, T. Dumitras, A. Mislove, A. Schulman, and C. Wilson. Analysis of SSL certificate reissues and revocations in the wake of heartbleed. In *Proceedings of the 2014 Internet Measurement Conference, IMC 2014, Vancouver, BC, Canada, November 5-7, 2014*, pages 489–502. ACM, 2014.

[165] W. Zhang et al. Shift-invariant pattern recognition neural network and its optical architecture. In *Proceedings of annual conference of the Japan Society of Applied Physics*, 1988.

[166] S. Zimmeck and S. M. Bellovin. Privee: An architecture for automatically analyzing web privacy policies. In *Proceedings of the 23rd USENIX Security Symposium*, pages 1–16, 2014.