KAREL KUBICEK

AUTOMATED ANALYSIS AND ENFORCEMENT OF CONSENT COMPLIANCE

DISS. ETH NO. 29874

AUTOMATED ANALYSIS AND ENFORCEMENT OF CONSENT COMPLIANCE

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES
(Dr. sc. ETH Zurich)

presented by

KAREL KUBICEK
Magistr in Information Technology Security, Masaryk University – Brno born on 09.06.1993

accepted on the recommendation of

Prof. Dr. D. Basin, examiner
Prof. Dr. S. Bechtold, co-examiner
Prof. Dr. N. Bielova, co-examiner
Dr. H. Harkous, co-examiner

2024

Karel Kubicek: *Automated Analysis and Enforcement of Consent Compliance,*
© 2024

DOI: 10.3929/ethz-a-

For my mother Marie

ACKNOWLEDGEMENTS

This thesis would not have come to completion without the invaluable support I received from those around me. I would like to express my gratitude first to the individuals who were directly involved in my PhD journey, and then to those who provided moral support.

First, I would like to express my gratitude to David Basin for providing me the opportunity and the freedom to pursue the research I was passionate about. I value that ETH Zurich supports research that protects users rather than pursuing corporate profits. Stefan Bechtold, your ability to bridge the worlds of law and computer science greatly influenced and contributed to various aspects of this research. Amit Zac, you walked me the CS-law bridge and also showed me the way to economics; I hope our research indeed makes it to the NY Times, as you often motivate! Jakob Merane, you shared the burdens of the PhD journey with me, and I sincerely hope that you will soon feel the weight lift from your shoulders as well. Carlos Cotrini, thank you for guiding me in the academic realm, and may our future reviewers be less ruthless. Ahmed Bouhoula, I appreciate your trust in my (now our) project, and I hope my guidance could be to your help. David Bernhard, you eased me the responsibility of managing our systems and turned maintenance into exploration of possibilities. To Alexander Stremitzer and Martin Monsch, thank you for your legal expertise on our projects, along with the help of numerous research assistants who diligently annotated datasets of registration forms and emails. Aileen Nielsen, if I were to start a second PhD, it would focus on Open Source compliance!

I am also indebted to all my students. Patrice Kast, thank you for placing your trust in my supervision twice and for providing us with computational resources. Dino Bollinger, without your contributions, CookieBlock would not have resulted into what it is today; I also appreciate your friendship and patience with me. Luka Lodrant, you taught me maybe more than I did you, namely thanks for introducing me to DevOps, so I am delighted to witness your career in that topic. Florian Turati, thank you for your unwavering patience, even when fate was against our project—ultimately for the greater good. Rita Ganz, Truong Long, Elias Datler, and Laura Soldner, each of you brought unique perspectives that enriched the thesis supervision.

Studying GDPR and ePrivacy Directive compliance from the non-EU island named Switzerland would not have been possible without Joachim Posegga

and the Bright Data Initiative, to whom I am grateful for providing VPN access to the EU and beyond.

Second, I express my gratitude to all the individuals who formed the environment where I could concentrate and find relaxation when needed. My fellow F58.1-mates, I wouldn't trade you for any other office-mates (again). Tobias, it was a pleasure to share so many experiences with you, we were like a big brother to me! Reza, our candid and vivid discussions opened my eyes on several occasions. Dhruv, your positivity shines even on the rainiest days—keep that light alive! Hoang, your kindness always warmed up our office! And Xenia, you might have been only shortly our F58.1-mate, but I am very grateful for your German translation of the abstract of this thesis. I also want to express my gratitude to the other members of the InfSec group; our board game nights and discussions were truly enjoyable. To all VMI members, thank you for organizing various activities; they were a lot of fun! I hope we made a difference together. Thanks also to my friends, our outdoor adventures helped maintain my life balance, and the memories we created will endure.

Last but not least, I would like to turn to my family. Mým milým rodičům děkuji za to, že jste ze mne vychovali muže kterým teď jsem, a mým sourozencům, hlavně Oli, děkuji za to, že jsem mohl sdílet všechny problémy, včetně těch akademických. Meine Schwiegerfamilie, ich danke euch für herzliche Willkommen. Und Nina, Du hast mich so viel unterstützt, ich bin Dir dankbar für jeden Tag zusammen. Ich liebe Dich!

ABSTRACT

The collection and processing of personal data by websites have due to their ubiquity become subject to privacy regulations. In the European Union (*EU*), the ePrivacy Directive mandates that any data collection not strictly necessary for service provision must be consented to by users. The General Data Protection Regulation (*GDPR*) further stipulates that such consent must be freely given, specific, informed, and unambiguous. Personal data is any information identifying individuals, yet our primary focus lies on email addresses and browsing behavior collected by cookies.

We evaluate the extent to which the EU's privacy regulations protect users from unsolicited emails. First, we define the legal properties of registration and newsletter sign-up processes. The combination of these properties forms a decision tree for predicting potential legal violations. We evaluate these violations on a dataset of 1000 websites that we annotated with these legal properties. Second, we train machine learning models on this annotated dataset, and together with our crawler automating the sign-up process, we scale our analysis to cover 660 202 websites. We report observations from both the manual and automated analyses, identifying websites sending marketing emails without obtaining valid consent during registration or sharing users' email addresses with third parties.

We also examine how websites request consent for non-essential cookies. By investigating a sample of 29 398 websites featuring specific cookie notices, we identify eight potential consent violations in a surprising 94.7% of these websites. These violations stem from discrepancies between the information declared in the consent notice and the actual usage of cookies on the website or from the usage of tracking cookies prior to or despite the user's negative consent.

Given the high prevalence of privacy violations, we propose automated methods for enforcing privacy compliance. We developed a browser extension, CookieBlock, which employs machine learning for client-side enforcement of cookie consent. Using an XGBoost model which attains competitive accuracy compared to domain experts, CookieBlock categorizes cookies by their usage purpose and filters them according to user preferences. Finally, we suggest utilizing the automated violation detection methods from both the email and cookie projects for notification campaigns to enhance website operators' awareness of privacy regulations, or for direct enforcement by data protection authorities.

ZUSAMMENFASSUNG

Das Sammeln und Verarbeiten von personenbezogenen Daten durch Webseiten wurde aufgrund ihrer Allgegenwärtigkeit Gegenstand von Datenschutzverordnungen. Die ePrivacy-Richtlinie der Europäischen Union (*EU*) verlangt die Zustimmung der Nutzenden für sämtliches Sammeln von Daten, das für die Erbringung einer Leistung nicht strikt notwendig ist. Die Datenschutz-Grundverordnung (*DSGVO*) schreibt weiter vor, dass diese Zustimmung freiwillig, spezifisch, informiert und unmissverständlich gegeben werden muss. Personenbezogene Daten sind sämtliche Informationen, die Einzelpersonen identifizieren. Unser Hauptfokus liegt hierbei auf E-Mail-Adressen und dem Onlineverhalten das von Browser-Cookies gesammelt wird.

Wir werten das Ausmass aus, mit dem die Datenschutzverordnungen der EU, die Nutzenden vor unerwünschten E-Mails schützen. Zuerst definieren wir die rechtlichen Eigenschaften einer Registrierung und dem Abonnieren eines Newsletters und kombinieren diese zu einem Entscheidungsbaum, um potenzielle rechtliche Verstösse zu identifizieren. Wir evaluieren diese Verstösse auf einem Datensatz von 1000 Webseiten, die wir mit den rechtlichen Eigenschaften annotierten. In einem zweiten Schritt trainieren wir ein Machine Learning Modell auf diesem annotierten Datensatz. Zusammen mit unserem Crawler (Programm zur Datensammlung), der den Registrierungsprozess automatisiert, skalieren wir unsere Analyse auf über 660 202 Webseiten. Wir melden Beobachtungen sowohl während unserer manuellen als auch der automatischen Analyse von Webseiten, die Marketing-E-Mails ohne gültige Zustimmung verschicken oder E-Mail-Adressen mit Dritten teilen.

Wir untersuchen, inwiefern Webseiten die Zustimmung nicht-essentieller Cookies anfordern. Unsere Untersuchung an 29 398 Webseiten, welche jeweils spezifische Cookie-Zustimmungen beinhalteten, konnten erstaunlicherweise acht potenzielle Zustimmungsverstösse an 94.7% der Webseiten identifizieren. Diese Verstösse ergeben sich durch Unstimmigkeiten zwischen den Informationen bei den Cookie-Zustimmungen und der tatsächlichen Verwendung von Cookies auf der Webseite, oder durch die Nutzung sogenannter Tracking-Cookies bevor oder ungeachtet der negativen Zustimmung.

Angesichts der höheren Häufigkeit von Datenschutzverstössen, schlagen wir automatische Methoden zur Durchsetzung der Einhaltung des Datenschutzes vor. Wir entwickelten eine Browser-Erweiterung, CookieBlock, die mit Hilfe von Machine Learning auf der Clientseite Cookie-Zustimmungen durchsetzt. Mithilfe eines XGBoost Modells, dessen Genauigkeit mit dem von Fachexperten vergleichbar ist, kategorisiert CookieBlock Cookies aufgrund ihrer Nutzungszwecke und filtert sie nach Nutzerpräferenz. Zuletzt schlagen wir die Nutzung der automatischen Methode zum detektieren von Verstössen aus den E-Mail- und Cookie-Projekt für Benachrichtigungskampagnen vor, um das Bewusstsein von Webseitenbetreibenden zu Datenschutzvorschriften zu vergrössern oder für die Durchsetzung von Vorschriften durch Datenschutz Behörden.

TABLE OF CONTENTS

1 INTRODUCTION

With the transition into the information age, data privacy has emerged as a paramount concern. The collection and processing of private personal data constitute foundational elements of internet-based businesses. This data collection comes in various forms, with our focus centered on cookies and email addresses, while recognizing the existence of more covert tracking methods such as browser fingerprinting [1]. Cookies are commonly employed by websites to track user behavior for the purpose of delivering targeted content and advertisements. This tracking is gaining in popularity, in 2012, approximately 80% of websites engaged in this practice [2], which increased to 90% by 2019 [3]. Alongside, email addresses serve as highly reliable identifiers, typically required during website registrations. Beyond tracking, email addresses are also leveraged for delivering personalized marketing, even when users are not actively engaged with the website.

The operators of websites employing these intrusive practices often exploit what is referred to as the "privacy paradox" [4]. It states that while users express significant concerns about their privacy, they frequently fail to take proactive steps to protect it. Given the scale of data collection, this can result in harms of individuals, but also of the whole society, such as in the case of Cambridge Analytica, which used illegally collected private data to influence elections around the world including the US 2016 presidential election [5]. Events like this are a call for action, namely to strengthen both the privacy regulations and their enforcement. Such a legislative shift is particularly evident within the European Union (*EU*), where policymakers took action to address the power imbalance between website operators and users regarding privacy.

This work focuses on two recent and active regulatory frameworks designed to counter emerging data collection practices: the 2002 ePrivacy Directive (*ePD*) with its 2012 amendment, and the 2018 General Data Privacy Regulation (*GDPR*). The ePD, sometimes referred to as 'The Cookie Law,' specifically targets technologies related to digital communication, including cookie usage and electronic communication. It mandates user consent for all data usage that is not strictly necessary for the functioning of a website. The GDPR defines this consent and requires that it must be freely given, specific, informed, and unambiguous. Although the GDPR permits alternative legal bases for personal data collection and processing, recent rulings and guidelines have underscored that active opt-in consent constitutes the legal base of a significant majority of private data collection and processing present on modern websites.

1.1 EMAILS

To investigate the compliance of consent with sending marketing emails, we must assess both the registration or newsletter subscription process and the received emails. For the registration process, valid consent is considered given when users actively mark a checkbox that explicitly asks for consent to receive marketing emails, accompanied by a clear explanation of its implications. For forms exclusively dedicated to newsletter subscriptions, the purpose of receiving marketing emails is implicit in the form's wording, and therefore including a checkbox becomes redundant. Nevertheless, in all cases, websites should respond to form submission by sending an activation email to verify the user's ownership of the registered address through a *double-opt-in* procedure. Furthermore, even when consent is granted, the ePrivacy Directive mandates the provision of user-friendly mechanisms for consent withdrawal. This implies that every marketing email should include an option to unsubscribe from the mailing list.

There has been limited research examining the compliance of these processes. Englehardt et al. [6] and Mathur et al. [7] investigated email privacy by subscribing to US e-commerce and political campaign newsletters. Their findings revealed that 30% of senders shared email addresses with third parties, engaged in email tracking, and employed manipulative political campaign tactics. However, these studies were conducted exclusively in the US predating the enactment of stronger privacy laws like the California Consumer Privacy Act (*CCPA*), and did not consider the EU's perspective. Senol et al. [8] automated the interaction with website forms to investigate data leakage prior to form submission, which violates the GDPR, as data collection occurs without consent on 3% of websites. Oh et al. [9] examined compliance of consent with privacy policies and terms of service, revealing that the majority of websites failed to meet GDPR requirements.

We contribute to empirical research by conducting two studies to inspect marketing consent compliance. We initially conduct a pilot study involving manual registration, a task we fully automate in a subsequent study.

We define 21 legal properties of registration forms, such as the presence of a checkbox for consenting to marketing emails. Based on these legal properties, we construct decision trees to predict potential consent violations within these forms. Additionally, we analyze the content of emails, identifying instances of failure to collect recipient consent through the double-opt-in procedure, the presence of user-provided passwords within email bodies, the absence of unsubscribe options and legal notices, and instances of email address sharing with third parties.

In the pilot study, seven research assistants registered on 1000 websites preselected as likely to contain registration forms. They successfully registered on two-thirds of these websites, with each registration accompanied by the annotation of the 21 legal properties by pairs of annotators. In cases of disputes, a third annotator resolved the discrepancies. This pilot study led to the identification of 119 consent violations and 162 email violations.

Subsequently, we scale up the pilot study using a crawler that automates the registration or newsletter subscription and using machine learning (*ML*) models that predict the 21 legal properties of forms and additional properties of emails. We evaluate both the crawler and violation detection on a crawl of 660k websites, registering or signing up for newsletters on 5.9% of them, which more than doubles the registration and sign-up rate compared to prior work by Drakonakis et al. [10]. Using ML classification of email types, we assess the presence of the double-opt-in email verification, finding that 59.8% of websites fail to follow this procedure. Since we generate a unique email address for each registered website, we discover that in 14.5% of the cases, we receive emails from domains other than the domain of registration. Our filtering suggests that up to 7.2% of websites share our email address with third parties, often without declaring this either on the registration form or in the privacy policy.

We then evaluate the accuracy of the automatically detected privacy violations. Unfortunately, this manual evaluation shows high false positive rates due to the complex propagation of misclassifications in our decision trees. However, this manual evaluation, which also generates more data usable for training better models, represents the first step toward the main goal of the project, which is enforcing the regulations of electronic marketing. Reliable methods detecting privacy violations and generating incriminating evidence would help the understaffed and overloaded data protection authorities (*DPAs*) to check violations faster using the evidence.

1.1.1 *Contributions*

In the manual pilot study, we have made the following contributions.

LEGAL TAXONOMY FOR MARKETING. We summarize the legal requirements for sending marketing emails based on the German implementation of the ePrivacy Directive. We further propose decision trees for detecting potential violations of the ePrivacy Directive's opt-in requirement and the GDPR's notion of consent in website registration forms.

VIOLATION STATISTICS. We observe at least one potential violation in 22% of the websites. Namely, 17.3% of the websites send marketing emails without obtaining proper consent and 17.7% of services send emails that are potential violations because content required by law is missing, passwords are sent in plaintext, or the service shares the email address with third parties.

ANNOTATED DATASETS. We offer the privacy research community a dataset with annotations of the legal properties of registration forms for 1000 websites. We release these annotations, the registration page source code, and post-processed features of the registration form upon request. We also release a dataset of 5000 emails labeled with their purpose. Both datasets are suitable for other studies, such as email or registration form tracking analysis, or marketing email content analysis.

As the main author of this study [11], my contributions include: designing the annotation tool and overseeing the annotation of legal properties across 1000 websites; managing the mailserver, confirming double-opt-in registrations, and annotating over 5k emails; and implementing legal decision procedures, processing all results, and presenting the findings.

In the subsequent automation study, we have made the following contributions.

REGISTRATION CRAWLER. We develop a crawler that achieves more than double the rate of registration and sign-up for newsletters than prior work. Our crawler enables the automated analysis of those parts of websites that require prior user authentication, enabling previously impossible privacy and security studies at scale.

VIOLATION DETECTION. We automate the detection of privacy and security violations using ML models that allow fully self-contained processing of crawled registration forms and received emails.

LARGE-SCALE EVALUATION. We present new results on how tens of thousands of websites potentially violate GDPR consent requirements in the user registration process. Namely 42.5%, that is 12 417, of websites send marketing emails despite insufficient consent. This demonstrates the usefulness of our crawler in analyzing the security and privacy of the registration process.

As the main author of this study which is currently submitted to The Web Conference 2024, I contributed by: supervising the development and co-developing the crawler; automating the prediction of legal properties using machine learning and heuristics; conducting manual evaluation of ML results and confirming registrations in subsequent crawls, leading to annotation of an additional 11k emails; and overseeing the execution of crawls, processing results, and drafting the entire paper.

1.2 COOKIES

Cookie notices represent one of the most apparent impacts of the EU's privacy regulations. Their ubiquity is often perceived as bothersome by users, with news media even dubbing them the 'biggest failure of the GDPR' [12]. Consequently, the aspect of compliance with these notices has attracted significant attention from researchers. Kampanos et al. [13] discovered that, in a sample of approximately 17k websites from the UK and Greece, fewer than half displayed a cookie notice to users, a notable contrast to the 90% of all websites employing tracking cookies. Moreover, websites with cookie consent notices frequently fail to adhere to even fundamental rules. Matte et al. [14] found that among a sample of 1426 selected websites, 10% recorded affirmative consent before users had made a choice, and 5% did not respect users' opt-out preferences. Trevisan et al. [15] reported that 49% of the 36k inspected websites set profiling cookies before obtaining user consent. An additional direction of research has explored deceptive design patterns, strategies employed to coerce or nudge users into consenting against their will. Nouwens et al. [16] observed that nearly 90% of 680 examined websites using supported Consent Management Platforms (*CMPs*) failed to meet GDPR requirements for valid consent or nudged users. Numerous user studies have demonstrated the effectiveness of these deceptive patterns in influencing user behavior [17–22].

Our research extends and improves upon past studies, confirming the lack of GDPR compliance. We assess the accuracy of the information displayed on cookie banners using a dataset collected from nearly 30k websites. Specifically, we identify incorrect category assignments, misleading cookie expiration times, and evaluate the overall completeness of the consent mechanism. We introduce six novel methods to detect potential GDPR violations and extend two methods from prior works. Among the selected domains, we find that 94.7% contained at least one potential violation. On 36.4% of the websites, we identified at least one cookie with an incorrectly assigned purpose, and on 85.8% was at least one cookie with a missing declaration or purpose. 69.7% of the sites assumed positive consent before it was given, and 66.4% of sites suitable for such an analysis created cookies despite receiving negative consent.

Our measurements reveal a wider prevalence of violations of EU privacy regulations than previously acknowledged, rendering regulatory authorities unable to effectively keep pace. We therefore provide users with a tool to enforce cookie consent on their web clients without the need for regulatory intervention. We have developed a browser extension, CookieBlock, which categorizes cookies by purpose and removes those that the user rejects. This empowers users to eliminate over 90% of privacy-invasive cookies without having to trust cookie banners or CMPs. Previous attempts to provide users with such control, like the P3P standard [23], failed due to a lack of willingness of website administrators to implement the required functionality. We circumvent this issue by not relying on website cooperation at all. CookieBlock is available on all major browsers and has over 13k installations.

1.2.1 *Contributions*

This work has the following main contributions.

VIOLATION STATISTICS. We identify inaccurate information in cookie banners, and apply this to a sample of approximately 30k websites, finding potential GDPR violations for 94.7% of them.

COOKIE CATEGORIZATION. We present a machine-learning classifier that infers purposes from cookies, reaching a performance that is comparable to that of human experts.

CLIENT-SIDE ENFORCEMENT. We develop a browser extension that automatically removes cookies according to users' preferences, which, unlike comparable approaches, is applicable to any cookie and does not require websites to cooperate.

INFRASTRUCTURE. We release our tools customized for researchers to perform follow-up studies and for web administrators to enable them to verify and improve the cookie consent compliance of their websites.

This publication [24] stemmed from my supervision of Dino Bollinger and extending his Master's thesis [25]. My contributions included: conceiving the initial idea of automating cookie consent using machine learning trained using available data labeled by selected CMPs; proposing methods to extract tabular features from cookie data; and defining privacy violations observable from the collected data.

This research direction also led to several subsequent works in which I contributed, although they are not included in this thesis. Notably, Bouhoula et al. [26] automated violation measurements for all types of cookie notices. Turati et al. [27] investigated and broke the privacy properties of FLoC, Google's "private-targeted" alternative to tracking cookies. Finally, Schöni et al. [28] conducted a user study of CookieBlock, identifying adoption challenges for such privacy-enhancing browser extensions among the general public.

1.3 THESIS STRUCTURE

The structure of this thesis is as follows:

IN CHAPTER 2, I provide legal background for our studies.

IN CHAPTER 3, I intertwine the studies regarding email consent, including both manual pilot and subsequent automation. This presentation provides valuable context and facilitates a more meaningful comparison between manual and automated results.

IN CHAPTER 4, I present our measurements related to cookie consent violations and introduce the privacy-enhancing browser extension CookieBlock, designed to address these issues from the client-side.

IN CHAPTER 6, I describe the literature overview related to both studies, including additional literature that emerged after our research publications.

IN CHAPTER 7, I conclude the thesis with a discussion of the findings and their implications.

2 EU PRIVACY LAWS

The interplay of the Privacy and Electronic Communications Directive 2002 (also known as the ePrivacy Directive, *ePD*) [29] and the General Data Protection Regulation (*GDPR*) [30] is complex and has been studied by legal scholars [31]. This complexity is further amplified when combined with local regulations. As a directive, the ePrivacy allows for varying interpretations by EU member countries,[1] and both laws also provide local data protection authorities certain freedoms in enforcement. To facilitate conciseness for the computer science audience of this thesis, we adopt a unified interpretation across all EU countries. The local guidelines, court cases, and other legal references we cite are assumed to be applicable in the whole EU. Article 3 of the GDPR defines the territorial scope to any user present in the EU, even when the visited website is hosted by a non-EU country.

In this chapter, we present a simplified explanation of the regulatory requirements.

2.1 KEY COMMUNICATING ENTITIES

Article 4 of the GDPR defines the following entities involved in data collection and processing. We establish connections between these entities within terminology used in CS and law. Despite narrowing this to the context of websites, the mapping is not strictly one-to-one.

DATA SUBJECT refers to the person identified by the data. Given the focus of this thesis on investigating data collection and processing from users who actively engage with a website, our primary data subjects are the users of the website.

CONTROLLER pertains to any entity that determines the purposes and means of data processing. We use this term to refer to website owners, operators, maintainers, etc.

[1] The GDPR is applicable in the European Economic Area (*EEA*), that includes all EU countries and Norway, Iceland, and Lichtenstein. In addition, compatible laws apply also in the UK and from September 2023 in Switzerland. When referring to EU users in our text, we encompass users in countries with privacy regulations equivalent to the EU's.

PROCESSOR applies to any entity that processes data on behalf of the controller. For websites, this could be the cloud provider hosting the site. In the context of emails, a company sending newsletters on behalf of the controller also qualifies as a processor.

THIRD PARTY is defined as any entity apart from the data subject, controller, and processor. This definition diverges from the conventional computer science notion of a third-party domain, which simply denotes a domain involved in loading a page different from the user's initially visited first-party domain. So under CS interpretation are processors, when they are present on another domain, also a third party. We maintain a distinction by using either the GDPR's meaning of a 'third party' or the computer science's meaning of a 'third-party domain.'

2.2 PERSONAL DATA

The GDPR exclusively applies to personal data, which it defines in Article 4.

> *1. 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*

This broad definition encompasses email addresses as well as behavioral data concerning browsing patterns. Browser cookies primarily serve as technology to gather behavioral data and, thus, fall within the scope of Article 5(3) of the ePD, which addresses data in a user's terminal equipment, such as a computer, used for purposes beyond those strictly necessary to provide a service. The European Court of Justice ruled in the Planet49 case [32] that tracking users through cookies cannot be categorized as strictly necessary processing to offer a service.

2.3 LAWFULNESS OF PERSONAL DATA PROCESSING

As per Article 6 of the GDPR, processing of personal data is lawful only when at least one of the following six legal bases is satisfied:

> (*a*) *the data subject has given consent to the processing of his or her personal data for one or more specific purposes;*
>
> (*b*) *processing is necessary for the performance of a contract to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;*
>
> (*c*) *processing is necessary for compliance with a legal obligation to which the controller is subject;*
>
> (*d*) *processing is necessary in order to protect the vital interests of the data subject or of another natural person;*
>
> (*e*) *processing is necessary for the performance of a task carried out in the public interest or in the exercise of official authority vested in the controller;*
>
> (*f*) *processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.*

In the context of websites monetizing personal data, such as email addresses or browsing histories collected using cookies, all except for one of these legal bases can be ruled out. First, data processing that benefits businesses does not align with legal bases (c), (d), or (e).

Second, the European Data Protection Board's decision 3/2021 [33] deemed Meta's use of the contract legal basis (b) for targeted behavioral advertising to be unlawful. Data collection and processing for behavioral advertising are not essential for the service's functionality, which users perceive as a communication tool. However, this legal basis remains valid for instances like product purchases where data collection or processing is necessary [34]. A parallel exception for purchased products is granted under Article 13(2) of the ePD. This has implications for our email project in Chapter 3, where we exclude services subject to payment, potentially involving the contractual legal basis.

Finally, the 'legitimate interest' legal basis (f) remains the most ambiguous. The recent clarification by the European Cookie Banner Taskforce [35] established that it is not a valid legal basis for collecting personal data for activities like 'Creat[ing] a personalised content profile.' The taskforce criticized practices where cookie notices obscure certain choices in settings, requiring users to provide negative consent multiple times to decline personalization. The taskforce emphasized that such data processing falls under the ePD's consent requirements. The final verdict is still to be given by Court of Justice of the European Union upon request of Rechtbank Amsterdam (Netherlands) court in Case C-621/22 [36].

Hence, for the personal data collection and processing studied in this thesis, consent remains the only legal basis that website operators should employ.

2.3.1 *Consent*

Consent is defined in Articles 4 and 7 of the GDPR.

> *'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her;*

We outline the following requirements for consent:

- Consent must be freely given, allowing the data subject a genuine choice regarding data processing. The concept of freely given consent is intricate and is further addressed by Articles 7, 8, and Recital 43 of the GDPR. For instance, consent cannot be bundled with contract performance, as this would coerce the user into consenting.

- Consent must be for specific and legitimate collection and processing purposes. When multiple purposes exist, separate consents are required. This requirement for purpose-specific consent forms the basis of our study of cookies in Chapter 4.

- Consent must be informed, necessitating the controller to effectively communicate data collection and processing practices to the data subject. The information must be easily accessible; linking to lengthy privacy policies that are rarely read [37] likely renders consent invalid.

- Consent must be unambiguous, demonstrated through affirmative action. In the context of websites, this typically involves checking a checkbox, as opposed to pre-selected options not actively chosen by the data subject. This aligns with the ePD's 'opt-in' requirement under Article 13, in contrast to the 'opt-out' approach of US privacy laws. However, checkboxes are not obligatory when the user's action unambiguously indicates consent, such as subscribing to a newsletter.

Consent has also received further clarifications in EDPB's guidelines 05/2020 [38].

2.4 TERMINOLOGY

Throughout this thesis, we report our observations as *potential* violations for three reasons. First, as a matter of legal formality, only a legal proceeding can determine a violation. Second, while we were conservative in defining the types of potential violations, and our analyses are informed by the relevant statutes, judicial precedent, and articles by legal experts, there remains some legal uncertainty as to how courts will decide specific cases. Third, we faced factual uncertainties during our assessment. This is addressed in the appropriate sections. We remain confident that possible labeling disagreements are not of a magnitude or type that should affect our reported results.

3 PRIVACY OF REGISTRATION

To register for web services, users generally must provide their email addresses. Unfortunately, this information can be used by companies to send unsolicited marketing emails [39]. This misuse, along with the sheer number of users' online accounts, leaves users with no idea why they received a particular marketing email and from where the sender obtained their email addresses.

To counteract unsolicited email advertising, regulations on privacy and unfair competition have come into force. The EU's ePrivacy Directive established the requirement of users' prior consent for sending marketing emails. The precise notion of consent is provided by the General Data Protection Regulation (*GDPR*).

We analyze how well websites sending marketing emails comply with legal requirements. While previous studies focused on only on newsletter subscription (in politics [7, 40] or e-commerce [6]), we generalized our analysis to any forms that collect our email address and therefore might need consent for sending marketing emails. We report the following three aspects. First, we study how registration forms and emails ask for consent to marketing emails. Second, we analyze the content of the emails sent by these websites. Lastly, we detect websites sharing users' email addresses with third parties. We elaborate on this by analyzing whether websites disclose this practice.

We conducted such an analysis twice. First, we manually registered to 666 websites, annotating their forms with 21 legal properties and emails as marketing or servicing. Second, we automate the registration procedure by a crawler, which is able to successfully submit the forms of 5.9% of websites. We then use machine learning (*ML*) to predict the same 21 legal properties. Based on the legal properties and presence of marketing emails, we defined a decision tree for detecting potential violations. We report the presence of potential violations in both the manual and automated studies, and we discuss their discrepancies.

Organization. This chapter is structured as follows: In Section 3.1, we describe our manual pilot study. Specifically, in Section 3.1.1 we review the legal requirements governing email marketing. Subsequently, we explain the annotation process concerning registration and provide insights in the content of the annotated datasets of websites in Section 3.1.2 and emails in Section 3.1.3. After the pilot study, in Section 3.2 we report automating and scaling up our manual procedures. In Section 3.2.1, we describe the development of a crawler designed to automate the website sign-up process. In Section 3.2.2, we explain the ML methods we employ for predicting legal properties, which required manual annotation during the pilot study. Afterward, we undertake a legal analysis, encompassing both manually and automatically processed datasets in Section 3.3. Finally, our automated findings are subjected to a manual inspection in Section 3.4.

3.1 PILOT STUDY

3.1.1 *Legal taxonomy*

Privacy legislation has been recently introduced in many parts of the world aiming to strengthen consumer rights and privacy in the digital era. In Europe, the specific rules that relate to marketing emails consist of a complex interplay of European and national law. However, an essential pillar of legislative efforts against unsolicited marketing emails was the adoption of an *opt-in* requirement, whereby marketing emails are prohibited in the absence of prior consent [41, p. 79]. While some member states like Germany adapted such a regime early on [42, p. 168], the ePrivacy Directive [29] has established the opt-in requirement in July 2002 at European level [43, p. 46]. In particular, Article 13(1) of the ePrivacy Directive provides the requirement of an *opt-in*. This EU provision was implemented in Germany by § 7(2) No. 3 of the Act against Unfair Competition (*UWG*) [44], which is a national legislation that aims to protect companies and consumers against unfair competition practices.

There is one exception to the opt-in requirement: the presumption that existing customers have given sufficient consent to receive marketing emails advertising similar products and services they had previously procured. The specific requirements are outlined in Article 13(2) ePrivacy Directive and § 7(3) UWG. The exception implies that a product or service was provided for money [45]. Although controversially discussed, providing personal data as payment for “free” services is insufficient to generally trigger the exception ([46] in discussion of [47]). To protect customers from unsolicited commercial communications, legal scholars and German courts have tended to interpret the exception strictly [48]. As a result, the exception is not relevant for our study.

In addition to the opt-in requirement, legislators have provided further and complementary measures in many different European and national laws, often with the aim of achieving transparency. Evaluating the legal landscape therefore involves further sources of laws, such as information requirements laid down in the e-Commerce Directive [49] or the German Telemedia Act (TMG) as the corresponding national implementation [50]. Furthermore, the EU's Directive on Unfair Commercial Practices (UCPD) [51] specifically bans persistent and unwanted solicitations by email [51, No. 26 of Annex I]. The UCPD has recently been amended in the context of the EU's "New Deal for Consumers." In the following, we focus primarily on Article 13 of the ePrivacy Directive because these sector-specific provisions prevail over the UCPD [52, p. 90].

We selected the German implementation of the ePrivacy Directive as Germany is the largest economy in Europe. It is worth noting that Article 13(1) of the ePrivacy Directive ensures a complete harmonization of national rules with respect to email marketing in a business-consumer context. For this reason, it is not expected that implementations vary widely among EU member states. The European Commission concludes in a report that member states have adequately implemented Article 13(1) of the Directive [53, p. 10].

3.1.1.1 *Valid consent under the GDPR*

The interplay between the ePrivacy Directive, the UWG, and the GDPR is complex [31], but it is clear that consent is required. What "consent" means is a question of the GDPR. With respect to the term "consent," the ePrivacy Directive refers to the former Data Protection Directive [54]. The reference to the repealed Directive is now construed as a reference to the GDPR. This view is confirmed by the German Federal Court of Justice (Bundesgerichtshof, BGH) which held in a judgment of 28 May 2020 that consent must be interpreted in accordance with the GDPR's notion of consent [55]. The European Court of Justice also agreed with this view in the underlying preliminary ruling [56].

In general, Articles 4(11) and 7 GDPR are the relevant provisions of the GDPR. Thus, Article 4(11) of the GDPR defines consent as: "any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data. . . ." Beyond this definition, Article 7 GDPR provides content-wise and formal requirements. In addition, we also consider the specific guidelines on consent adopted by the European Data Protection Board (EDPB) [57].

FREE, SPECIFIC, AND UNAMBIGUOUS CONSENT First, consent must be freely given. Users should therefore have a genuine or free choice to refuse consent. The GDPR prohibits in Article 7(4) to condition the performance of a contract on an unnecessary consent declaration (so-called *bundling*). The sending of marketing emails is hardly ever necessary for the performance of the main service. Accordingly, the consent declaration for marketing emails should be unbundled from the main registration [58, paragraph 24].

Second, the declaration of consent must be specific. As early as 2008, well before the GDPR entered into force, the German Federal Court of Justice (BGH) held in its Payback judgment relating to § 7(2) No. 3 UWG that a separate declaration of consent, relating only to marketing emails, is required [59]. Although the BGH has recently ruled that a consent declaration can include several advertising communication channels (such as telephone, e-mail, and text messages), the requirement of a specific and separate declaration of consent is still established case law [60]. The mere acceptance of general terms and conditions or privacy policies is also deemed insufficient [57, paragraph 81].

Lastly, consent must be unambiguous. In the context of marketing emails, consent must be given through an affirmative act or declaration. According to Recital 32 of the GDPR, actively ticking an optional checkbox can constitute a clear affirmative act. Conversely, inferring consent from inactivity, presenting users with pre-checked boxes, or other opt-out solutions are considered ambiguous [56, 61]. It must be obvious that the user has consented. Nudging users to provide consent with visual features such as color tricks or hidden consent declarations is also not enough to fulfill this requirement.

3.1.1.2 *Legal taxonomy*

To operationalize the legal requirements of free, specific, and unambiguous consent, we have developed a legal taxonomy. We have tested the taxonomy in an exploratory pilot (see appendix A.1.1.1 for more information about the pilot study), and refined the legal properties accordingly. In Section 3.3.2, we present a decision method that determines whether a website potentially violates the legal requirements based on an evaluation of these properties.

Let $A = \{\mathrm{ma}, \mathrm{pp}, \mathrm{tc}\}$ be a set of pre-/suf-fixes for marketing, privacy policy, and a terms and conditions checkbox, respectively. Also let $a \in A$ denote a single checkbox type. We define the following legal properties.

Marketing consent (ma_consent): The website asks for consent from the user for marketing emails on the registration page.

Marketing purpose (ma_purpose): Registering with the website is only, or mainly, for receiving marketing emails.
Marketing checkbox (ma_checkbox): There is a checkbox that the user must tick to give consent for marketing emails.
Privacy policy checkbox (pp_checkbox): There is a checkbox for consent for the website's privacy policy.
Terms and conditions checkbox (tc_checkbox): There is a checkbox for consent for the website's terms and conditions.
Pre-checked checkbox (a_pre_checked): The corresponding checkbox is already ticked by default.
Forced checkbox (a_forced): It is required to tick the corresponding checkbox to successfully register. This is often indicated with asterisks on the registration forms.
#tying_b: There is only one checkbox asking for (tying) two or three consents together. Therefore, $b \in \{\text{ma\_pp}, \text{ma\_tc}, \text{pp\_tc}, \text{ma\_pp\_tc}\}$.
#forced_c: The website does not ask for consent to the privacy policy and/or terms and conditions, but assumes it through the registration process. Hence $c \in \{\text{pp}, \text{tc}, \text{pp\_tc}\}$.
#settings: Refusing consent requires more clicks, therefore the consent is assumed by default.
#age: The user's age or the date of birth are required for registration.
#colortrick: The colors on the website nudge the user to consent. For example, giving consent is highlighted with green, while refusing it is red.
#hidden: The declaration of consent can be easily missed by users.

All these legal properties are Boolean, i.e., either a website has the property or not. We call the properties with a hashtag sign *hashtags*, and the remaining *checkboxes*. Note that the last two properties are subjective. We have therefore provided the annotators with many examples, so that their annotations will be more in agreement. Annotators can also comment on annotations, which clarify the annotation of the subjective properties.

3.1.2 *Website annotation*

We manually collected a training dataset of 1000 annotated websites. For each website, we retrieved its registration form and manually annotated it based on how it asks users for consent to marketing emails and for agreement to the website's privacy policy and terms and conditions. To the best of

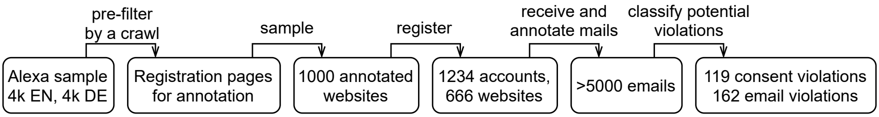


Figure 3.1: Overview of steps of our pilot study and results.

our knowledge, this is the first dataset on registration practices across the Internet.

In this section, we describe in detail the process we use for creating this dataset. We start with a short summary (see also Fig. 3.1):

1. We collected a set of websites from Alexa's ranking (Section 3.1.2.1).
2. We designed a website annotation procedure (Section 3.1.2.2).
3. We had a group of six legally-trained annotators execute this procedure on the set of websites (Section 3.1.2.3).
4. We had each website annotated a second time by a second annotator. This allowed us to measure the annotators' consistency. Any conflicts were subsequently resolved by a third annotator. (Section 3.1.2.4).

3.1.2.1 *Website collection*

Alexa (alexa.com) ranks websites according to page views and site users, and maintains a list of the most popular websites based on this ranking for the last three months. We used Alexa's top 1 million websites worldwide from May 25th, 2020.

Our goal is to inspect websites with varying popularity, so we split this set into four groups: the top 1000, the next 9000, the next 90 000, and the rest. From each group, we randomly selected 1000 unique websites. This sampling ensures that we analyze many of the most popular websites, in contrast to an entirely random selection. We call this the EN set of websites, as it is the starting point for detecting websites in English.

Considering that the underlying legal analysis uses German law and court cases as an example of the implementation of the EU's ePrivacy Directive, we focused on websites that allowed registration for people located in Germany. Therefore, we also created a separate set of 3694 websites, the DE set, by taking from Alexa's top 1 million, those websites with the domain ".de." Since the notion of consent in German law is interpreted according to the GDPR, our dataset is still likely representative of how websites across Europe ask users for consent.

TABLE 3.1: Website selection process for the manual study.

Processing step	Size EN	Size DE
Sampled	4000	3694
Pre-filtering crawl	662	436
Randomly sampled for annotators	607	393
Registered successfully	343	325

Based on the study by Chatzimpyrros et al. [62], who observed that only one third of websites have login or registration forms, we did not expect to find more websites with available registration in our selected languages. To reduce the number of annotations where registration was not possible, we pre-filtered both the EN and DE sets of websites using a crawler. This crawler filtered websites that are not available in English or German, malfunctioning websites, and websites without a registration. Table 3.1 shows the website selection process.

We analyzed 100 filtered websites to inspect whether the filtering causes a bias in our study. From 50 randomly selected DE and 50 randomly selected EN websites that were filtered out, it was possible to register for thirteen of them and subscribe to one of them (seven EN and seven DE websites). These websites were mostly rejected due to advanced bot detection (seven websites),[1] which can cause under-representation of more complex websites. However, these websites were uniformly distributed in the Alexa rank. The authors manually registered to all fourteen filtered websites and found no statistical deviation from any presented observations in this study. The Bachelor's thesis by Kast [63], which was working with the crawler used for the pre-filtering, provides similar analysis of the filtered websites. Its results are aligned with ours.

3.1.2.2 *Annotation procedure*

Every website was manually annotated with the legal properties described in Section 3.1.1.2. To determine these, a human annotator would register for the website, using fictitious personal information like name, address, or phone number. Only the email address provided is real, as we use its inbox to detect unsolicited marketing emails. In addition to the properties, annotators marked the registration as either successful or unsuccessful, depending on whether

[1] Confirmed by the Wayback Machine, which was also unable to visit these websites.

they successfully registered to the website. When unsuccessful, they provided the reason for not completing the registration, for example, by stating that there was no registration form on the website, or that the registration required a payment.

We developed a support tool to facilitate the manual process of registration and annotation. Our tool features a graphical interface for recording the legal properties, according to the legal taxonomy defined in Section 3.1.1.2. Our tool uses Firefox, which we extended by Selenium to also help annotators by automatically filling in registration form fields with the generated credentials. We describe this tool in appendix A.1.1.2.

For each website, our support tool retrieved the HTML source of the entire page and the registration form's HTML subtree. If the webpage contains multiple forms, such as a login and a registration form next to each other, we detect the form with which the annotator interacted and collect only its HTML subtree. All Internet traffic was routed via a German VPN endpoint, so our requests appeared to originate from Germany.

3.1.2.3 *Annotators*

Six scientific research assistants, all with a law degree, annotated the 1000 websites. The annotators were compensated fairly, according to the hourly wage for teaching assistants. To avoid biasing them, we did not inform them about our research objectives.

The annotators were randomly assigned the websites from the EN and DE datasets. The amount of work each annotator performed depended on their availability, and ranged from 95 to 453 annotated websites per annotator.

The website annotation process was manual, but it was precisely defined by instructions we provided. These included legal and technical guidelines and examples of 22 annotated websites with justifications for the annotations. We had previously tested the instructions in an independent pilot study.

3.1.2.4 *Resolving disagreements*

Following empirical social science standards, every website was validated by a second independent annotator [64, p. 114]. The second annotator was randomly chosen for every website and was different from the first annotator, but from the same group of six annotators. We observed only a single website that changed the registration form by the time the second annotator annotated the website, so website modifications were not a significant source of inter-annotator disagreement.

In case of inconsistencies between the annotations, we provided a third annotator with screenshots of the registration forms seen by the first two annotators and their annotations. He would then choose one of the two annotations and, if necessary, he could modify the selected annotation. The third annotator was not part of the original set of annotators and also had a law degree.

We measured the agreement between annotators with Cohen's κ [65]. Like a correlation, it takes values between -1 to 1, where $\kappa = 0$ indicates the absence of agreement, $\kappa = 1$ indicates perfect agreement, and $\kappa = -1$ indicates perfect disagreement. For legal properties that were satisfied by at least 10% of the websites, the average κ in our sample was 0.74. All the individual κ's are given in appendix A.1.1.3.

Our annotation procedure was more rigorous than those procedures used in most other related studies. For example, in Zimmeck et al. [66], 350 policies were labeled by two law students. Only 35 of them were doubly annotated and their Krippendorff's α was 0.78 (text labeling requires this metric for inter-annotator agreement, but it has the same range and a similar interpretation as Cohen's κ). In Bannihatti et al. [67], a law student labeled 2692 opt-out statements from privacy policies. Only a subsample (50) was labeled independently by two additional annotators. The inter-annotator agreement was measured with Fleiss' κ, and its value was 0.7 (in this context, Fleiss' and Cohen's κ are identical). To the best of our knowledge, the only other study with an annotation procedure as rigorous as ours is Wilson et al. [68], who used two law students to annotate 115 privacy policies with an average Krippendorff's α of 0.71, and had a third law student resolve any inconsistencies.

3.1.2.5 *Resolved annotations*

For 666 of the 1000 websites, the annotators agreed on successful registration. The most common reasons for unsuccessful registration was that there was no registration form (9%), the registration required a membership (7%), or the registration required payment (5%). We report the reasons for other failed registration in Fig. A.2 in the Appendix. Fig. 3.2 depicts the resolved annotations for websites with successful registration. Each bar represents the percentage of websites satisfying that property. Note that more than half of the websites do not mention marketing emails in the registration form. Only 6.6% (44) of websites provide for marketing email subscription (`mark_purpose`), which indicates the number of websites we can expect to send us marketing emails with properly granted consent.

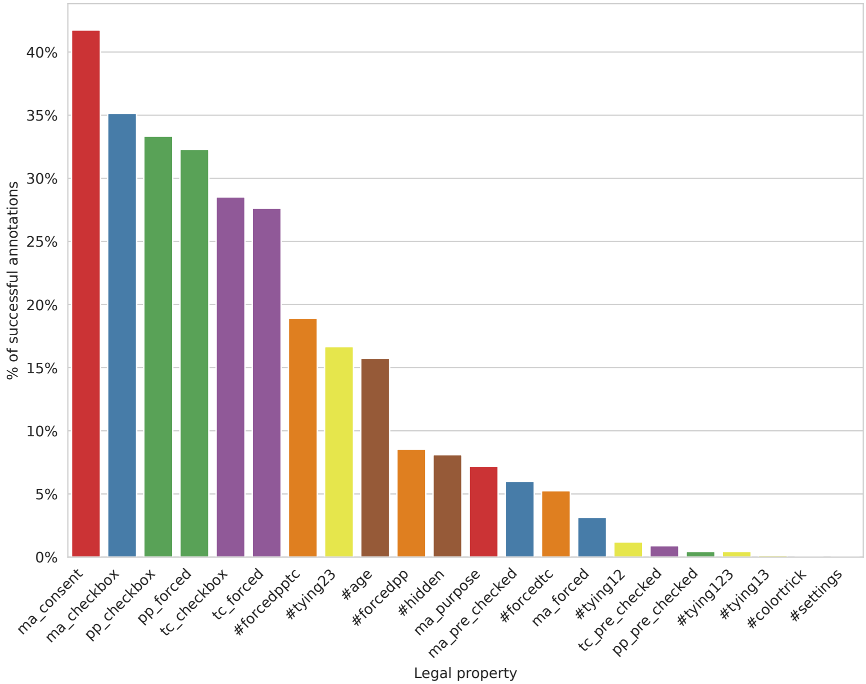


FIGURE 3.2: The number of observed legal properties (defined in Section 3.1.1.2) in the successful registrations.

3.1.2.6 *Ethical consideration*

Informed by ethical considerations, we adhered to the following protocols, as we created the website dataset. We did not register for websites where we would order products or services while not honoring the contract. Moreover, the annotators were instructed to skip illegal services or content. As we did not use real persons during registration, we do not harm the privacy interests of the annotators. We ensured that these credentials do not match any real person. Finally, we provide our datasets only for research and replication purposes.

3.1.3 *Email annotation*

We registered for websites using a real email address with a fictional identity. To analyze whether the website shares the email address with a third party, we generated a unique email address for each website. We hosted these email addresses privately at `infsec-server.inf.ethz.ch`. All annotated emails were fully loaded and rendered, including any tracking mechanisms confirming the email account activity to the sender.

For most of the websites, we registered accounts in both registration rounds, so we receive emails to two unique addresses by the same sender. However, we also analyze the websites where we registered only once. In total, we generated 1234 unique email addresses. During the eight months of the study, 987 of these addresses received at least one email. This corresponds to 568 different services, which serves as the baseline for this section. While each address received around five emails on average (the median was one), one service sent us over 200, and the top 10 senders jointly sent us over 1000 emails. In total, we collected and annotated over 5000 emails.

In this section, we explain this procedure in more depth. This includes the following steps.

1. We define marketing and servicing emails and show their distribution in our dataset.

2. We present the *double-opt-in* procedure and report that fewer than 60% of websites follow this best practice.

3. We check the content of servicing emails for passwords in plaintext, finding 2.3% of websites send the user-provided password in plaintext via email.

4. We check the content of marketing emails for unsubscribe options and legal notices, observing that 16% of websites do not meet at least one requirement.

5. We check whether companies share the registered email address to third parties, finding that 4.1% of our addresses receive emails from multiple senders.

Overall, from the 568 websites that sent emails, over 20% sent at least one email that potentially violates the legal requirements described in this section. This number does not include the 36% of websites that send emails without following the best practice procedure of double opt-in.

3.1.3.1 *Marketing and servicing emails*

In order to detect emails falling within the EU regulatory framework, we distinguish between marketing and servicing emails [41, p. 7].

Marketing emails typically advertise specific products or services. Examples include product-related newsletters or vouchers. It is settled case law of the German Federal Court of Justice that the term “marketing” is interpreted in a broad sense and in accordance with Article 2(a) of the EU’s Directive on misleading and comparative advertising [69]. This case law was last affirmed by the BGH in 2018 [70]. Therefore, marketing also covers indirect sales promotion such as non-product-related image advertising, customer surveys, and birthday and holiday letters.

Servicing emails are ad-free and not intended to promote products or services. Often these are transactional emails triggered by the user. Examples are registration confirmations, invoices, and updates on changed terms and conditions. As our only interaction with the website is the registration and its confirmation, the number of servicing emails is limited.

We annotated the dataset of over 5000 emails with these email types, and we present their distribution in Section 3.1.3.1. The annotation was done by one of the authors and one research assistant using the email’s subject and body and information from the annotator’s website registration.

3.1.3.2 *Double opt-in*

Double-opt-in emails require an additional user action after registration to activate the account. This action serves as the user’s proof of ownership of the provided email address and can be implemented in several ways. The email typically contains unique information, such as an activation link, a one-time password, or a verification code. Alternatively, it may require the user to initiate the account activation by sending an email, a less common method observed on fewer than 0.5% of the websites where we registered. Marketing emails can only be sent after obtaining consent through these preceding actions. In contrast to the *single opt-in* process, the double-opt-in procedure effectively prevents users from registering, either unintentionally or maliciously, with an email address that is not under their control. The company offering registration must ensure that the email addresses belong to the registered users and must keep clear records of consent.

Following this explanation, we categorized servicing emails into three distinct groups: double-opt-in emails, confirmation emails (encompassing both confirmations related to the double-opt-in procedure and those stemming

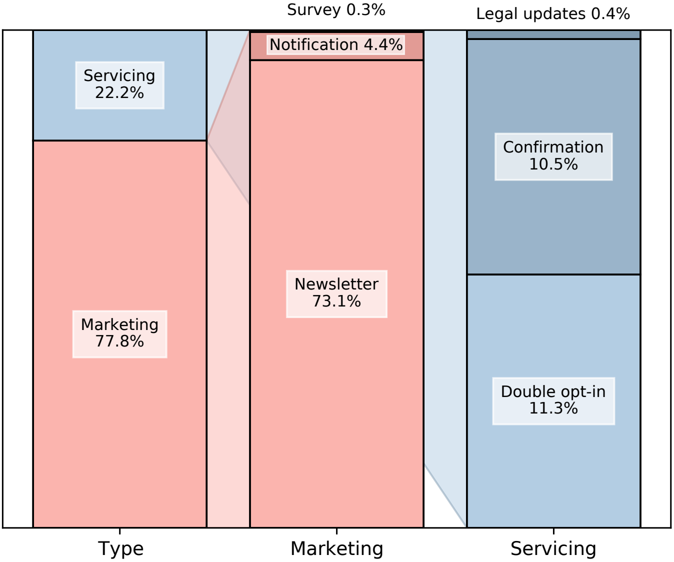


FIGURE 3.3: Email classification of the 5030 annotated emails, where we zoom into the marketing and servicing subclasses.

from single-opt-in processes), and other servicing emails, such as notifications pertaining to changes in the privacy policy. The training dataset consists of 570 double-opt-in emails, 531 single-opt-in emails, and 18 remaining servicing emails.

3.1.3.3 *Design of marketing emails*

There are specific provisions that govern the content of marketing emails. We focus on how websites perform two common practices. The first is letting users unsubscribe from marketing emails and the second is informing users about the origin of the email by legal notice. While we present the German legal background, it's important to note that both provisions are derived from EU Community legislation, specifically Article 13(4) of the ePrivacy Directive and Articles 5 and 6 of the e-Commerce Directive [71].

Marketing emails must contain a method for users to unsubscribe from subsequent emails. According to § 7(2) No. 4 (c) of the UWG, the method must be clear, unambiguous, and free of costs other than the transmission costs under the basic rates. Additionally, GDPR Article 7(3) emphasizes that

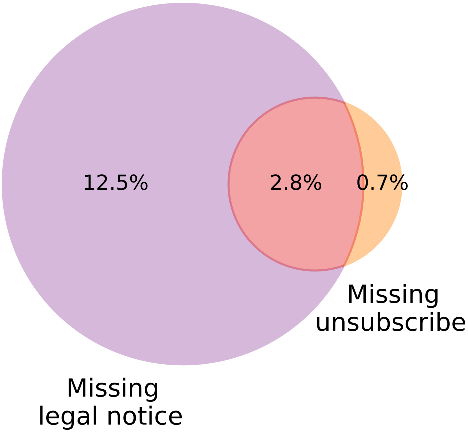


Figure 3.4: Venn diagram of emails missing legal notice and unsubscribe method. Percentages are relative to all marketing emails. The remaining 84.0% of emails contained both legal notice and unsubscribe method.

opting out should be as straightforward as opting in. Furthermore, according to § 7(2) No. 4 of the UWG and with reference to § 6 of the German Telemedia Act (*TMG*), marketers must not disguise or conceal their identity. Companies sending marketing emails must include some company details, known as the *legal notice*, in their emails based on § 5(1) of the TMG [72]. We inspect a selection of the required company information, including the company's name, the company's address, and the email address. Note that these requirements are not exhaustive, but they are among the most common and generally applicable in various jurisdictions.

Our analysis involves the inspection and annotation of both the presence of an unsubscribe method and the inclusion of a legal notice. We combine both pattern matching and manual inspection to detect missing email content. The most common unsubscribe method is an unsubscribe link, typically placed either in the email body or in the X-Headers. An alternative method requires users to send an email to the service provider to unsubscribe. Legal notices are typically located in the email footer and include information such as the company name and service domain, which allows for the identification of the legal notice.

Due to the specific nature of these requirements in German law and the technical challenges associated with recognizing email components, we report

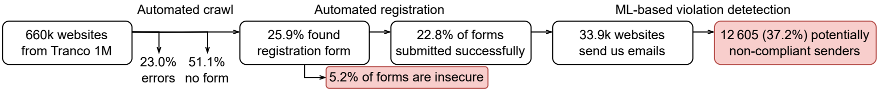

Automated crawl
Automated registration
ML-based violation detetection
660k websites from Tranco 1M
23.0% errors
51.1% no form
25.9% found registration form
22.8% of forms submitted successfully
5.2% of forms are insecure
33.9k websites send us emails
12 605 (37.2%) potentially non-compliant senders

Figure 3.5: Overview of steps of our automated study and results.

violations related to missing email elements only within the dataset from the pilot study. Fig. 3.4 illustrates the portion of the email dataset that lacks either an unsubscribe method, a legal notice, or both. Of the emails reviewed, 84.0% adhered to both the unsubscribe and legal notice requirements, while 2.8% were deficient in both aspects. It is important to note that these reported numbers are conservative because we based our calculations on the number of services that sent us emails, while we assessed the design requirements exclusively within marketing emails.

Finally, the email dataset can reveal additional information about marketing emails, such as insight into the marketing trends, which are out of the scope of this work. We present examples in appendix A.1.2.4.

3.2 LARGE-SCALE STUDY

In this section, we describe the steps required to automate the process of the pilot study, namely automating the registration process (Section 3.2.1) and then the classification of legal properties (Section 3.2.2). We illustrate these steps and the associated statistics in Fig. 3.5.

3.2.1 *Crawling infrastructure*

We developed an infrastructure for crawling websites and automating user registration. For each website where the crawler registers, we provide a unique email address for a (simulated) user. Our infrastructure then analyzes the received emails to evaluate how the website uses the user's email address.

Websites vary significantly in both their appearance and implementation, primarily due to the flexibility of JavaScript and CSS. Since all registration options must adhere to the same laws regardless of the technologies used, we focus on registration using an email address. We therefore do not attempt to register using single sign-on, which was covered by other compliance studies [73].

Below we discuss the crawling steps. First, the crawler navigates through the website to find pages containing a registration form, which it then fills out and submits. Afterwards, it checks the registration state and finishes the double opt-in when it is requested by email.

3.2.1.1 *Crawler implementation*

To simulate users' browsing patterns, our crawler utilizes a real browser orchestrated by Selenium. Since existing frameworks such as OpenWPM [74] or webXray [75] are not designed for the complex crawling that our task demands, we do not use them. To represent the majority of web users, we crawl websites using Chrome, but Firefox is supported as well.

To maximize the chances of successfully loading websites, we employ several techniques to evade bot detection, which we describe in Appendix A.1.3. We have tested that our crawler is not flagged by any major Content Delivery Network (*CDN*), including Cloudflare, Fastly, Amazon CloudFront, and Akamai.

Our crawler successfully loads 90.6% of websites, as opposed to 70% without bot evasion techniques. In comparison, Le Pochat et al. [76], successfully crawled 85% from URLs of a similar list (the intersection of the Tranco and Chrome UX report lists). Their crawler did not actively evade bot detection. We suspect that many of the websites that they report as successfully loaded actually flagged their crawler as a bot and presented a simple warning page.

3.2.1.2 *Crawler navigation*

After loading each website with a fresh cache, the crawler determines the page's language using the `polyglot` Python package. If the language detection fails, we rely on the <html> tag. If English is not the detected language, the crawler tries to switch to the English version, if one exists. We keep browsing the website regardless of the switch to English since we support the majority of European languages (see Appendix A.1.4).

KEYWORD MATCHING The detection of a link or button to change the language is based on matching keywords in the visible text, the 'alt' attribute of <img> tag, or the URL. We curated phrases for determining the purpose of page elements, such as a privacy policy link or marketing consent checkbox. Native speakers translated these phrases to all the supported languages. The curation was guided empirically by example websites. The matching procedure works as follows. First, we remove stop words from both the website and the keyword phrase. Then we lemmatize both texts, using the SpaCy [77] or lemmagen3 [78] lemmatizers, depending on the language support. Next, we map characters with accents or Cyrillic to lowercase ASCII counterparts. Finally, the processed keywords and phrases are matched. This keyword matching approach is also used for other navigation aspects, which are described below.

NAVIGATING WEBPAGES Our crawler uses a priority queue to determine the order of visiting pages of the site. The priority represents the likelihood that a given link leads to a registration or a newsletter form. We order the link categories starting with the highest priority as follows: the registration page, login page, privacy policy and terms and conditions, and others. Links within a category are ordered by their matching score.

From each loaded page, the crawler collects at most three links per category. The links labeled as "other" are selected randomly, and they prevent the crawler from getting stuck on a page with no other links, such as cookie walls or an empty landing page with "Entry" links used by adult websites. The privacy policy and terms and conditions links are collected after registration given their potential relevance in the further legal evaluation.

The crawler is restricted to visiting at most twenty pages and the registration page is typically reachable within the first five pages. We allow the crawler to navigate beyond the original TLD+1 domain,[2] but only for a single step, i.e., links found on external domains are not considered for subsequent crawling. This allows registration on an affiliated website directly accessible from the original site. However, it restricts the crawler from navigating away from the original site and identifying unrelated registration forms. Moreover, the keyword-matching algorithm penalizes external domains.

PAGE CONTENT CLASSIFICATION When we load a page, we classify it according to the presence and thereby type of a <form> tag. We apply the decision tree depicted by Fig. 3.6 to classify the form as registration, login, subscription, contact, search, or other. We evaluated this procedure on a manually annotated dataset collected from 1000 randomly selected English websites from Tranco 1M,[3] containing 426 forms. There were 12 contact, 32 login, 139 subscription, 163 registration, and 80 other forms. Procedure from Fig. 3.6 detected 74% of the registration forms and 94% of the subscription forms, yielding an overall accuracy of 82%.

[2] TLD+1 refers to the registered domain name preceding the top-level domain. For example, in both bbc.co.uk and bbc.com, the string 'bbc' represents the TLD+1.

[3] From an older crawl using `https://tranco-list.eu/list/89WV/1000000`.

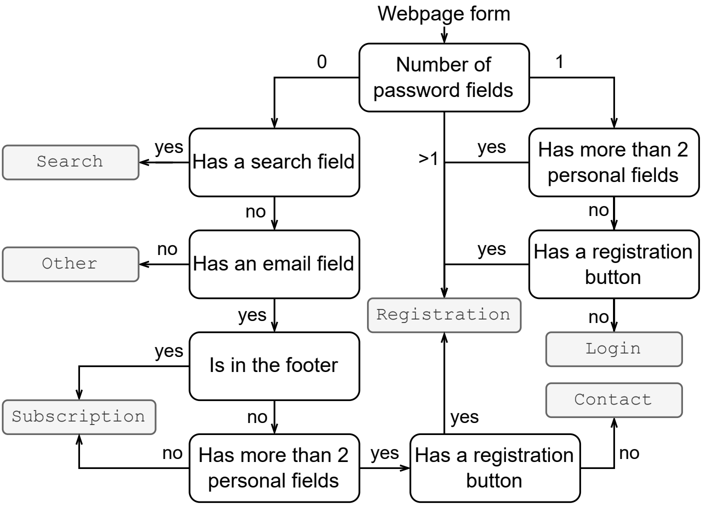


FIGURE 3.6: Crawler's form classification procedure.

3.2.1.3 *Crawler form interaction*

Once we detect a registration form, or a subscription form when no registration form is found, we interact with it. We first extract the entire subtree of the <form> tag, which we process using the Beautiful Soup library. We use a similar keyword-matching method as in Section 3.2.1.2 to detect the type of input fields. We search for matches in the corresponding <label> tag and visible text, and in attributes such as autocomplete, type, label, placeholder, and value.

Once we determine the input type, we check which input fields must be filled as indicated by the presence of the "required" attribute, an '*,' or a bold label. Then we fill all the required inputs by simulating typing. We ensured that our fictitious credentials including an EU address seem plausible. This, together with VPN in the EU, should indicate for the website that EU privacy laws are applicable, which we further discuss in Chapter 5. Most importantly, we generate a unique email address for every website.

CHECKBOXES AND FORM SUBMISSION We interact with every required checkbox and <select> tag. The latter is usually used for the birth date to ensure that the person's age exceeds some threshold. Once the form is filled, we submit it using any detected submission button or by simulating pressing the Enter key. After submission, we look for a redirect or a change in the website content to detect the registration state. We compute the difference in the website's visible content and the form code to distinguish the following outcomes. The text differs and contains keywords indicating a 'successful' or 'failed' registration. The form remains unchanged, usually indicating a 'failed' registration. The form is changed after a redirect, indicating a multi-step registration. None of the above applies and we denote this an 'unknown' state.

If the registration failed but the same form is still present, we try filling in the credentials again, but this time we confirm all checkboxes. This increases the probability that a required checkbox like "I agree with the terms and conditions" is checked. However, it also increases the probability of consenting to sending marketing emails, which could be detrimental to the objective of our consent study.[4] Then the form is submitted again, possibly many times when the form changes and our heuristic detects a multi-step registration.

CAPTCHA SOLVING During any of the crawling steps, we might encounter a CAPTCHA. This usually happens during registration or when loading an index page is intercepted by CloudFlare or a similar DDoS-mitigation service. The crawler observes the type of CAPTCHA by the JavaScript that loads it. For reCAPTCHA or hCAPTCHA, we load a template substitute JavaScript that prevents crashes due to website changes of the CAPTCHA invocation. Image CAPTCHAs are detected by keywords directly in the forms. We use an external service that solves CAPTCHA using humans.

A third of crawled websites use CAPTCHAs, namely 75% of them are ReCaptcha v2, 20% are ReCaptcha v3, 2% are hCaptcha, and 3% are image CAPTCHA.

SELF-HOSTED MAILSERVER We self-host generated email addresses at `sybilmail.de`, configured to only receive emails using the Mail Delivery Agent implemented with the Python Maildir library.

3.2.1.4 *Registration confirmation*

Once the crawler determines that the registration state is either 'successful' or 'unknown,' it waits for a confirmation email. As shown in [11], only 80% of

[4] Checking all checkboxes hinders detecting the 'marketing email despite user did not consent' violations. By skipping such a requirement, our crawler improves the registration rate by more than 10%, which is relevant for application of our crawler to other than consent compliance studies.

websites send emails to registered users and, of those, 59% send double-opt-in emails requiring activation. If we receive an activation email, we extract the activation link or code. The crawler visits the activation link or inserts the code into the open registration.

Since letting the crawler wait for an activation email is computationally expensive, our crawler does so for up to 30 seconds. If an activation email is received after this period, we activate the registration using a standalone script that processes the incoming emails from all the crawlers running in parallel. However, this script lacks the registration page session, such as cookies, which reduces its success rate compared to the stateful crawler within the 30-second period. We analyzed the distribution of confirmation emails over time in our crawl and observed that less than half of the activation emails arrived within this 30-second period. To achieve a higher success rate for account activation, we recommend waiting for five minutes in future work, since 97.7% of websites that send activation emails do so within this period. Further increasing the waiting period to, say, fifteen minutes would only marginally improve this rate to 99.0%. The longer waiting time, however, comes at the expense of crawling time. Specifically, waiting for five minutes doubles the crawling time, while waiting for fifteen minutes almost quadruples it. Emails with activation codes are typically prompt, so late email confirmation uses only activation links, not the codes.

Unfortunately, due to technical issues the independent confirmation script was malfunctioning for about half of the crawl. The combination of a shorter period of waiting by the crawler and the faulty script results in lower confirmation rates. This causes the presented results in Section 3.3 to be more conservative. Namely, websites that violated the consent in the form but then complied with the double-opt-in requirement and never sent us a marketing email are falsely considered compliant.

3.2.1.5 *Deployment*

We evaluated our crawler by visiting the Tranco 1M list[5] [79], generated on 15 June 2022. We selected the Tranco list to enable an accurate comparison with prior work that utilizes a similar crawling list. However, Ruth et al. [80] have observed that Tranco represents less accurately users' browsing patterns than the Chrome UX Report (*CrUX*) list. Hence we also evaluate the subset of Tranco that is present in the CrUX list. Unfortunately, due to a processing error, we crawled one million websites that were uniformly randomly sampled

[5] Available at `https://tranco-list.eu/list/82Q3V`

with replacement, rather than crawling all the websites. For this reason, our results are only based on 660 202 unique domains, corresponding to the first crawl.

The crawl was conducted from June to September 2022, averaging 10k websites per day on a server equipped with four Intel Xeon E7-8870 CPUs. We ran 60 Chrome browsers in parallel each within a separate docker container, using a freshly launched browser for every website. We used 16 IP addresses provided by the German Research Network ensuring that the traffic originates in the EU.

The crawler collected evidence in the form of HTML code from the index and registration pages, as well as extracted text from the privacy policy and terms and conditions. Additionally, we obtained screenshots of each step taken during registration and recorded all the observed cookies. Finally, the crawler collected information regarding the registration status, which we describe below.

3.2.1.6 *Crawling results*

Fig. 3.7 shows a Sankey diagram of the crawling process. From the 660 202 websites, 504 509 websites were successfully loaded in a supported language. Among the loaded websites, our crawler detected a registration or subscription form on 33.6% (169 765) of them. Furthermore, our crawler estimated the success rate of form submissions defined in Section 3.2.1.3. The estimation indicates that 30.2% of form interactions were successful (51 290), 38.4% failed (65 220), and 31.4% resulted in an undefined state (53 255).

The form submissions state detection is prone to false positives. Hence we manually investigated the correctness of the crawler determined registration state by inspecting 200 websites and testing the used credentials. The analysis revealed three newsletter subscriptions deemed successful by the crawler and nine registrations, seven of which were correctly identified as successful by the crawler. Two registrations were successful, despite the crawler assigning them an ‘unknown’ and ‘failed’ state. We suspect that newsletter forms were underrepresented in this sample and as nearly half of the received emails resulted from newsletter subscriptions. Further observations from the manual analysis are presented in Appendix A.1.6.

We also analyze the results based on whether the websites are included in the CrUX list. Note that Tranco 1M and CrUX have only 51.9% overlap. The crawl was significantly more successful for the CrUX websites. Specifically, 90.6% of the websites present in both lists were successfully loaded, in contrast with 65.3% for non-CrUX websites. Among the websites in the CrUX list,

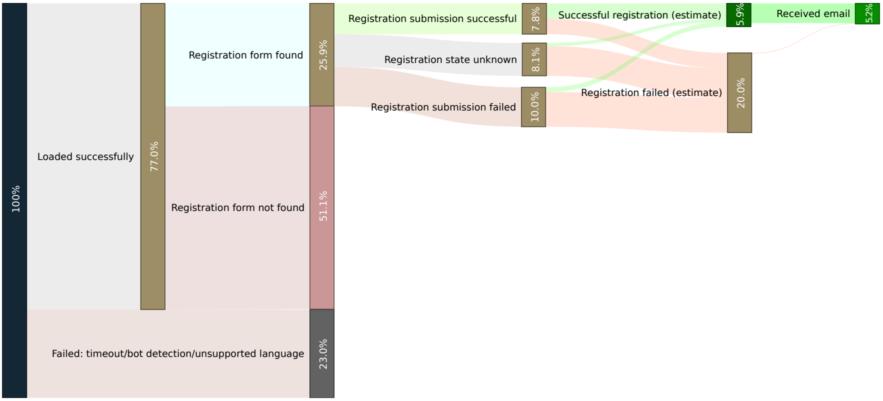


FIGURE 3.7: Sankey plot of the crawler's intermediate results.

registration was detected as successful in 11.7% of cases (3.9% for non-CrUX websites). Our list choice supports a comparison with [10], relying on the DNS-based Alexa list with domains as `WindowsUpdate.com` without HTTP(S) endpoint. In the future, we recommend crawling the CrUX list to prevent unnecessary computations.

3.2.1.7 *Ethical considerations*

We have identified the following three risks of our study. 1) *Legal risks arising from crawling*: we considered various legal regimes and concluded that our research does not violate laws such as fraud, trespass, or breach of contract as our intentions are the pursuit of good-faith privacy research. We also used Google Safe Browsing to skip crawling potentially hazardous websites. 2) *Risks to website owners*: our single crawl negligibly impacts each individual website's capacity. Moreover, the registration rarely results in a manual action by website owners, as the vast majority of emails are automated. In Section 3.3, we present only aggregated results, preventing harm by wrongful accusation of individual websites for privacy violations. For that reason, we refrain from publicly disclosing our dataset of identified violations, except in cases where parties explicitly provide consent to adhere to the same ethical standards we uphold. 3) *Risks to CAPTCHA solvers*: we employed a third-party CAPTCHA solving service, where we cannot influence the remuneration provided to their workers. Given the substantial prevalence of CAPTCHAs, accounting for one-third of our successful registration, and their prevalence on often higher-profit services, omitting CAPTCHA solving would introduce a significant bias. Finally, our university's ethics board determined that our project does not require ethics approval since it does not involve human subjects.

3.2.2 *Classifying legal properties*

In this section, we automate the prediction of the legal properties defined in Section 3.1.1. Using the annotated datasets from Sections 3.1.2 and 3.1.3, we train two types of ML models: for emails and forms. For each type of model, we describe the feature engineering step, how models are trained, and the results.

3.2.2.1 *Features of emails*

The training dataset consists of 5725 mostly German and English emails. To reduce the complexity of dealing with multiple languages and to utilize all the training samples, we translate the subjects and bodies into English using LibreTranslate. From each translated email, we further process the headers, subject, and body.

HEADERS Email headers constitute a set of key-value string pairs, such as 'Date,' 'Reply-To,' or 'List-Unsubscribe.' While several headers are standardized, there are many, often prefixed with 'X-,' that are custom to specific email servers. We define the *supported keys* as the set of all header keys in the training dataset. This resulted in 76 headers without the 'X-' prefix and 488 headers with it. For each email, we denote whether there is an entry for a given key, whether it contains an email, URL, other text, or whether it is empty.

Our feature analysis confirms headers usefulness. In particular, 'List-Unsubscribe' or 'X-CSA-Complaints' are relevant for detecting the compliance of the marketing emails with the privacy regulations.

SUBJECT The translated subjects are processed with TF-IDF encoding[6] that we fit to the training dataset. In addition to this encoding, we use a universal sentence encoder [82]. This pretrained NLP model transforms sentences to an embedding in $\mathbb{R}^{512}$.

[6] Term Frequency-Inverse Document Frequency (*TF-IDF*) is a variant of the Bag-of-Words text representation model that accounts for the total number of words. It outperforms Bag of Words in common classification tasks [81].

BODY We extract both the TF-IDF encoding of the translated body and several manually-defined numeric features. These features include the number of characters or sentences of the email text, number of URLs, images, and links.

3.2.2.2 *Training ML models for emails*

Given that our features correspond to tabular data, we use the XGBoost model [83]. XGBoost is well-suited as it outperforms other training algorithms for datasets with few annotated samples but high dimensionality of the feature space.

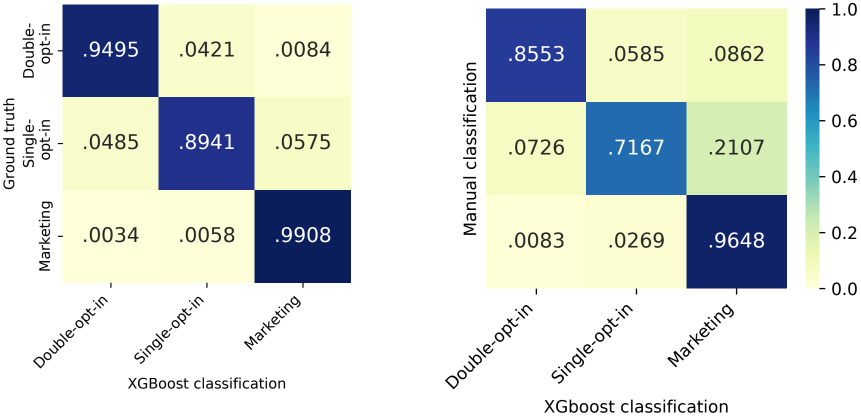
We train the model using an established ML pipeline. We perform a stratified split of the dataset dedicating 75% for training, saving 25% of the unseen data for validation. We adjust for class-imbalance by sample-weighting. The models optimize the weighted 'multi:softmax' metric for multi-class classification and 'binary:logistic' for binary classification. All reported results are based on four-fold cross-validation. Given data scarcity, we skip hyperparameter tuning, which would require a further data split, and we use the default XGBoost hyperparameters.

We trained models that predict two distinct legal properties of emails. Our first model predicts whether an email is a marketing email (i.e., newsletters, notifications promoting service monetization, and surveys), a servicing double-opt-in email, or another kind of servicing email (confirmation emails or service updates). Our second model detects whether an email contains a method to unsubscribe, which we evaluate only on marketing emails.

In Fig. 3.8a, we present the confusion matrices of the mail-type model. The mail-type model achieves 97.7% balanced accuracy, while in the simplified task of deciding only whether email is marketing or servicing (aggregating double opt-ins with confirmations and legal updates), the balanced accuracy increases to 99.2%. The same balanced accuracy of 99.2% is achieved by the model predicting presence of the unsubscribe options.

3.2.2.3 *Features of forms*

To transform forms of unlimited length to tabular features, we aggregate the form inputs by the crawler's keyword-based element classification. We group semantically similar inputs, such as the first and last name, full name, and username, see Appendix A.1.5 for details. We also reduce the complexity by excluding inputs irrelevant to legal classification, such as CAPTCHA. From all inputs, we extract whether they are required or optional, and from checkboxes



(A) Results on test set (accumulated over all CV folds) after training using the dataset of 5k emails from the pilot study.

(B) Prediction on unseen three years newer additionally labeled 11k emails.

FIGURE 3.8: Confusion matrices of mail type classification.

also their default values. We concatenate texts, such as corresponding labels, and translate them to English. Finally, we include the form type (registration or subscription) as a categorical feature.

We then process the form texts similarly as emails. Note that checkbox labels often consist of complex and nuanced statements, such as “I don’t want to receive special offers about [company name] products.” To better capture the meaning of these statements, we extract both sentence embeddings and TF-IDF representations with a limit of 500 words. However, for other form inputs, which tend to have shorter labels like “Your email,” we skip sentence embeddings and only use TF-IDF with a limit of 50 words.

The feature extraction produces 5839 tabular features: 69 numerical features about form’s input fields, 3154 TF-IDF columns, and five sequences of $\mathbb{R}^{512}$ sentence embeddings.

3.2.2.4 *Training ML models for forms*

Similarly as with the email classification, we trained an XGBoost model for each of the 21 binary legal properties annotated by [11]. Note that the training dataset consists of only 666 annotated forms. To address this data scarcity, we also conducted experiments using the Tabnet model [84], a neural network model optimized for tabular data. One notable advantage of Tabnet over

XGBoost is its ability to perform unsupervised pretraining on unlabeled data, allowing it to capture the distribution of classified data. For the pretraining phase, we provided the extracted features of 30k websites where the crawler detected registration or subscription forms. The pretraining process took 32 minutes on an Nvidia 3080 Ti GPU and resulted in a model with a loss of 1.319. Note that Tabnet is an order of magnitude slower than XGBoost in training but just twice as slow in prediction.

Table 3.2 presents the results of XGBoost with predictions based solely on the crawler's keyword-based classification of form content. However, the crawler's prediction is unavailable for some legal properties, so for space reasons we skip such rows together with Tabnet as its performance is aligned with that of XGBoost. The table provides a summary of the macro-averaged F1 score, precision, and recall, while the last column indicates the percentage of positive samples in the training dataset. Note that the overall performance is highly dependent on the number of positive samples, making scarce properties insufficient for making legal judgments. To mitigate the risk of falsely predicting a privacy violation, we combine the ML predictions with the crawler's keyword-based deterministic prediction. When the presence of a legal property implies a violation, we combine predictions using AND and conversely when it implies compliance, we use OR. We further reduce false positives by conditioning predictions when possible, such as 'marketing checkbox forced' requires 'marketing checkbox present' in the first place.

Table 3.2: Performance of legal properties models. 'Deterministic' model stands for the crawler's prediction.

Property	Model	F1	Precision	Recall	Support
Marketing consent	Deterministic	77.58%	80.43%	76.87%	41.92%
	XGBoost	82.33%	82.88%	82.08%	
	TabNet	**85.04%**	**85.69%**	**84.65%**	
Marketing purpose	Deterministic	**68.06%**	**64.95%**	**74.65%**	7.04%
	XGBoost	63.15%	61.71%	66.21%	
	TabNet	57.40%	56.41%	63.41%	
Marketing checkbox present	Deterministic	79.01%	83.23%	77.33%	35.18%
	XGBoost	81.67%	82.95%	81.04%	
	TabNet	**84.48%**	**85.79%**	**83.58%**	
Marketing checkbox pre-checked	Deterministic	**71.74%**	**73.26%**	**70.44%**	5.84%
	XGBoost	57.66%	57.58%	58.43%	
	TabNet	54.87%	55.73%	68.85%	
Marketing checkbox forced	Deterministic	55.67%	59.67%	54.22%	3.14%
	XGBoost	**58.94%**	**59.84%**	58.38%	
	TabNet	56.91%	56.94%	**90.43%**	
Policy checkbox present	Deterministic	81.75%	**84.23%**	80.29%	32.93%
	XGBoost	**83.58%**	83.60%	**83.62%**	
	TabNet	82.64%	83.59%	81.90%	
Policy checkbox pre-checked	Deterministic	59.39%	55.81%	82.21%	0.45%
	XGBoost	**99.70%**	**99.40%**	**100.00%**	
	TabNet	**99.70%**	**99.40%**	**100.00%**	
Policy checkbox forced	Deterministic	65.52%	**84.64%**	64.58%	31.89%
	XGBoost	**84.22%**	83.79%	**84.78%**	
	TabNet	80.84%	80.56%	81.16%	
Terms checkbox present	Deterministic	76.45%	75.43%	78.32%	28.44%
	XGBoost	**84.29%**	**84.56%**	**84.13%**	
	TabNet	80.75%	80.75%	80.75%	
Terms checkbox pre-checked	Deterministic	**65.32%**	**60.26%**	84.28%	1.05%
	XGBoost	49.74%	49.48%	50.00%	
	TabNet	56.81%	55.26%	**94.85%**	
Terms checkbox forced	Deterministic	62.13%	76.06%	61.19%	27.40%
	XGBoost	79.44%	79.56%	79.66%	
	TabNet	**81.65%**	**80.32%**	**83.87%**	
Tying marketing and policy checkboxes	XGBoost	**48.77%**	49.07%	48.48%	1.65%
	TabNet	47.30%	**53.00%**	**85.67%**	
Tying policy and terms checkboxes	Deterministic	71.16%	71.48%	70.86%	16.77%
	XGBoost	77.71%	**78.10%**	77.92%	
	TabNet	**80.40%**	77.53%	**85.32%**	
Tying all checkboxes	Deterministic	**51.84%**	51.51%	64.49%	0.45%
	XGBoost	49.70%	49.70%	49.70%	
	TabNet	50.62%	**52.17%**	**93.37%**	
Forced policy	XGBoost	**74.16%**	**74.34%**	**74.07%**	26.95%
	TabNet	67.51%	67.39%	71.50%	
Forced terms	XGBoost	**74.05%**	**80.28%**	70.99%	5.24%
	TabNet	67.57%	64.30%	**74.30%**	
Forced policy and terms	XGBoost	**72.55%**	**72.16%**	**73.25%**	18.41%
	TabNet	63.71%	62.69%	66.76%	

3.3 POTENTIAL VIOLATIONS

In this section, we present the methods for detecting potential violations of the EU regulations, we justify them by references to specific sections of the General Data privacy Regulation (*GDPR*), ePrivacy Directive (*ePD*), other laws, court cases, and guidelines when available. We then present the measurements based on both the manual pilot study and the automated large-scale study.

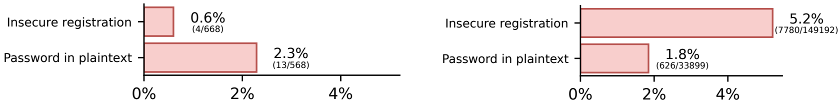
We first present the security violations of the registration procedure, then the potential privacy violations in registration and subscription forms, followed by potential violations in emails, and we conclude by discussion of the overall observations and their possibility of enforcement in the future.

3.3.1 *Security violations*

Using our automated methods, we investigate websites' adherence to security best practices in private data protection mandated by GDPR Articles 25 and 32. We focus on the collection of personal information through user registration and newsletter sign-up processes. We present our findings in Fig. 3.9.

3.3.1.1 *Insecure registration form*

GDPR Article 32(1)(a,b) mandates that the data controller implements appropriate technical measures to ensure the confidentiality of data processing. These measures should consider state-of-the-art methods that are economically viable. The widespread adoption of Let's Encrypt has significantly reduced the costs and technical hurdles associated with implementing encryption via HTTPS. Consequently, the adoption rate of this protocol has surpassed 95% [85]. To simplify our evaluation, we identify insecure registration forms by detecting forms that collect sensitive data over an



(A) Results from the manual pilot study.

(B) Results from the automated study.

Figure 3.9: Security threats of registration to websites.

HTTP (non-HTTPS) connection. Note that this approach may yield false positives in cases where websites employ alternative encryption methods, although these are rare. False negatives may occur when websites use HTTPS but the connection is compromised due to outdated protocols or weak or compromised keys.

During our manual annotation process, we encountered only four insecure forms, which accounts for a mere 0.6% of successful registrations. An additional six insecure registration forms were present in our annotated datasets but marked as failed registrations. Two of these websites were excluded because they were in unsupported languages, while three website forms required information that annotators were unable to provide, such as credit card numbers or membership data. One form required the use of a third-party app to generate a confirmation code.

Our automated web crawler identified 5.2% of websites with forms collecting email addresses or passwords via unsecured HTTP connections. The higher prevalence compared to the manual study[7] can be attributed to several factors. First, the manual study focused on German and English websites, which are countries with a higher adoption rate of HTTPS, as observed by Felt et al. [86]. Our measurements indicated that insecure forms were twice as likely to appear on Russian, Turkish, or Ukrainian websites, while being three times less common on German websites. A more comprehensive analysis can be found in the Appendix in Fig. A.8. Second, HTTP websites are more likely to be outdated or broken, and hence resulting in an unsuccessful registration. Third, our automated approach evaluated any forms collecting email addresses or passwords, including login forms excluded from the manual study, making the automated method more sensitive.

For comparison, Utz et al. [87] found such violations on only 2.85% of websites. The disparity may arise from our more in-depth selection of forms for inspection and differences in the crawling lists.

3.3.1.2 *Sending passwords in plaintext*

After registration, some servicing emails contain either a user-provided password, a generated password, or a password reset link. Sending users the user-provided password by email risks exposure of the user's potentially reused password to anyone capable of reading the emails. Moreover, and quite

[7] We compare the results from the pilot and large-scale studies using Fisher's exact test and apply the Holm–Bonferroni correction to the p-values. We reject the hypothesis that results come from the same distribution when the p-value < 0.001. We perform such analysis for all reported results, results are summarized in Table A.4.

disturbingly, if the server can send the user-provided password for recovery, it implies that the password is not protected by, for instance, hash-and-salt, as recommended by PKCS #5. By not following secure password storage best practices, the service provider risks that a service compromise will expose user passwords that are likely being reused. Non-compliance can also constitute a potential violation of Article 32(1) GDPR. A German Data Protection Authority imposed a fine on a social media provider and held that hashing the passwords of users has been the state of the art for many years [88].

MANUAL PILOT STUDY RESULTS. We inspect how many services send passwords in any of the emails, typically in the confirmation emails right after the registration. We distinguish four cases of what the service sends us: a user-provided password in plaintext (2.3%); a service-generated password in plaintext (3.2%); a password set/reset link (6.0%); and the rest without any passwords (88.5%). When a service sends the user-provided password, we inspect if the same password is sent by when the user requests password recovery. We observe that 20% of these websites send the original password in plaintext.

The various dangers of the account recovery, such as man-in-the-middle attacks on the service-generated password or password set/reset links in plaintext, have been studied extensively (e.g., [89–91]). Also, a list of websites that send passwords in plaintext is curated at `https://plaintextoffenders.com`, although it did not contain any of the websites where we detected this practice. Our study is the first to evaluate the proportion of websites that send user-provided passwords by email. The occurrence of this phenomenon underscores the importance of using password managers to prevent the leakage of reused passwords.

LARGE-SCALE AUTOMATED STUDY RESULTS. We observed 1.8% of websites that send us an email included the user-provided password in plaintext in the email. This is aligned with observations of the manual study, as statistical tests cannot reject the hypothesis that these observations resulted from sampling the same distribution.

3.3.2 *Registration form violations*

In the previous sections, we have described the datasets of emails and websites. In this section, we combine these datasets, using the unique email address as the identifier, and present an overview of overall compliance. Through the combination of the annotated legal properties, we propose a decision tree for the detection of potential violations of the ePD's opt-in requirement and the GDPR's notion of consent.

OPT-IN VIOLATIONS OF EPD. Under the ePrivacy Directive, marketers must obtain an individual's consent (opt-in) before they can send marketing emails. Fig. 3.10 illustrates the opt-in requirements. Websites that engage in email marketing require this consent, which we have annotated using the `ma_consent` legal property. Implicit consent is applicable when newsletter subscription clearly constitutes the primary purpose of registration. In other cases, we must further examine the consent requirements under the GDPR, as we explain below.

GDPR CONSENT VIOLATIONS. As mentioned in Section 3.1.1.1, GDPR mandates that consent must be freely given, unambiguous, and specific. Based on these principles, we present selected potential GDPR consent violations. We describe the combinations of legal properties that lead to a potential violation in Fig. 3.11.

Initially, we identify consent obtained without the provision of a specific marketing email checkbox as unspecific. Moreover, in alignment with case law, we classify the bundling of marketing email consent with other purposes, such as terms and conditions, as unfreely obtained. Furthermore, practices like pre-checked marketing checkboxes and the use of nudging techniques with visual features are classified as ambiguous consent (see Section 3.1.1). Nudging is a typical example of a dark pattern, and we summarize the similarities between potential violations observed in our study and dark patterns in appendix A.1.1.4.

Our decision tree identifies a selection of potential violations. Note that when our procedure identifies no potential violations, it does not necessarily imply compliance with consent requirements. For instance, our procedure does not analyze the specific language used in consent declarations. Nevertheless, our procedure has proven effective in detecting a substantial number of potential violations, as demonstrated in the subsequent sections, using datasets from both manual and automated studies.

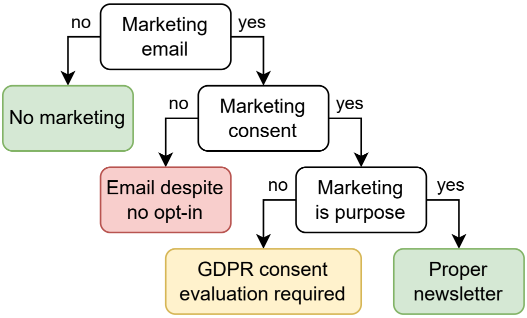


FIGURE 3.10: Decision tree about opt-in validity based on legal properties.

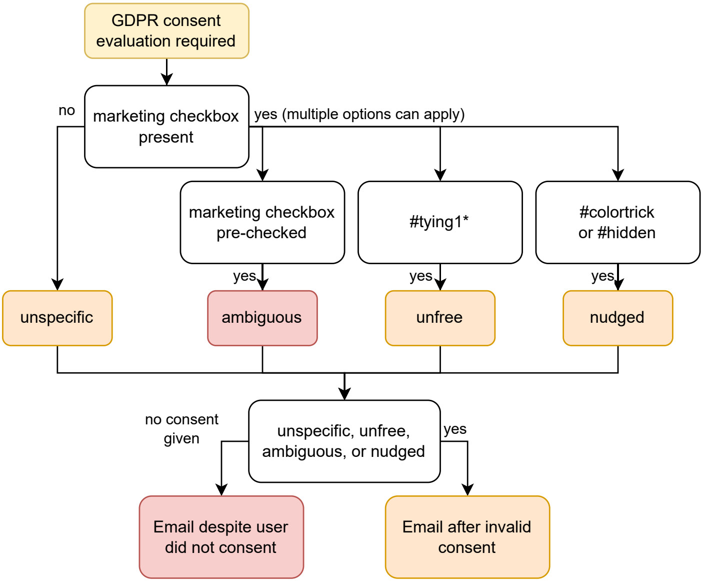

GDPR consent evaluation required
marketing checkbox present
no
yes (multiple options can apply)
marketing checkbox pre-checked
#tying1*
#colortrick or #hidden
yes
yes
yes
unspecific
ambiguous
unfree
nudged
unspecific, unfree, ambiguous, or nudged
no consent given
yes
Email despite user did not consent
Email after invalid consent

FIGURE 3.11: Decision tree about consent validity based on legal properties.

3.3.2.1 *Manual pilot study results*

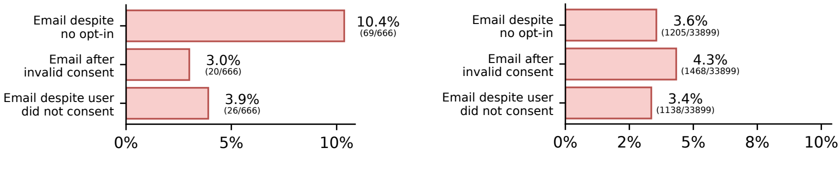
We summarize the marketing consent violations in Fig. 3.12a. Among the annotated websites, 80% of them never sent us marketing emails initially, so they do not require consent in the first place. Among the remaining websites analyzed, 52.3% sent marketing emails despite their registration forms not mentioning marketing emails at all ("Email despite no consent" in Fig. 3.10). This potentially constitutes a violation of Article 13(1) of the ePD. For 12.9% of websites that sent marketing emails, a newsletter subscription was the primary purpose of the registration ("Proper newsletter" in Fig. 3.10). The remaining 34.8% required further assessment for consent requirements under the GDPR, as elaborated below.

Of the websites that sent marketing emails, at least 43.5% did not meet one of the GDPR consent requirements ("Email after invalid consent" in Fig. 3.11). Interestingly, we received marketing emails even from websites that did not violate any of our selected consent requirements. Since annotators were instructed not to provide consent during registration, it is likely that these marketing emails lack valid consent ("Email despite user not opt-in" in Fig. 3.11).

3.3.2.2 *Automated large-scale study results*

In Fig. 3.12b, we present findings from the automated study using the decision trees outlined in Figs. 3.10 and 3.11. Note that the baseline of reported incidence is 33 899 of websites that send any email.

More than 43% of registrations resulted in websites that did not send any marketing emails, potentially influenced by issues related to account activation (refer to Section 3.2.1.4). Additionally, up to 46% of the marketing emails we received resulted from newsletter subscriptions, reflecting the



(A) Results from the manual pilot study. (B) Results from the automated study.

FIGURE 3.12: Portion of senders that violate at least one marketing consent requirement.

crawler's higher success rate with subscription forms compared to registration forms. We found that at least 3.6% of senders violated the opt-in requirement of the ePrivacy Directive by sending marketing emails without any indication of marketing email consent. Furthermore, at least 4.3% of websites violated GDPR consent requirements by not including a marketing checkbox, pre-checking the checkbox by default, or linking the checkbox with privacy policies or terms. In 3.4% of cases, we received marketing emails despite rejecting consent, where the checkbox was neither pre-checked nor checked by the crawler.

It is important to acknowledge that the differences in violation statistics between the manual and automated studies, which are statistically significant only for 'Email despite no opt-in,' cannot be attributed to a single factor. The main contributing factors likely include the (in)accuracy of predictions, particularly the presence of marketing emails, which appears to be significantly higher in the automated study, as well as the detection of marketing consent, which is an abstract legal property and therefore complex to capture by ML.

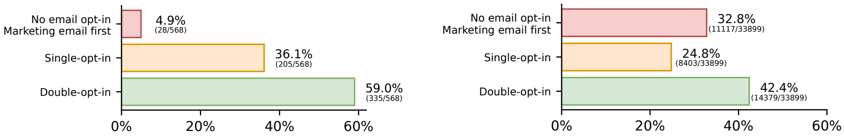
3.3.3 *Email privacy violations*

In this section, we delve into potential violations detected entirely within the emails. Initially, we inspect how websites adhere to the double-opt-in requirement, followed by an examination of whether websites share emails with third parties. We present findings from both the manual and automated studies.

3.3.3.1 *Double opt-in*

In case of legal disputes, companies sending marketing emails must be able to demonstrate that recipients provided informed consent [58, paragraph 6]. To address this requirement, the *double-opt-in* procedure has been established as a best practice, although it is not legally mandatory. Nonetheless, it is highly recommended by legal scholars and the marketing industry [92]. Alternative procedures, such as requiring users to send an email to the service to finish the registration, are not widely adopted. Implementing such procedures can only be partially automated through "mailto" links, which can compromise the usability of the registration process.

For the purpose of this study, we conservatively classify services that only employ a single opt-in as GDPR compliant, even though they fail to follow best practices. Conversely, services that directly send marketing emails without any confirmation email are classified as potential GDPR violations.



(A) Annotated in the pilot study.

(B) Predicted in the automated study.

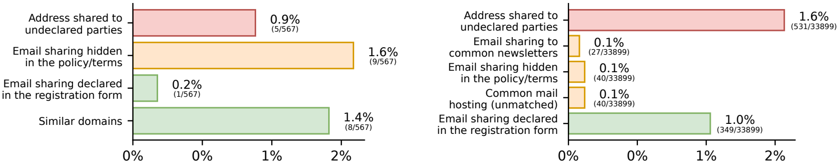
FIGURE 3.13: Classification of the first email from the service.

However, there is a growing body of case law that considers proper double opt-in as a legal obligation. In a recent Austrian case [93], a minor was registered for a dating website by a third party, resulting in the website sending targeted marketing emails to the minor without confirming the email address beforehand. The Austrian Data Protection Authority ruled that such a sign-up procedure did not meet the requirements outlined in Article 32 of the GDPR.

MANUAL PILOT STUDY RESULTS. In Fig. 3.13a, we present the first email received from each service. Only 59% of websites that sent us at least one email initially adhered to the double-opt-in procedure. Moreover, 5.5% of services sent unsolicited marketing emails without any confirmation or double-opt-in email.

AUTOMATED LARGE-SCALE STUDY RESULTS Using the machine learning model described in Section 3.2.2.2, we classify the first email received from websites. The results, as presented in Fig. 3.13b, reveal that 42.4% of websites adhere to the double-opt-in best practice, while 24.8% of websites only send a confirmation email, not conforming to the double-opt-in practice. The remaining 32.8% of websites initiate the communication directly by sending marketing emails to users. The statistically significant higher prevalence of websites directly sending marketing emails compared to the pilot study is likely attributed to differences in website selection, with half of the sample consisting of German websites. This suggests that websites within the EU are more inclined to adhere to the double-opt-in process. Another reason likely includes the misclassification of our methods, as we discuss later in Section 3.4.

Although the double-opt-in procedure is not a legal requirement, its significance is amplified in the presence of registration crawlers like ours. Without this verification, our crawler could be exploited to subscribe arbitrarily email



(A) Annotated in the pilot study. (B) Predicted in the automated study.

FIGURE 3.14: Observed types of email sharing to a third party.

addresses to thousands of newsletters without the owners' consent, potentially leading to the "Bomb attack" [94].

3.3.3.2 *Third-party email sharing*

The collection of email addresses and their sharing with third parties for marketing purposes are governed by the same legal restrictions mentioned in Section 3.1.1.1 [95]. Therefore, third parties sending marketing emails must be able to demonstrate that prior consent was obtained. This requires that users must be specifically informed about whom their email address is shared with and for which marketing purposes [96]. Third parties must therefore be specifically named.

We evaluate whether all emails originate from addresses with first- and second-level domains matching the visited domain. We treat combined top-level domains such as `co.uk` as the first-level domains. However, not every third-party domain (as per the CS interpretation of the term "third party" defined in Section 2.1) corresponds to a legal third party, as domain names may not reflect the legal entities involved. Many websites use dedicated domain names for sending emails, such as `facebook.com` using `facebookmail.com`. In our violation detection, we therefore apply a more lenient approach.

MANUAL PILOT STUDY. We group sender domains and distinguish the following scenarios. First, we observed only one sender domain. Second, sender domains differ only in their top-level domain. Both of these cases we consider as compliant. Third, there are multiple domains that differ in TLD+1.[8] In Fig. 3.14a, we focus on websites sending emails from at least two entirely different domains, as the remaining 95.9% of websites are clearly compliant. We first intuitively inspect whether the domains are similar. Then we examine

[8] Matching was conducted using the `tldextract` Python package.

how websites disclose the sharing of user email addresses with third parties for marketing purposes. Specifically, we manually inspect the content of registration forms, the website's privacy policy, and terms and conditions. If none of these sources inform users about the observed third-party domains, we investigate whether all sender domains are operated by the same group of companies, relying on publicly available sources such as corporate annual reports, Crunchbase, or the WHOIS database.

We conclude that services often send emails from similar domains. Very few services disclose the practice of sharing email addresses with subsidiaries openly in registration forms. Most disclose this only in their terms and conditions, which is legally insufficient. Furthermore, it is well known that such documents are rarely read by users [37, 97, 98]. Over the fourteen months of our study, we observed that one of our email addresses received emails from nine different domains, some of which were not stated in the registration form or terms and conditions. From another service, we received fraudulent emails after a data breach without any notification of the breach by the service.

AUTOMATED LARGE-SCALE STUDY. To mimic the manual inspection of senders with different domains, we developed a heuristic. This heuristic covers a broader range of email sharing types, including common newsletter services. However, we acknowledge that automated methods cannot perform as thorough checks as manual inspections, which include examining corporate annual reports, Crunchbase, or the WHOIS database.

For a given registration, we extract a set of TLD+1 domains from which we receive emails. We then match these domains to other domains found in various sources documenting how websites declare these domains. We consider two domains to match if their longest common subsequence is at least half the length of the shorter domain. This threshold of 0.5 was determined through empirical evaluation of 200 domain matches, resulting in 91% accuracy with 2.5% false negatives (wrongly predicting that domains are not similar) and 7.5% false positives.

For each sender domain, we identify how the website discloses it. We take the first of the following outcomes, ordered from the most to the least disclosed. (1) The domain name where we registered and any domains that are similar are marked as 'registration domain.' (2) The domain of the first received email is marked as 'first sender.' (3) Any common email host (e.g., gmail.com) is marked as 'similar email host' if the name preceding the @ symbol is similar to the registration domain. (4) Any domain declared on the registration page is marked as ‘in form.’ (5) Any common host that was not matched previously as ‘dissimilar email host.’ (6) Domains in the privacy policy and terms and conditions, are marked as ‘in policy/terms.’ (7) Domains that belong to common newsletter senders such as Mailchimp are marked as ‘newsletter sender.’ (8) If all these checks fail, the domain is marked as ‘undeclared.’ We list other methods we considered for third-party sharing detection in appendix A.1.7.1.

If there are at least two senders and one of them is marked as ‘dissimilar email host’ or higher in the ordering above, we consider the website to be sharing the email address without a proper disclosure. As shown in Fig. 3.14b, 1.6% of our email addresses received emails from undeclared domains, including one website that shared our email address to 56 undeclared domains. Additionally, 0.3% of websites sent emails through common newsletter senders such as MailChimp or from domains that were only declared in the policy or terms, which are rarely read [37]. Finally, 1.0% of senders are correctly defined directly in the form, and the remaining websites sending emails do so from expected domains. The prevalence of this violation is comparable to results of the manual study, as statistical tests cannot reject the hypothesis that these observations result from sampling the same distribution.

3.3.4 *Summary and future work*

In Fig. 3.15 we present the potential violations in emails from Section 3.1.3 together with those presented in this section, thereby depicting how many potential violations websites have in total. We aggregate individual missing parts of the legal notice into a single potential violation, while GDPR consent requirements are counted separately. We found 281 potential violations in total, where 148 websites contained at least one potential violation. One website was responsible for five different potential violations, namely they sent marketing emails without opt-in in the registration form, the first email was directly marketing, they shared the address to a third party, and the emails did not contain both unsubscribe method and legal notice.

Although our results show a serious number of potential violations of consent to marketing emails, they do not suggest that such practices are reasons for the majority of unsolicited emails. We propose several explanations that should be inspected in future work. First, studies in the US showed much higher rates of email sharing, namely Mathur et al. [7] found that 12.4% of websites about 2020 US election campaign shared the address. This might be specific to the elections or to the US weaker privacy laws. The latter is

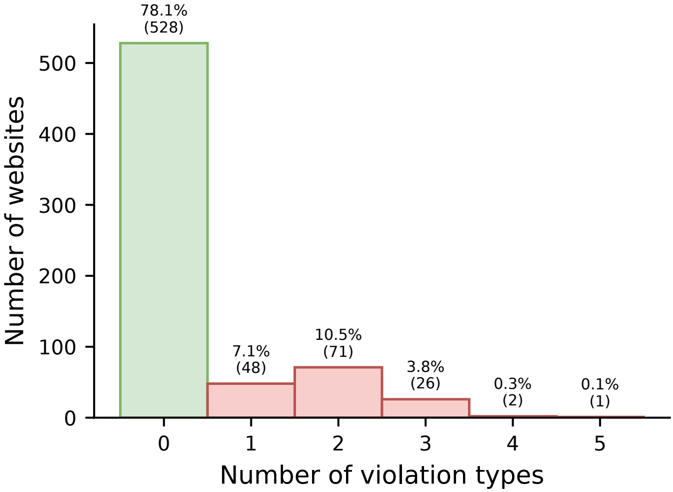


FIGURE 3.15: Histogram showing number of websites with the given number of potential violations. We report the potential violations from 676 websites, i.e., the union of websites where the annotation resulted in successful registration and websites that sent us an email. Note that we are conservative in determining potential violations, so the reported number does not imply that 78.1% of websites are fully compliant.

more likely given that Englehardt et al. [6] observed such violations from 30% of the US e-commerce websites. Future work should inspect comparable samples of US and EU websites to decide whether the GDPR and ePD truly protect users better than US laws, and moreover inspect differences among individual countries, which seem indicated by differences in our manual and automated study. Second, our uniquely generated email addresses limited external factors such as websites guessing our email address or other users registering us. Schneider et al. [94] subscription to services by malicious users as a form of DDoS attack, but such actions could also in a smaller rate come from simple typos during registration by other users. Future research could analyze emails of real users. Finally, our study might need more for observing emails stemming from data breaches, mergers or company rebranding and other events that might cause users forgetting that they ever subscribed to the service.

To complete the compliance picture, we need to consider perspectives from other sciences inspecting further statistics. In appendix A.1.2.2, we explore the violations depending on the website popularity. We present results only from the manual study, which have not found any statistically significant observations due to limited sample size. Such an issue will be addressed by utilizing the large dataset from the automated study, which is subject of interest for future work exploring the compliance of websites depending on various attributes, such as the popularity, topic, or location.

In the future work, we also propose to enforce the law by reporting results to website operators and data protection agencies that can fine website operators for such violations. Our crawler is able to collect contact emails addresses, which we employed in a notification study by Soldner [99]. Before we can start reporting the observations found by our automated methods, we have to establish the trustworthiness of them in the first place, which we do in the next section.

3.4 MANUAL EVALUATION OF THE DETECTION METHODS

The automated violations detection methods of consent to marketing emails depend on a complex combination of predicted legal properties. Computing the overall violation detection accuracy, precision, and recall is impossible as we do not know the correlation of misclassifications. We therefore evaluate the violations empirically.

We manually analyzed a random sample of 100 websites that had sent us at least one email. We selected this sample for two reasons. First, it maximizes the number of websites for which our crawler has successfully filled out the form. Second, websites that had sent us emails serve as a baseline for reporting the majority of violations. Among these 100 websites, our crawler submitted one contact, 54 newsletter subscription, 45 registration forms. Our crawler misclassified six subscription forms as registration forms and one registration and contact form as subscription forms.

Out of the registrations or newsletter sign-ups, the crawler was unable to complete 25 double-opt-in procedures. Note that our evaluation of failed double-opt-ins is conservative since we classified any lack of email confirmation as a failure, regardless of whether the website actually sends such an email. Nonetheless, considering that almost half of the websites use double-opt-in, email confirmation should be improved in future work. Additionally, two registrations were incomplete, but the websites reminded us to finish the registration—a behavior that was studied by Senol et al. [8]. Finally, the crawler successfully submitted the remaining 73 forms.

We examined the email opt-in violations and found that the first emails from 83 websites were correctly classified. Unfortunately, the model misclassified that the first email was for marketing rather than single- or double-opt-in in nine and five cases, respectively. For a subsequent legal study, we completed double-opt-ins manually, which allowed us to inspect 11k classifications of initial email, which we summarized in a confusion matrix in Fig. 3.8b. Note that this dataset is largely skewed to single- and double-opt-in emails, since we were only manually inspecting the first three emails. Nevertheless, the comparison shows that the generalizability of the model is not ideal, which is likely the result of overfitting the model to only 568 websites sending the emails. As future work, we will incorporate the larger annotated dataset for training to improve the mail-type model's robustness.

Regarding form interface violations, our sample contained 17 marketing consent violations. Our method detected 11 of them, with an 86% accuracy, 82% precision, and 50% recall. The two false positives were misclassification of servicing emails for marketing, but the method correctly identified the form interface problems.

For insecure registration and passwords sent via email, the sample had two violations each, and their prediction was accurate. We expect false positives to occur only if we misclassify a form. We evaluated third-party sharing on 50 websites sending emails from multiple different domains. This sample contained 13 violations. Our method achieved a recall of 85% (two short sender domains were falsely detected on the registration page) and a precision of 79% (three senders used multiple domains belonging to the same company which can be observed only from the email content).

In conclusion, while our results reasonably represent the landscape of violations, individual violations are sometimes incorrect. Therefore, individual violations should not be blindly trusted without inspecting the evidence we collected. Still, using our detection methods as a tool for privacy enforcement can considerably streamline the detection of violations, as it presents enforcement agencies with a set of potential violations alongside the evidence needed to manually check whether the violation actually took place.

4 PRIVACY OF COOKIES

Browser cookies are the most common method for tracking the session state of websites. While some cookies are necessary for a website to operate, such as authentication cookies that keep users logged in, the majority of cookies are used for user tracking and advertising (as we show later in Fig. 4.2). Despite the existence of stateless tracking techniques such as browser fingerprinting [100], stateful tracking using cookies remains the primary tracking method. In 2019, Solomos et al. report that almost 90% of all websites use tracking cookies [3], an increase from 80% observed in a study by Roesner et al. from 2012 [2].

Similarly to personal data collection during registration, also user tracking is in the European Union addressed through regulations, namely General Data Protection Regulation (*GDPR*) [30] and the ePrivacy Directive [29]. These regulations specify that usage of all but strictly necessary cookies requires consent, which must be freely-given, unambiguous, specific, and informed, and consented data usage should be limited to concrete and minimal required purposes. Websites must thereby inform users about the purposes cookies are used for, and they must provide users with the option to deny consent for specific purposes.

These regulations have created a demand for prepared consent solutions, from which a new "consent as a service" industry has emerged [101]. The companies offering these services, called consent management platforms (*CMPs*), provide websites with cookie banner implementations that handle the collection of consent from users [102], and offer detailed descriptions of all the purposes that cookies are used for. Unlike the simpler cookie notices, which only inform users about the mere use of cookies, CMPs promise to provide users with more control over their personal data, fulfilling the GDPR's requirements in this area. However, prior studies showed that majority of websites fail to meet the legal requirements [13, 14, 16, 17].

Our analysis. We confirm the lack of GDPR compliance by extending and improving upon past research. We analyze the accuracy of the information displayed on cookie banners, using a dataset collected from almost 30k websites. Specifically, we identify incorrect category assignments, misleading cookie expiration times, and assess the overall completeness of the consent mechanism. We define six novel methods to detect potential GDPR violations

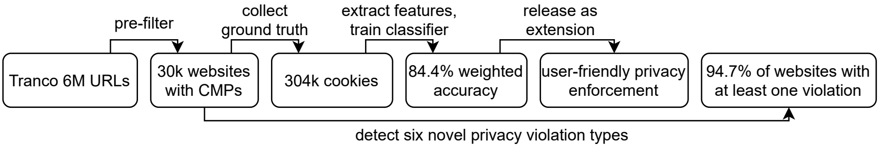

pre-filter
collect ground truth
extract features, train classifier
release as extension
Tranco 6M URLs
30k websites with CMPs
304k cookies
84.4% weighted accuracy
user-friendly privacy enforcement
94.7% of websites with at least one violation
detect six novel privacy violation types

FIGURE 4.1: Overview of the process involved in our study and the results.

and extend two methods used in prior works. For the selected domains, we find that 94.7% contained at least one potential violation.

Browser extension. Based on evidence from prior works and our own measurements, cookie consent practices violate the GDPR so frequently that regulatory authorities cannot hope to keep up. We therefore provide users with a tool to enforce cookie consent on their web clients, without regulatory intervention. We develop a browser extension, CookieBlock, that classifies cookies by purpose, removing those that the user rejects. In this way, users can remove over 90% of all privacy-invasive cookies, without having to trust cookie banners or CMPs. Previous attempts to provide users such control, like the P3P standard [23], failed due to a lack of willingness of website administrators to implement the functionality required. We sidestep this problem by not relying on websites' cooperation at all.

To classify cookies, CookieBlock uses an ensemble of decision trees model, trained using the XGBoost [103] library. We gathered a training dataset of cookies from 29 398 websites that display cookie banners from a specific set of CMPs. Each CMP maintains its own cookie-to-purpose mapping, which we use to define the ground truth class-labels for the cookies in our dataset. Our trained XGBoost model is competitive with the performance achieved by human experts, showing that it is possible to automatically classify cookies by purpose using only the information available in the cookies themselves.

Organization. The remainder of the chapter is structured as follows. Section 4.1 describes our approach to collecting cookie information from CMPs, including the purposes used for training and the data used for the website analysis. Section 4.2 presents our website analysis and demonstrates our approaches to detecting potential GDPR violations. The rest of the chapter is dedicated to addressing these violations. Sections 4.3 and 4.4 describe the features we extract from cookies and our classifier model. Section 4.5 describes and evaluates the browser extension.

4.1 DATASET COLLECTION

In this section we describe how we collected our dataset of cookies annotated with the ground-truth class-labels. This dataset is then used for both the classifier in Section 4.4 and the GDPR compliance analysis presented in Section 4.2.

We collect cookie purposes from consent management platforms (*CMPs*). In contrast to Cookiepedia, these purposes are chosen by the website administrators who control which cookies are created in the users' browsers [104, 105]. As such, we collect the ground truth from parties that have full knowledge about the purposes of cookies, rather than a third-party who may not know the full context. This also allows us to assign categories to cookies that are rare and may be unknown to Cookiepedia. In Section 4.4.1, we show that more than 20% of the collected cookies could not be identified by Cookiepedia.

Our first step is to select CMPs that list cookies with their purposes (Section 4.1.1). Then, from a set of six million domains, we detect the presence of the selected CMPs (Section 4.1.2). For each website where a CMP is used, a web crawler gathers both the cookies declared by the CMP and the cookies that are created in the browser when interacting with the website (Section 4.1.3). Finally, we combine the declarations with the cookies, and obtain the training data for use with our classifier (Section 4.1.4).

4.1.1 *Suitable CMPs and cookie categories*

There are a plethora of CMPs, each offering its own website plugin [102]. These plugins range from simple notifications to elaborate cookie banners that allow users to choose from dozens of possible category options [106]. The purpose assignments we intend to collect can only be retrieved from a small subset of all CMPs. In this section, we describe the criteria we used to select them.

Our first criterion is that the CMP must publicly and reliably list purposes for each cookie on every website where the plugin is correctly implemented. This is essential for collecting the purpose labels that we take as the ground truth. On certain websites, CMPs may offer category choices, but they do not display which cookies belong to which category. Our second criterion is that, when this mapping exists, it must be accessible in a way that can be automatically processed, ideally hosted remotely on a server by the CMP itself. Some websites list the cookie-to-purpose mapping in their privacy policy.

This is generally not useful as the HTML structure of such policies varies greatly between sites, and thus would require a specialized data extraction for each case.

In Table 4.1 we list the CMPs with the highest market-share worldwide, as reported by the technology trend database BuiltWith [107]. For each entry, we list how suitable they are for data collection, based on our criteria. We selected the CMPs OneTrust, OptAnon, Cookiebot, CookiePro, and Termly, here displayed in boldface, which we will use for all subsequent steps of data extraction and analysis.

Table 4.1: Listing of CMPs and their market share in the top 1M websites as reported by BuiltWith [107]. In the third and fourth columns, we evaluate the CMPs with respect to two criteria for collecting purpose labels.

CMP	Market share	Remote?	Labels?
Osano	2.25%	✓	✗
Cookie Notice	1.29%	✗	✗
OneTrust	1.17%	✓	✓
OptAnon	1.08%	✓	✓
Cookie Law Info	0.95%	✗	✗
Cookiebot	0.77%	✓	✓
Quantcast CMP	0.68%	✓	✗
UK Cookie Consent	0.33%	✗	✗
TrustArc	0.26%	✓	✗
WP GDPR Compl.	0.20%	✗	✗
Moove GDPR Compl.	0.18%	✗	✗
tarteaucitron.js	0.16%	✗	✗
Usercentrics	0.16%	✓	✗
CookiePro	0.15%	✓	✓
Borlabs Cookie	0.12%	✗	✓
EU Cookie Law	0.12%	✗	✓
PrimeBox CookieBar	0.09%	✗	✗
Cookie Script	0.07%	✓	✓
Cookie Information	0.06%	✓	✓
Termly	0.05%	✓	✓
Cookie Info Script	0.05%	✓	✗
Easy GDPR	0.04%	✓	✗

4.1.1.1 *Cookie purpose categories*

No law defines which set of cookie purposes the CMPs must declare. Only cookies that are strictly necessary for website operation are recognized, which as per Article 5(3) of the ePrivacy Directive do not require consent from users, and may therefore be set before interaction with the cookie banner.

Given that the categories are not regulated, this selection varies across CMPs. For instance, the Transparency and Consent Framework 2.0 (*TCF*), an industry standard defined by IABEurope, proposes a set of 12 purposes for cookies [108]. Others, like OneTrust, even support the definition of custom categories by the website administrator [104]. In this work, we restrict ourselves to the following four categories, as originally defined by the UK's International Chamber of Commerce [109]:

1. (Strictly) Necessary cookies, which cannot be omitted without breaking the website's main functionality, such as authentication cookies.
2. Functional cookies, which allow for website customization without collecting user data, and are not required for essential services. Examples include user-specific localization and layout customization.
3. Analytics cookies, which serve to track and analyze users' behaviors on a single domain, and are used for aggregated data collection. Google Analytics cookies are common examples from this category.
4. Advertising cookies, which serve to deliver targeted advertisements by tracking users across multiple different domains. DoubleClick or social media websites are common origins for tracking cookies.

In addition to these categories, we also identify unclassified cookies, which will be used for the analysis in Section 4.2. The advantages of the above four categories are that they represent an ordering from the least to most privacy-invasive types of cookies, and that they represent clearly distinct functions. This makes it easier for users to select and distinguish them.

To map the purposes listed in cookie banners to the categories we use internally, we use the keyword mapping shown in Table 4.2. Purposes that do not contain any of the keywords are recorded as 'Other', and are neither used for training the classifier nor for our analysis.

4.1.2 *CMP presence crawler*

After selecting which CMPs to target, we need to find domains that use these CMPs to show cookie banners. To do so, we implemented a fast website scanning procedure using the Python `requests` library to concurrently fetch

Table 4.2: Keywords used to map purposes in CMPs to the selected categories, with the percentage of declarations matched. By * we group multiple suffixes of similar words. The "Other" category contains the cookie declarations that did not match a category, including non-English category names.

Category	**Fraction**	**Keywords**
Necessary	13.2%	essential, mandatory, necessary, required
Functional	8.7%	function*, preference, secure, security, video
Analytics	11.4%	anonym*, analytic*, measurement, performance, research, statistic*,
Advertising	60.9%	ad, advertis*, ads, ad selection, fingerprint*, geolocation, market*, personalis*, personal info, sale of data, target*, track*
Unclassified	3.9%	uncategorize*, unclassified, unknown
Other	1.9%	-

the index page of multiple target websites and scan them for the presence of the desired CMP. If the CMP is used, the website is recorded as being a potential candidate for retrieving cookie labels, and otherwise, the site is filtered out.

Because of the relatively low percentage of websites that use the selected CMPs, and to maximize the amount of collected data, we initialize the presence crawl using a set of nearly six million distinct domains. Our primary source is the Tranco ranking [79] of May 5th, 2021,[1] which lists domains ranked by their estimated worldwide popularity.

Our scan was performed on an AWS EC2 server instance located in Germany, with 32 vCPUs, 64 GB of RAM, and a 10 Gigabit connection. Special care was taken to perform the scan from within an EU country, as previous works have shown that there is significant geographic discrimination

[1] Available at: `https://tranco-list.eu/list/P63J/full`

with regards to GDPR enforcement. Cookie banners are generally less likely to be shown to non-EU visitors [110, 111].

In total, we find 37 587 ($\sim$ 0.63% of 5.94M) candidate domains for the next step of our data collection process.

4.1.3 *Scraping cookie consent information*

The second stage of the data collection process is to extract the cookies and their corresponding purposes from the candidate domains. To do so, we utilize the OpenWPM framework, version 0.12.0 [74, 112], which runs multiple concurrent Firefox browser instances via Selenium. OpenWPM instruments the browser such that all cookie creations and updates are recorded. We call these cookies the *observed cookies*.

We extend OpenWPM to handle data extraction from the CMPs. The gathered information includes at least the declared name, domain, expiration time, and purpose description, as well as the purpose category of the cookie. We will refer to this data as the *declared cookies*. The exact method for retrieving the declared cookies is specific to the CMP implementation. Common to all approaches is that we retrieve the information directly from the JavaScript files that define the consent mechanism. As such, the gathered information should directly relate to which cookies are accepted or rejected depending on the users' choices in the cookie banner.

Our crawl then proceeds as follows: For each domain, after arriving on the landing page, the crawler detects which CMP is actively present on the site. Then the set of declared cookies are extracted. If this proceeds without error, the subsequent steps are intended to trigger the creation of cookies in the browser. First, the crawler consents to all cookie purposes in the cookie banner using the Consent-O-Matic extension [16, 113]. This is required, as otherwise, the lack of consent would prevent cookies from being created. Afterwards, the browser visits random links leading to subpages of the domain, scrolling down to the bottom of each page and performing random cursor movements for each subpage. Urban et al. [114] reported that browsing subpages increases the number of observed cookies up to 36%. As a trade-off between crawling speed and the amount of collected data, we visit ten randomly selected subpages for each site.

The consent crawl was performed on the same AWS EC2 instance described in Section 4.1.2, and took approximately 36 hours for the $\sim$ 37.5k candidate domains. In total, we successfully extracted $\sim$ 2.2 million declared cookies from the cookie banners of 29 398 websites ($\sim$ 72 cookies per site). In addition,

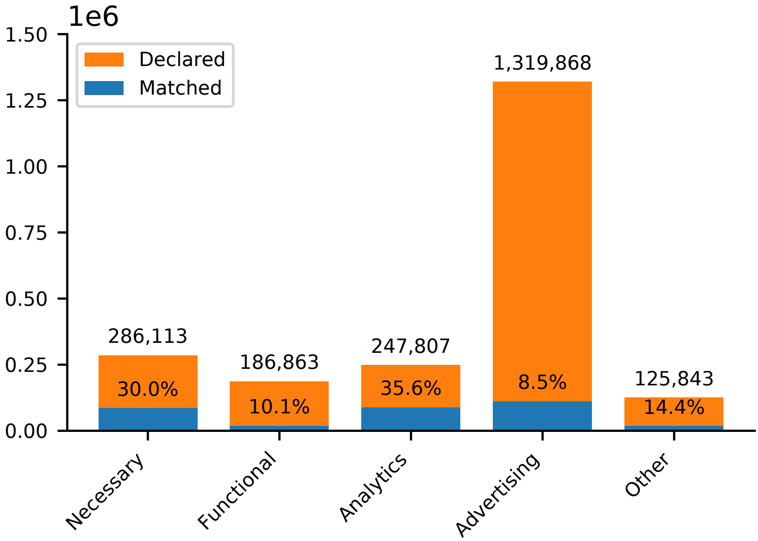


FIGURE 4.2: The total number of cookie declarations with the ratio of observed cookies that match, separated by category.

we extracted 602k observed cookies from those same websites ($\sim$ 22 cookies per site). We find that 81.2% of the declared cookies are third-party entries, while only 46.3% of the observed cookies stem from third-parties.

There exists a discrepancy between the number of declared and observed cookies, which we explain as follows:

Limited automated interaction with the website. Our crawler does not register an account, login or modify the website settings, which can lead to fewer necessary and functional cookies being observed.

Overabundance of declarations. CMPs may list significantly more cookies in their cookie banners than there are actual cookies to be found on the website. Papadopoulos et al. [115] find that users will encounter approximately $\sim$ 12 cookies per site. We observe a mean of $\sim$ 22 cookies, indicating that we do not observe significantly fewer cookies than the related work in the area.

4.1.4 *Obtaining the training dataset*

Our training dataset consists of the observed cookies, with purposes derived from matching cookie declarations. Each cookie is uniquely identified by its name, host, and the target domain of the crawl, and these values are used as the key to join observed and declared cookies. This produces a total of 304k cookie samples for training, of which 28.2% are necessary, 6.2% are functional, 29.0% are analytics, and 36.7% are advertising. An additional 18k cookies are unclassified, or declared a purpose that could not be assigned to any of our categories.

Fig. 4.2 shows the total number of declarations per category, together with the ratio of observed cookies. It is important to note that the category of functional cookies is underrepresented, which we compensate for by weighting the samples when training the classifier. Moreover, despite the overabundance of declarations, out of 602k observed cookies, only 53.6% could be matched with a declaration. This implies that there may be many cookies present on websites that are unknown to the cookie banner. We will discuss this topic in more detail in the next section.

4.2 VIOLATIONS

Article 7 and Recital 32 of the GDPR require that consent must be freely given, specific, informed, and unambiguous; hence any cookie banner that displays misleading or false information may violate the law. In this section, we present an analysis on the data displayed by selected suitable CMPs, performed on a dataset of cookies from 29 398 websites, the collection of which we described in Section 4.1. For these websites, we assess the correctness of the cookie-to-category assignments shown on the cookie banner, the claimed expiration time of cookies, as well as the completeness of the cookie banner. These approaches encompass six novel analysis methods not explored in prior work.

Additionally, we extend the studies of Nouwens et al. [16] and Matte et al. [14] by making use of the cookie purposes collected from CMPs. Namely, we analyze whether websites assume implicit cookie consent or respect the users' consent choices. We accomplish this by observing which types of cookies are set in the browser.

In summary, out of 29 398 websites, 94.7% contain at least one issue, while 77.3% have at least two. A detailed breakdown of the results is given in Figs. 4.3 and 4.4. The following subsections will elaborate on the analysis in greater detail.

4.2.1 *Incorrect cookie purposes*

The CMPs we selected in Section 4.1.1 declare purposes for the corresponding cookies. We inspected the accuracy of these declarations using several complementary methods.

Incorrect purpose for well-known cookies. Google Analytics cookies, such as `_ga`, `_gat`, and `_gid`, occur commonly throughout the web and have a well-known purpose. There nevertheless exist numerous websites that do not declare these cookies as analytics. In the case of Google Analytics, 8.2% of the 29 398 examined websites assign an incorrect purpose to these cookies. Moreover, 2.7% of all websites declare at least one GA cookie as necessary, which the EU Court of Justice previously ruled to be a violation of the GDPR, as decided on the Planet49 case [32].

Incorrect purpose based on the majority opinion. In the collected dataset, we observe that for identical third-party cookie identifiers, different domains may disagree on the purpose. We use it to detect outlier purpose assignments, which likely indicates an incorrect declaration. We find that 30.9% of websites contain at least one third-party cookie with a purpose that disagrees with a corresponding two-thirds majority.

This serves as a lower bound on the number of potential violations. In the event where the majority class is false, the number of potential violations would be even greater. Because this is only a lower bound, each case detected using this method requires manual analysis to determine whether it constitutes a true misclassification.

Cookies with multiple labels. An ambiguity occurs when the same website labels a cookie multiple times for different or even contradictory purposes. We observe this in 2.3% of the examined websites. This ambiguity may deceive users, as it is not well-defined whether rejecting only one of the purposes suffices to prevent the cookie's creation. In practice, we observed websites creating cookies with one purpose accepted and one rejected. Moreover, in 0.7% of the sites, the cookie is declared both as necessary and another purpose, which means that these cookies cannot be rejected at all.

4.2.2 *Unclassified and undeclared cookies*

The CMPs we target in our study offer a cookie scan service that detects cookies on a website and suggests purposes based on a database lookup. Those cookies that cannot be annotated in this fashion must have their purposes assigned manually by the site administrator [104, 105].

We find two problems with this process. First, when the web administrator neglects to assign a purpose, the cookie becomes unclassified. Second, when the CMP scan fails to detect cookies, or the cookies are added after the scan, those cookies are undeclared and are missing from the cookie banner. The website's visitor can reject neither the undeclared nor unclassified categories, which means that the consent is both uninformed and not freely given.

Unclassified cookies. We find unclassified cookies in 25.4% of the examined websites. These websites contain on average 11 unclassified cookies. Surprisingly, we find 45 websites that contained more than 200 unclassified cookies.

Undeclared cookies. We detect undeclared cookies by identifying which observed cookies do not have a matching declaration. When matching on name and domain, we find undeclared cookies in a staggering 82.5% of the examined websites. Of the 496k cookies, 40.2% were undeclared. Similarly to unclassified cookies, we find 71 websites with more than 100 undeclared cookies.

4.2.3 *Incorrect expiration time*

Article 13(2)(a) of the GDPR requires websites to declare the expiration time of personal information. The EU Court of Justice in the Planet49 case decision [32] clarifies that this also applies to cookies. We therefore compare the true expiration time of the observed cookies with that of the corresponding declaration. If the true expiration time is 50% longer than the declaration states, with a minimal difference of one day as threshold, then we consider it a potential violation. Additionally, we also identify all persistent cookies that are declared as session cookies, and vice-versa. In total, 9.1% of all sites show at least one expiration time discrepancy, 3.8% declare a persistent cookie as a session cookie, and 3.1% declare a session cookie as persistent.

4.2.4 *Extension of previous approaches*

The following two approaches extend methods defined in the works of Matte et al. [14] and Nouwens et al. [16]:

Cookies set prior to user's consent. Article 5(3) of the ePrivacy Directive states that only necessary type cookies may be created prior to the user's interaction with the CMP. By crawling the website without interacting with the cookie banner, we inspect if websites set any cookies with a purpose that is not declared as necessary. We find that 69.7% of the examined websites

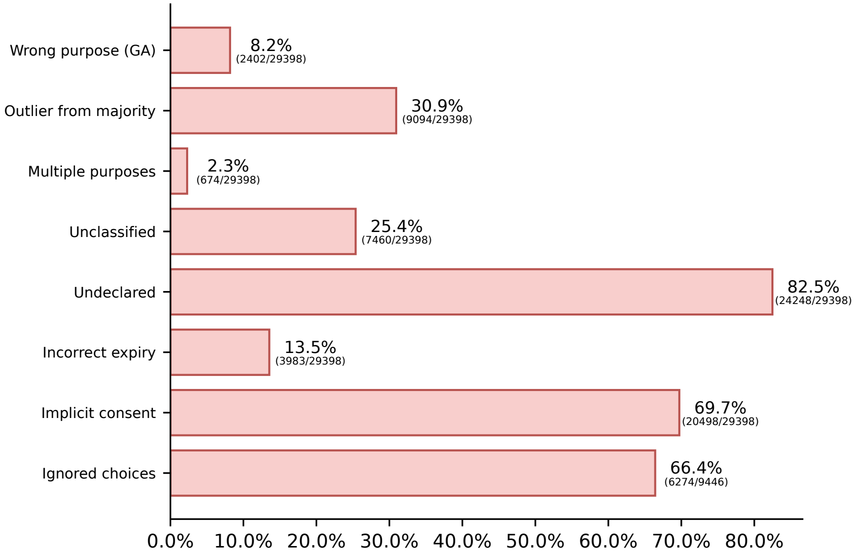


FIGURE 4.3: Number of websites that show the respective type of violation. The first six are novel and have not been explored in prior work.

set such cookies, and hence use implied consent. In contrast, Nouwens et al. [16] found implicit consent in the form of a missing button on 32.5% of 680 sites and Matte et al. [14] only found implicit consent by inspecting the consent string recorded by the CMP on 9.9% of 1426 analyzed websites. Our approach is more fine-grained than both of these works, as we analyze the actual cookies set in the browser before the consent is granted.

Cookies set despite negative consent. Using the Consent-O-Matic browser extension [113], we reject all purposes other than necessary. We then verify that the recorded consent status of the CMP is indeed negative, and identify which of these websites still set non-necessary cookies. We do this only for Cookiebot, as for this CMP we can verify whether the cookie banner was interacted with. For the 9446 Cookiebot domains, 66.4% set at least one cookie with a rejected purpose. For comparison, Matte et al. [14] found that 4.8% of 560 analyzed websites store user's positive consent to categories that the user rejected, but their measurement is dependent on presence of CMP's consent string and is therefore less sensitive than ours.

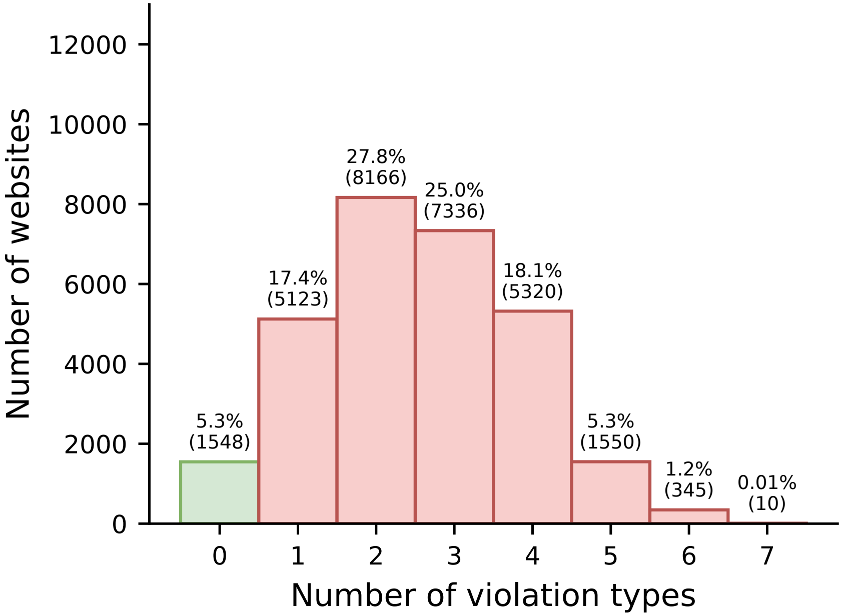


FIGURE 4.4: Histogram that shows the distribution of violation types per site. This does not include repetitions of a single violation type by multiple problematic cookies. The green bar represents the compliant websites.

4.2.5 *Summary*

Fig. 4.3 summarizes the number of potential violations for each of the types we described above. In Fig. 4.4, we present how many different violation types are present on websites in our dataset. The histogram shows that the median number of violations is 2 and the average is 2.5.

The first six bars in Fig. 4.3 represent analysis methods that, to the best of our knowledge, have not been explored in prior works. The latter two extend analyses previously performed by Nouwens [16] and Matte [14], yet our approach is more fine-grained and direct, as we directly detect the cookies created in the user's browser, based on the purposes declared in the cookie banner. Our sample size of websites is also much larger than in both their works.

For the case of unclassified and undeclared cookies, we believe that the issues usually stem from neglect rather than malice. The cause is likely the lack of enforcement and web administrators who are not sufficiently familiar with the legal requirements. This can be addressed with the methods described in this paper. Regulatory authorities can improve enforcement of the GDPR by automatically determining which websites violate the law. Moreover, CookieBlock and the corresponding web crawler can help web administrators inspect the compliance of their website by detecting undeclared cookies, and predicting purposes for currently unclassified cookies.

We also repeated the experiments after one year, which we illustrate in Fig. A.9. The new results suggest that while adoption of our selected CMPs is increasing, the number of violations is mostly stationary. This supports the representatives of our results.

Our analysis has shown that the vast majority of websites are violating the legal requirements and therefore the consent implemented on server side cannot be trusted. In the next three sections, we describe the process needed for implementing the browser extension CookieBlock that enforces consent on the client side.

4.3 FEATURE EXTRACTION

Cookies have multiple attributes, including a name, domain, path, value, expiration timestamp, as well as flags for the "HttpOnly," "Secure," "SameSite," and "HostOnly" properties. There is no straightforward relationship between these attributes and the cookies' purpose. Therefore, we extract statistically-rich, domain-specific features so that a machine-learning model can extract a potentially complex, meaningful relation from the data.

We define more than 50 feature-extraction steps that represent a cookie as a real-valued sparse vector. We provide a high-level account of these steps below. More details are provided in Appendix A.2.2 and the full description is given in the extended report [25] and documentation.[2]

Top-500 most common names and domains. A very effective method for identifying a cookie's purpose is detecting whether the cookie name or its origin domain are among the most common identifiers found online. Using a representative random sample of websites from our Tranco list, we collect a ranking of the 500 most common cookie names and domains. The intuition is that web modules use first-party cookies with predefined names and purposes,

[2] The feature documentation and classifier are available at: `https://github.com/dibollinger/CookieBlock-Consent-Classifier`.

such as PHPSESSID in the case of PHP, and that cookies originating from the same domain usually have a common purpose.

Value type, encoding, and length. Several of our features indicate the presence of specific data types in the cookie content. This ranges from scalar types such as Booleans or integers to composite types such as CSV or JSON. We also record the number of entries for composite types, as well as the length of the content in bytes as ordinal features. We furthermore distinguish between decimal and hexadecimal integers, as well as base64 and URL encoded strings. The intuition is that by identifying the types of data stored in a cookie, the classifier can better distinguish which cookies are used for tracking. For example, long hexadecimal strings are more likely to be used for uniquely identifying a user than short decimals.

Dates, timestamps, UUIDs, URLs, or locale strings. These values may provide hints about the purpose of a cookie. Intuitively, dates, UUIDs, and timestamps may be used as unique identifiers for tracking, while locales and URLs are more commonly used with functional cookies, for example to alter the display language or input method.

Update features. Cookies are dynamic, and can be frequently updated by HTTP requests or through events in JavaScript code. As such, we not only consider features for a single state of the cookie, but also for changes that occur over time. Examples are the total number of times a cookie is updated over a fixed time interval, or the edit distance between cookie updates.

Cookie entropy. The entropy of the cookie's content, for example computed using Shannon's method, can provide information about its randomness. The intuition is that tracking identifiers often include a randomly generated component and hence have high entropy, thus potentially allowing the classifier to detect tracking cookies.

Note that not all cookie features can be used in all settings. For instance, in our dataset, advertising cookies are updated more rarely than other types of cookies. While this property could be used as a feature for training, it is highly dependent on the user's browsing pattern. Any features that are based on such patterns are unreliable in the setting of a browser extension, and may cause false predictions that cannot be observed during the model validation. For CookieBlock, we therefore only use those features that are agnostic to browsing patterns. Nevertheless, such properties may still be used for offline settings with a fixed browsing behavior, such as studies involving automated web-crawlers.

4.4 ML FOR COOKIE PURPOSES

In this section, we present the design and evaluation of our cookie purpose classifier. We first describe the baseline, which is the manually constructed repository Cookiepedia (Section 4.4.1). Next, we explain our choice of model (Section 4.4.2) and the selected hyperparameters (Section 4.4.3). We explain the impact of different types of misclassifications (Section 4.4.4), and present our model's performance, comparing it with the selected baseline (Section 4.4.5). Finally, by estimating the degree of noise in the data, we estimate the best possible classifier performance for this dataset (Section 4.4.6).

4.4.1 *Baseline*

We compare our model's performance to that of a manual classification by experts in the field. Namely, we query cookie purposes from the public cookie repository Cookiepedia [116]. Cookiepedia reportedly stores data for over 30M cookies, of which a large portion has been labeled with purpose categories. These categories match the ones we have chosen in Section 4.1.1.1. For our dataset, Cookiepedia provides purposes for 79.2% of the cookies.

To use Cookiepedia as a classifier, we query it for each cookie name in our dataset and obtain the corresponding purposes from the repository. These purposes are then compared to the class labels we collected from the CMPs. To validate Cookiepedia as a classifier, we split the cookie dataset into 5 equally-sized chunks and compute the average accuracy, precision, and recall. In Table 4.3 we present the results.

Our measurements show that Cookiepedia achieves a mean balanced accuracy (i.e., macro-recall) of 84.7%. It achieves a high precision for both necessary and advertising cookies, but has particularly low precision for functional cookies. This can be explained through the class imbalance we find in the validation data. Due to the low number of samples for the functional ground truth, any error that assigns this category to other cookies will have a much greater effect on the precision of this class than it would have for the other categories.

4.4.2 *Model selection*

Our chosen model for the task of classifying cookies are ensembles of decision trees. We train them using the XGBoost library [103], which uses a sparsity-aware gradient tree boosting method developed by Chen and Guestrin. We use boosting because ensembles of decision trees can be as competitive as neural networks and have achieved top performance in several machine-learning competitions and benchmarks [117–119].

In the setting of multi-class classification, XGBoost creates a classifier model with a forest of decision trees for each purpose class. Given a sparse input vector representing a cookie, the model produces a probability for each purpose that indicates how likely the cookie belongs to it. Using a Bayesian Decision function, we transform these probabilities into a discrete prediction. For our evaluation, we apply a simple `argmax` decision, i.e., the purpose with the highest probability is chosen as the prediction.

4.4.3 *Training parameters*

The dataset we use consists of 304k labeled cookies, of which 277k are used for training. The 27k cookies we filter out are cookies created by CMPs to track users' interaction with the cookie banner. With this filtering, we aim to remove training bias as these cookies are always present on the sites we crawled, but are not common outside the chosen websites.

To find good hyperparameters, we applied a randomized grid-search with 5-fold cross-validation. The performance of each model is validated using the multi-class cross-entropy loss, as well as the balanced accuracy, due to the training dataset being imbalanced. The most impactful parameters were the learning rate and the maximum tree depth, for which we selected a rate of 0.25, and a depth of 32, respectively. Further increasing the depth leads to a decrease in the validation performance. We trained each model for a maximum of 300 boost rounds, with early stopping after 20 rounds with no increase in validation score. For the final model, there are 12 to 29 trees per forest, with the average size being 22 trees. The complete set of parameters is shown in the Appendix in Table A.6.

4.4.4 *Impact of misclassifications*

As mentioned in Section 4.1.1.1, our selected purpose categories can be interpreted as an ordering, with necessary being the least and advertising the most privacy-invasive. Using this ordering, a misclassification of a functional cookie into the necessary category has reduced privacy impact, as the functional cookie is close in the ordering, and unlikely to be used for user tracking. A wrong assignment of an advertising cookie to necessary represents a greater

TABLE 4.3: Performance metrics for the Cookiepedia lookup. Evaluated using 277k cookies, as an average over 5 folds.

Cookiepedia	Necessary	Functional	Analytics	Advertising
Precision	94.5%	38.1%	84.2%	94.9%
	±0.2%	±0.6%	±0.2%	±0.1%
Recall	88.5%	78.7%	93.0%	79.0%
	±0.1%	±1.1%	±0.1%	±0.2%

Cookie coverage: 79.2%

Accuracy: 86.1% ± 0.1%

Balanced accuracy (macro-recall): 84.7% ± 0.3%

TABLE 4.4: Performance metrics of the XGBoost classifier in categorizing cookies, trained on 277k samples and evaluated with 5-fold cross-validation.

XGBoost	Necessary	Functional	Analytics	Advertising
Precision	87.3%	52.9%	89.8%	93.6%
	±0.2%	±0.5%	±0.3%	±0.2%
Recall	81.7%	76.3%	89.7%	89.8%
	±0.5%	±0.5%	±0.2%	±0.3%

Cookie coverage: 100%

Accuracy: 87.2% ± 0.23%

Balanced accuracy (macro-recall): 84.4% ± 0.27%

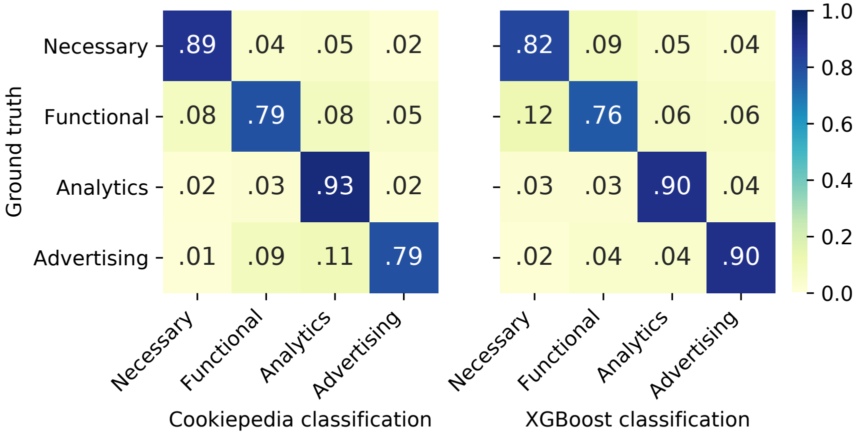


FIGURE 4.5: Confusion matrices of the Cookiepedia baseline and XGBoost. Each entry C_{ij} shows the ratio of cookies with ground truth i that were assigned purpose j.

privacy threat as these categories are far apart in the ordering, with tracking cookies potentially being unconditionally permitted.

Similarly, we also consider the potential of websites breaking due to misclassifications. When a necessary cookie is predicted as advertising, and thereby removed, it may break an essential service on the site, and drastically reduce the quality of the user experience. Assigning the class functional to a necessary cookie has a reduced impact as users are less likely to reject this purpose due to it being less privacy-invasive.

The probability with which advertising cookies will evade detection can be identified using the recall metric of the advertising class. The potential to break essential functionality on websites can be found in the recall of the necessary category. The closer either performance metric is to 1, the lower the privacy threat, respectively the less likely a website is to break.

4.4.5 *Evaluation*

Fig. 4.5 compares the performances of XGBoost and Cookiepedia. Table 4.4 presents the performance metrics for our XGBoost model. We discuss them next.

XGBoost attains higher privacy protection. In accordance with Section 4.4.4, we first consider the potential privacy protection through the recall of the advertising category. Here, the recall measures the fraction of advertising cookies correctly identified as advertising by our classifier. XGBoost's recall is almost 9% higher than that of Cookiepedia. In Fig. 4.5, we see that Cookiepedia's misclassifications in this regard occur mainly because it assigns advertising cookies to the analytics or functional class.

XGBoost preserves necessary and functional cookies. We consider the potential for websites breaking. The recall for necessary cookies for the XGBoost classifier is 81.7%, almost 7% lower than what Cookiepedia achieves. For functional cookies, we have a recall of 76.3%, roughly 2% lower than Cookiepedia. Fortunately, as we see in Fig. 4.5, most of the misclassifications of necessary are assigned to the functional purpose, and vice-versa. Therefore, if users accept both necessary and functional cookies, the extension will retain approximately 91% of the necessary and 88% of the functional cookies. We verify this empirically in Section 4.5.3.

XGBoost is as competitive as human experts. Our automated XGBoost model performs very similarly to the manually curated Cookiepedia in the remaining metrics. Both have a reduced precision and accuracy in functional cookies, which occurs due to the class imbalance. Additionally, both achieve a high recall for the analytics class, with XGBoost achieving an improved precision by more than 5%.

To summarize, Cookiepedia achieves a balanced accuracy of 84.7% on our dataset when queried for each cookie name. Our automated, XGBoost-trained classifier achieves a balanced accuracy of 84.4%, thus attaining a performance that is comparable to the performance achieved by human experts. While Cookiepedia is more accurate in the necessary category, XGBoost performs better with advertising cookies. Our deficit in necessary cookies can be counterbalanced by using an alternative Bayesian cost function, which penalizes misclassifications of necessary cookies more strongly than others. We can also provide users of CookieBlock with ways to correct the classification, which we describe in Section 4.5.

Finally, the number of cookies that Cookiepedia can classify is limited. For our dataset, Cookiepedia is able to provide a category for 79.2% of the cookies, while our classifier can predict a class for every cookie.

4.4.6 *Performance upper bound*

In this section, we try to estimate the theoretically best classifier performance on our dataset. The cookie labels we collected are noisy, as different websites can use the same third-party cookie, but they do not necessarily agree on its purpose. We reported this disagreement also in Section 4.2.1 as Outlier from majority violation. The presence of this disagreement means that it is impossible to achieve 100% accuracy on this dataset, as some cookies will be indistinguishable despite differing purposes. To estimate the percentage of cookies in the dataset for which this is the case, we collect the majority class for each third-party cookie name and domain, and compute the percentage of cookies with a deviating class. This gives us a lower bound of 7.2% of labels that are noise among the third-party cookies.

If we assume that the noise of the first- and third-party cookies is similar, we can conclude that we have an upper bound of roughly 92-93% in overall accuracy. With an overall average accuracy of 87.2%, we argue that our classifier is close to the best possible performance on this dataset.

4.5 BROWSER EXTENSION

In this section, we describe the design and implementation of CookieBlock.[3] It is an extension for Firefox and Chromium-based browsers that automatically classifies cookies into purpose categories, and allows users to deny consent for selected purposes. By using the classifier described in Section 4.4, we provide users with a tool to enforce the GDPR and protect their own privacy when handling cookies. Over the 2.5 years since its release, CookieBlock has attracted over 13k installations. Demir et al. [120, Sec. 7] showed that among other cookie consent privacy extensions, “CookieBlock showed the best performance regarding blocking (deleting) unwanted cookies.” They however recommend combining it with extensions blocking other tracking technologies, such as uBlock Origin, which we also justify below.

We first discuss the goals and features of CookieBlock (Section 4.5.1). Then we present its design and implementation (Section 4.5.2). We conclude the section with an empirical evaluation on a set of 100 websites that estimates how CookieBlock affects users’ browsing experience (Section 4.5.3).

[3] The code is available at `https://github.com/dibollinger/CookieBlock`, along with links to browser extension stores.

4.5.1 *Goals and Features*

The objective of CookieBlock is to give users control over their privacy, a practice that is neglected by the majority of websites. Table 4.1 indicates that out of the top 1M websites, only an accumulated total of 3.5% use CMPs providing consent choices according to cookie purpose. Many of those that do deceive users either by dark patterns, as shown by Nouwens et al. [16], or by providing wrong information, as we show in Section 4.2. Hence CookieBlock provides users with a means to control their cookie consent on any website they visit, without the risk of being deceived. CookieBlock offers the following features:

USER-DEFINED COOKIE POLICY. CookieBlock's central feature is that users specify which of the four categories they give or deny consent to. All cookies belonging to a purpose for which consent was denied are then removed from the browser's storage.

DOMAIN EXCEPTIONS. For domains that the users trust, they can define an exception. The extension will not remove any cookies originating from exempted domains, regardless of their purpose.

CUSTOM COOKIE CLASSIFICATION. Users can reclassify cookies, which can be used to correct individual mistakes made by the model.

Note that while CookieBlock imitates the behavior of a CMP, it is not intended to interact with or remove the cookie banners shown on websites. This function is already fulfilled by existing browser extensions, such as I don't care about cookies, uBlock Origin, or Consent-O-Matic, which can be used in conjunction with CookieBlock. CookieBlock also does not act as a replacement for the cookie banner in the legal sense, and its use is not a justification for websites to skip the gathering of user consent.

4.5.2 *Design and implementation*

CookieBlock is built using the WebExtensions API, and supports Firefox as well as Chromium-based browsers. An overview of its design is given in Fig. 4.6.

4.5.2.1 *Background process*

On initialization, CookieBlock begins listening for cookie events. When a cookie is created or updated, the cookie's current state is appended to a local

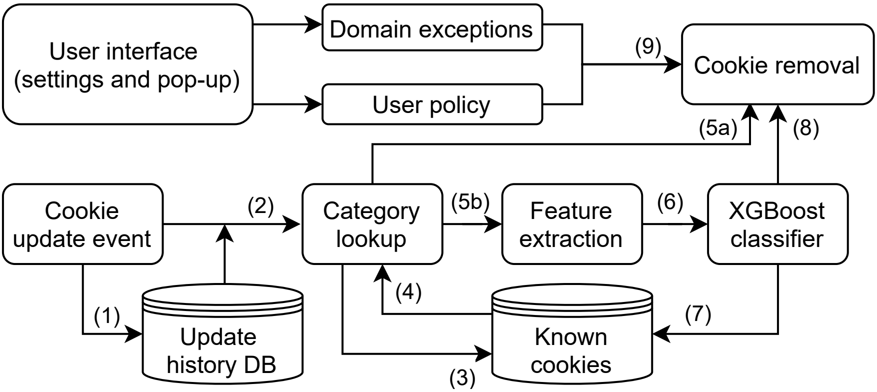


FIGURE 4.6: Outline of CookieBlock's design.

cookie history (1), and the full list of previous updates for that cookie is retrieved (2). This history allows CookieBlock to track the evolution of a cookie over time, a property which is used in the feature extraction.

Afterwards, CookieBlock checks its storage to determine whether the cookie has been encountered recently or whether it has been assigned a pre-defined category (3). In this case, it retrieves the existing purpose label (4), and skip directly to the policy enforcement step (5a). If the cookie does not have an existing label stored, then we proceed to the feature extraction (5b). This transforms the cookie object into a sparse vector representation (6). It then runs the precomputed XGBoost model on this vector input, which predicts a purpose label for the cookie. The predicted label is then cached in the extension storage for a short duration (7). Finally, the predicted label is passed on to the policy enforcement procedure (8), which decides whether to keep or remove the cookie.

To decide whether to keep or remove a cookie, CookieBlock takes into account the user's cookie policy and domain exceptions (9). If the origin domain of the cookie matches a domain in the set of domain exceptions, the policy enforcement will always retain the cookie.

4.5.2.2 *User interface*

The user interface is structured into four distinct components:

FIRST-TIME SETUP. The first-time setup page of the extension allows the user to define a user-policy, and requests consent to the collection of a local cookie history. This is the minimal setup required to initialize the extension.

SETTINGS PAGE. The settings page allows users to change their consent preferences at any time and add individual website exceptions.

TOOLBAR POPUP. The toolbar popup offers a quick method to pause the cookie removal and to add an exception for the domain in the address bar.

COOKIE CONFIGURATION. The cookie configuration page allows users to define custom categories for previously encountered cookies and to correct misclassifications.

For both the settings page and the first-time setup, CookieBlock allows the user to consent to the functional, analytics, and advertising purposes. The necessary category cannot be rejected as doing so would break websites.

We designed the interface to be simple to use and unobtrusive. Unlike cookie banners found on different websites, CookieBlock requires only a single setup, after which the users' cookie preferences will be enforced on all websites. This prevents the issue where privacy is neglected due to user fatigue or annoyance from cookie banners [121, 122], as well as the violations that we report in Section 4.2.

4.5.2.3 *Cookie update history*

As described previously, CookieBlock collects a cookie update history. This allows it to track how cookies change over time, enabling predictions based on these differences. It also allows CookieBlock to remember past purpose assignments by recognizing which cookies have been encountered before.

Since this cookie history may contain potentially sensitive user information, including information about the browsing history and authentication tokens, the history is kept local to the browser extension at all times. In addition, CookieBlock asks the user to opt-in to the collection of this history at setup time. If rejected, CookieBlock can still classify cookies, but it will not be able to remember previous labels or extract features from past updates, which may reduce its accuracy.

4.5.2.4 *Cached purposes*

CookieBlock caches labels for a short period after a prediction is made. This minimizes browser slowdown in case a website continuously regenerates cookies after they have been removed. After the grace period expires, the cookie will be reclassified using newly collected cookie updates.

4.5.3 *Empirical evaluation*

As noted in Section 4.4.5, our classifier has a recall of 81.7% on necessary cookies, meaning that potentially every fifth cookie required for the operation of the website could be misclassified. Since CookieBlock uses the computed model as a predictor, many necessary cookies may inadvertently be removed, causing websites to malfunction. However, due to the noise in the dataset, it is unclear how severe this issue is in practice.

To quantify the impact CookieBlock has on the browsing experience, we manually visit and examine a sample of 100 websites for possible malfunctions. We acknowledge that this evaluation is limited in that it does not constitute a full usability study. However, because the extension acts as a background process, it should ideally require very little interaction with the user. We therefore focus on evaluating whether a website breaks due to misclassification, which is the critical aspect of usability in this case.

We randomly sample websites from the Tranco list from Section 4.1.2 using an exponential distribution. This allows us to examine both popular as well as niche websites. Furthermore, this website selection is not restricted to those that use specific CMPs.

We use a clean installation of CookieBlock, configured to allow necessary and functional cookies, which is the recommended setup. For each website, we attempt to make use of its primary services as best as possible, recording any defects we encounter in the process. We also attempt to change website settings, such as the language or style, and we attempt to register an account and perform the login procedure where available. Finally, we also interact with and close cookie banners, recording whether any appear again on page reload. A reappearing cookie banner can be very annoying for the user, but it does not prevent the site's use, and therefore these are likely misclassified functional cookies. If we encounter any unexpected behavior, we determine whether this was caused by CookieBlock by disabling the cookie removal.

Our results show that out of the examined 100 websites: 85 showed no obvious malfunctions, 7 had a cookie banner that reappeared because of CookieBlock, 7 showed an authentication failure where the user was immediately logged out, and in one case, we could not change the website language. As such, the rate of serious defects is less severe than expected. Furthermore, all issues were resolved by defining an exception for the current site, or by correcting the cookie's assigned purposes in the extension interface.

We also measured the time it takes for CookieBlock to make a policy decision for a cookie. We ran CookieBlock on the Firefox browser on Linux, and it processed a total of 5561 cookies observed from real-world websites. Each decision took on average $\sim$ 20ms, with a maximum time of 4.3 seconds. This outlier was caused by asynchronous execution in the browser. The Firefox browser also reports a “low” energy impact for the extension.

4.5.4 *Subsequent user study*

Schöni et al. [28][4] conducted a user study on CookieBlock and performed an expert evaluation comparing CookieBlock to other privacy extensions. The user study, involving $N = 42$ participants, focused on understanding users’ mental models of CookieBlock, how they dealt with website breakages caused by CookieBlock, and their overall user experience ratings. Additionally, $N = 4$ experts evaluated the privacy-usability trade-offs of browser extensions based on their functionality and implementation.

During the user study, participants installed CookieBlock and interacted with two e-commerce websites. On one of these websites, CookieBlock’s misclassification of an authentication cookie caused website breakage and required users to identify the issue and resolve it by granting an exception or pausing cookie removal. After the installation, interacting with each website (in a randomized order) and completing the experiment, participants filled out a survey regarding their mental model and provided a System Usability Scale (*SUS*) [123] rating for CookieBlock.

Involving participants representative of the general population revealed significant challenges associated with using CookieBlock. Approximately half of the participants constructed incorrect mental models of CookieBlock, leading to a lack of awareness about potential website breakage caused by the extension. Nearly half of the participants were unable to log in to the website affected by CookieBlock’s misclassification. These participants were informed after 5 minutes that CookieBlock was causing the issue; however, three participants still struggled to resolve the problem.

The experience of encountering website breakages had a notable impact on participants’ perceptions of CookieBlock’s usability. Initially, users rated the extension as having good usability (with an SUS score of $M = 76$ out of 100). After experiencing website breakage, the rating dropped to an average usability level ($M = 61$). Surprisingly, participants who required assistance did not assign lower SUS scores. These usability scores remained high when

[4] Despite my involvement in this study, I refer to it in the third person, and the contributions remain attributed to Lorin Schöni. This reference is provided here to offer additional context absent in our original study.

compared to previous user studies on privacy extensions, including SUS scores of 79 for Ghostery, 60 for DuckDuckGo Privacy Essentials, and 62 for Privacy Badger [124].

The identified issues related to mental models and breakage resolution remain significant challenges for CookieBlock. Although this user study amplified the breakage incidence (Section 4.5.3 suggests that only 7% of websites are prone to such breakage by CookieBlock's misprediction), the results suggest that the extension targets mostly advanced users. To improve users' mental models, Schöni et al. [28] implemented various interface modifications aimed at better clarifying intended usage. Additionally, they proposed a heuristic as future work to warn users when cookie removal might lead to website breakages.

4.5.5 *CookieAudit*

Given the results from the CookieBlock user study, it is unrealistic to expect widespread adoption of client-side privacy enforcement tools. Therefore, we have also released a modified version of the extension, CookieAudit [125], which targets website operators seeking compliance with privacy regulations but encountering difficulties in implementing cookie consent free of violations. This extension allows website operators to conduct more comprehensive self-audits of their websites compared to the automated crawler described in Section 4.1. Currently, CookieAudit supports only CookieBot and OneTrust CMPs.

CookieAudit guides users through interactions with the audited website to identify various privacy violations. It supports a quick scan that identifies dark patterns in cookie notices, including missing reject buttons and pre-selected choices. It also detects non-essential cookies being used prior to consenting cookies or after rejecting them. In the advanced scan, it instructs users to interact with the notice in specific ways to identify undeclared cookies and cookies with wrong category or expiry. After the scan, CookieAudit generates a report, including suggestions for addressing identified issues and explanations of privacy regulations and the implications of violations.

While we still cannot answer the question of why so many websites violate cookie notice requirements, as discussed in Section 4.2, CookieAudit addresses one potential knowledge gap, namely, when website operators are unaware of what constitutes a privacy violation. Investigation into the reasons for non-compliance is left as a topic for future research.

5

LIMITATIONS

In this chapter, we discuss the limitations of our studies and tools, and provide avenues for future research to address these limitations.

5.1 BIAS

Our observations of privacy violations may be susceptible to selection bias induced by web crawling. A similar bias can also affect the quality of predictions made by models trained on such biased datasets, potentially impacting the generalizability of our findings.

The registration crawler introduces bias into the reported statistics of marketing consent violations. As detailed in Section 3.4, our crawler exhibits greater success in signing up for simple websites and forms, such as newsletters, compared to complex registration processes. However, it is possible that form complexity and website compliance are correlated, which means our results may not fully represent the entire population of websites visited by users.

To address this limitation, we propose involving real users in parts of the process, similar to the pilot study. For example, semi-automated techniques can be employed for email confirmation, ensuring that humans accurately handle the various double-opt-in processes used by websites. We employed such techniques for our subsequent crawls. Additionally, violation detection can be inspected similarly to our approach in Section 3.4.

The dataset used to train CookieBlock may also be biased for several reasons. First, we collect cookies only from websites that use the services of a CMP and assign purposes to individual cookies. Cookies used by such websites can differ from those found on generic websites, but Bouhoula et al. [26, Table 1] have evaluated differences in cookies between different CMPs and found no significant bias. Second, our web crawling process underrepresented functional cookies, leading to reduced precision for this class. Involving the registration crawler could significantly expand the representativeness of the collected dataset and potentially improve classifier performance by reducing bias. However, the overlap between websites with supported CMPs and those where our crawler can register is small, not justifying the significant engineering efforts of combining OpenWPM and a custom Chrome-based crawler. Third, the features collected during an automated crawl can differ from those resulting from user browsing patterns. To address this, we exclude features that depend on browsing patterns, such as cookie updates. Fourth, if the websites can detect our crawler as a bot, they can serve different data to the crawler than to a real user. We employ the bot evasion technique by OpenWPM, although we acknowledge they might not be sufficient. Lastly, the model should be kept up-to-date, otherwise the validity of the training data can become outdated. We address this by simplifying the process of collecting training data and model updates. We have updated the model once, with the first model used from the initial release in May 2021 until version 1.1.0 was released in August 2022 with an updated model. We plan to incorporate changes stemming from the usability study by Schöni [28] together with a new model trained from a fresh crawl into a release 2.0.

5.2 TRUSTWORTHINESS OF VIOLATIONS

All our findings are subject to potential misclassification. Therefore, it is crucial to view all violations as potential violations. In cases where our methods exhibit low precision in identifying violations, caution should be exercised when using the results for enforcement purposes. Two complementary solutions can help address this issue. First, a careful examination of violation evidence in the form of screenshots and website source code, similar to our approach in Section 3.4, is necessary for attributing violations to concrete websites. Second, constructing a larger training dataset by rectifying misclassified violations and adjusting corresponding legal labels can improve our models in the future. This is particularly crucial for properties with few positive samples, such as the pre-checked marketing checkbox.

Lastly, our methods do not serve as a complete audit as there may be additional unaddressed violations. Detecting email sharing may require a prolonged observation period to capture incriminating events.

TERRITORIAL APPLICABILITY OF EU PRIVACY LAWS. While we access websites from Germany and register a user located in the same country, note that websites with only a few EU visitors may not be obligated to comply with EU regulations. To ensure the enforcement of EU law, future studies can restrict their analysis to lists that rank websites by the origin of visitors, such as CrUX or Similarweb. As we found in Section 3.2.1.6, the registration rate is favorable when crawling such lists. By utilizing these lists and considering additional factors, such as the website's language, one can estimate whether a website is targeting users located in the EU and, consequently, whether their privacy rights must be respected.

5.3 ADVERSARIAL WEBSITES

ML models are susceptible to adversarial ML methods, which becomes even more critical when models are included in browser extensions to block content used for website monetization, as demonstrated by Tramèr et al. [126].

Within CookieBlock, we did not address the possibility of websites altering the content of their cookies specifically to counteract CookieBlock's cookie policy enforcement. For example, an adversarial website could change the cookie's name to a randomly generated value, use a proxy domain (CNAME cloaking [127]) to alter the cookie's host field, or obfuscate the cookie's content. However, CookieBlock did not reach wide-scale spread, and for websites, it is easier to employ other tracking technologies that do not involve cookies, which we do not consider in this work. Nonetheless, some websites, including those that use the CookieBot CMP, also declare other tracking resources like localStorage or tracking pixels. Therefore, it is possible to extend CookieBlock with a classification of these alternative tracking methods. We attempted to explore this direction in a thesis by Ganz [128], but this effort was unsuccessful given the rare declarations of other tracking methods than cookies. We also found that existing privacy filters and studies like WebGraph [129] address these tracking resources sufficiently.

Regarding ML-based detection of violations related to consent with marketing emails, website operators could potentially modify their forms by including input fields or text labels that are invisible to users. We assume that websites do not engage in such practices since we have not published our violation detection models, making it challenging for websites to exploit their weaknesses to evade detection. Moreover, our classification relies on both machine learning and the crawler's keyword-based form classification, making it challenging to evade our detection without access to the crawler's source code. The crawler's source code will not be published due to ethical risks. Finally, when the concerns would increase in future, we can employ methods proposed by Chen et al. [130] to enhance the robustness of decision tree models against adversarial modifications.

5.4 ENFORCEMENT

CookieBlock removes cookies after their creation rather than blocking the requests that generate them. This may not be sufficient to prevent cookies from fulfilling their intended purposes. We rarely observe cookies that are created and removed by the website more quickly than the $\sim$ 20ms required for CookieBlock to process the cookie. One example is the cookie `GoogleAdServingTest`, which records which advertisements have been displayed to the user. Fortunately, such cookies are rare. This limitation arises because it is not currently possible to prevent cookie creation within the WebExtension API. We can only remove a cookie after it has been stored in the browser.

6 RELATED WORK

6.1 PRIVACY OF REGISTRATION

ANALYSIS OF NEWSLETTERS Studies that analyze email content either rely on publicly available email datasets or require the authors to gather such datasets by signing up for services, a method similar to ours. The most closely related publications to our research are the following three studies. Englehardt et al. [6] subscribed to 902 newsletters by crawling 15 700 shopping and news websites. They analyzed how opening emails or following links within them results in information leakage. They observed that 30% of the emails leaked recipients' email addresses to third parties. They also investigated tracking protection in email servers and clients, proposing new privacy measures. In contrast, our studies focus on the legal aspects associated with sending marketing emails, particularly those from websites where the registration serves other purposes than only subscription to newsletters. Englehardt et al. exclusively subscribed to emails from websites that we have annotated as serving a marketing purpose (ma_purpose).

In a second study, Hamin [40] analyzed the content of election campaigns. Her crawler visited 4487 campaign websites and successfully subscribed to 1778 newsletters. A subsequent study by Mathur et al. [7], focusing on the 2020 US elections, found that 348 out of 2800 email campaigns shared email addresses with third parties. Importantly, only 25% of these campaigns disclosed their email-sharing practices. Both of these studies also examined email content but with a specific emphasis on manipulative tactics and political implications. As with the previous paragraph, these two studies targeted a narrow group of email senders who send emails to a) subscribed users, b) interested in elections, and c) located in the US. In contrast, our studies take a more generic approach, primarily focusing on registration forms, with newsletter subscription constituting only 10% of successful form submission in the manual pilot study and about half in the automated large-scale study. Moreover, our studies observed five times fewer email sharing than Mathur et al., with a high statistical significance. This difference could result from EU privacy regulations offering more protection to users than their US counterparts or from political campaigns being more aggressive in sending emails compared to a generic sample of websites that mainly advertise products, which form the focus of our studies.

AUTOMATED REGISTRATION Drakonakis et al. [10] automated the registration process to detect insecurely configured cookies on over half of the websites. Their crawler successfully registered on 1.6% of the Alexa top 1M websites, whereas our crawler achieved registrations on 5.9% of websites from the comparable Tranco list, although nearly half of our registrations stemmed from newsletter sign-ups. In contrast, Drakonakis et al.'s method relies on Single Sign-On (SSO), a component unsuitable for our mail violation detection, which requires a unique email address for each registration. We attempted to re-evaluate their results but were unsuccessful due to their code's dependence on an outdated Google SSO API. Besides, Zhou et al. [131] focused on registering and inspecting vulnerabilities on websites employing Facebook SSO, making their work even less aligned with our study objectives. Dimova et al. [73] examined the personal data shared from identity providers to websites using OAuth, highlighting that a substantial portion of shared data is redundant, as revoking access to it does not hinder the login process.

A generic registration crawler was also proposed by Chatzimpyrros et al. [62]. They claim that their crawler successfully registered on 26.4% of websites, encompassing 80% of websites featuring any type of registration form. However, their claims are questionable. First, they regard login forms as equivalent to registration forms. Second, they classify registration as successful immediately upon form submission. Lastly, they do not report the number of email senders, except for 0.03% of websites that sent emails without the crawler's form submission. Senol et al. [8] similarly investigated the detection of private data exfiltration prior to form submission. They observed that nearly 3% of websites extracted private inputs, such as email addresses. This can lead to potential misuse of password managers that automatically fill password fields, as reported in the Mozilla issue tracker [132, 133] and studied by Acar et al. [134] and Xin et al. [135].

Jonker et al. [136] developed a crawler that logs into websites using a legitimate crowd-sourced database of credentials called BugMeNot. They were able to log in to 14.3% of approximately 50k websites present in the BugMeNot database. However, they do not present any privacy or security results. While Jonker et al.'s approach is more effective at logging in compared to our crawler, it is constrained by the size of the BugMeNot database. Consequently, their approach is unsuitable for detecting violations during the registration process or within emails.

ANALYSIS OF CONSENT COMPLIANCE While our studies focus on consent with marketing emails, websites must obtain consent for various other processing purposes. Oh et al. [9] proposed four conditions on consent according to the GDPR. They evaluated these conditions both manually on 500 websites and by crawling 10 000 websites. They observed their crawler's decisions aligned with human judgments in 96% of cases. Their study partially overlaps with our inspection of GDPR consent violations in Section 3.3. However, our study delves deeper into the legal aspects of marketing emails, whereas their study is dealing with consent to privacy policies. Our decision procedure requires the observation of data misuse (such as receiving unsolicited marketing emails), which mandates completing a registration, which is a challenging task to automate. In contrast, their crawler detects violations solely by observing the registration form without any interaction and before the occurrence of data misuse.

Hasan Mansur et al. [137] automated the detection of dark patterns across websites and apps, including the identification of pre-checked boxes as a default choice. Their findings however underscore the difficulty in detecting this type of violation [137, Sec. 5-C]. A similar yet manual study was conducted by Gunawan et al. [138].

6.2 PRIVACY OF COOKIES

COOKIE CLASSIFICATION. In [139], Hu et al. propose a cookie purpose classifier that uses a Multinomial Naive Bayes model, which takes as input n-gram tokens extracted only from the cookie names. They train their model on 11.5k cookies with ground-truth labels taken from Cookiepedia, and state an F1-score of 94.6%. They also report a confusion matrix for one fold, which achieves a conflicting F1-score of only 86.7%.

Their work shares similarities with ours, but both works were developed simultaneously, with neither party being aware of the other. Our approach differs in two main respects. First, rather than using just the cookie name as a feature for training, we extract features from all cookie properties, including those that are observed between cookie updates. While the cookie name is simple to alter, the value and domain are restricted by the implementation requirements, e.g., a tracking cookie requires a minimum amount of entropy. This fact makes spoofing Hu et al.'s model by an adversarial web developer much easier than our model. Moreover, their model cannot distinguish cookies with the same name (e.g., user_id) but with different purposes and originating from different domains. Calzavara et al. [140, Sec. 5.2.1] showed that many cookies use naming conventions for unexpected purposes, which is not reflected by Cookiepedia's use of a single classification.

Secondly, our model is trained on ground truth collected from CMPs, while Hu et al. use ground truth labels collected from Cookiepedia. We elaborate on the advantages of our choice in Appendix A.2.1. Their classification task is also not affected by noise, which allows for a higher theoretical performance bound. This is because Cookiepedia will always report the same category for the same cookie name. By replacing the CMP labels with Cookiepedia labels on our dataset, our model accuracy increases from 87.2±0.23% to 89.2±1.3%. We provide additional details on these results in appendix A.2.1.2.

Calzavara et al. [140] used ML models to detect authentication cookies. They used a training sample of 2.5k cookies with 332 authentication cookies. They propose feature extraction from both the cookie name and other attributes, such as the entropy and the length of the cookie value, the expiry or whether the cookie is "HttpOnly." All their features or equivalent ones are included in our feature extraction. Their binary classification achieves an F1-score of 83% in classification tailored towards the high recall of 89%.

WEBSITE PRIVACY ENFORCEMENT TOOLS. There exists various proposed privacy enforcement tools by academia and industry. The Platform for Privacy Preferences (*P3P*) [23] is a framework for visualizing privacy policies on the client-side and enforcement of user preferences on the server-side. Although it was proposed as a W3C standard, it was never widely adopted, and Google and Facebook even bypassed P3P [141]. Another project, now discontinued because of lack of interest by websites, is the "Do Not Track" HTTP request header [142, 143]. Unlike these attempts to protect user privacy, the success of CookieBlock does not depend on the cooperation of the visited websites.

Major browsers are addressing user tracking by various means. Firefox introduced "Enhanced Tracking Protection" with controls such as blocking social-media tracking cookies [144] and "Total Cookie Protection" for partitioning third-party cookies per origin websites [145]. The Chromium Project proposed "Privacy Sandbox" [146] that plans to deprecate third-party cookies by 2022. These projects face similar issues as our project, in that they also need to white-list necessary third-party cookies, such as those for single sign-on. Compared to these efforts by browsers, CookieBlock also allows blocking for first-party cookies, such as Google Analytics.

The browser extensions Consent-O-Matic [113] and Cliqz Autoconsent [147] have similar goals as CookieBlock. They enforce users' cookie policies on websites by automating interaction with the CMPs. However, they are limited to websites that use supported CMPs, and they also depend on the website's honesty to follow the consent. In Section 4.2 we showed that dependence on the CMP's implementation still leads to multiple potential privacy violations. By being universally applicable, CookieBlock can provide stronger privacy guarantees.

Another privacy enhancing browser extension is Privacy Badger [148]. This extension logs third-party requests that perform fingerprinting or set cookies containing enough entropy to be used for tracking. When such tracking information is found in multiple websites' requests, Privacy Badger adds this third party to a blocklist. The construction of the blocklist used to happen individually in browsers, but this is prone to fingerprinting [149]. Hence Privacy Badger developers bundle the same blocklist constructed from an automated crawl to all users. CookieBlock focuses only on cookies, and it blocks them individually, compared to Privacy Badger which blocks the whole domain once it is detected to perform tracking. Unlike CookieBlock, Privacy Badger cannot prevent tracking using first-party cookies.

STUDIES OF COOKIE CONSENT COMPLIANCE. Researchers are continuously scrutinizing cookie consent compliance. Kampanos et al. [13] analyzed 17k websites in the UK and Greece and found that roughly 45% have a cookie banner. They also find that most of the websites nudge users into accepting all cookies. Matte et al. [14] inspected 1426 websites that use CMPs that are part of IABEurope's Transparency and Consent Framework. They find that 10% of these websites set consent before user action, and 5% of the manually inspected 560 websites store positive consent despite the user's choice to opt-out. In addition, Matte et al. develop a browser extension called "Cookie Glasses" that detects dishonest CMP implementations. Trevisan et al. [15] found that 49% of the inspected 36k websites set profiling cookies before users consent to them.

The study by Santos et al. [150] provides extensive legal background on cookie consent in EU jurisdictions. They define 22 requirements on valid cookie consent, some of which we inspected in our study.

There are several analyses of dark patterns of cookie consent notices, often supplemented with a user study. Nouwens et al. [16] found that almost 90% of an examined 680 websites using supported CMPs do not meet the GDPR requirements for valid consent. A user study by Utz et al. [17] inspected how the design of consent popups from 5k websites nudge users into uninformed consent. Since the field of the dark patterns is very active, we list further studies [18–22], and refer the reader to Dark Patterns workshop at ACM CHI.

6.2.1 *Publications succeeding our study*

In this section, we examine papers published since the completion of our CookieBlock study in September 2022. To identify such publications, we employed these three strategies.

- We considered all publications from the following conferences: S&P, NDSS, USENIX, CSS, WWW, PETS.
  - This included 343 S&P, 179 NDSS, 343 USENIX, 414 CCS, 1066 WWW, and 251 PETS publications.
- We considered publications that cited the following studies on cookie consent compliance measurements: [14–16, 24, 151].
  - These works were cited by 998 publications (with overlaps).
- We conducted a Google Scholar search using the query: "cookie consent (compliance OR violation OR measurement) AND (notice OR banner OR popup)."
  - This search yielded 994 publications.

From these sources, we selected publications based on their titles and if necessary, a quick inspection of the abstract. We applied the following filtering criteria.

- The work should be published in computer science or law conferences or journals or on preprint servers like arXiv, and it should be written in English.
- We excluded workshop works and incomplete works such as study pre-registrations.
- The work should have international impact and should not focus solely on a single small country (e.g., compliance of cookie notices in Ireland).
- We excluded usability studies of cookie notices that did not include measurements of website practices. While these studies contribute to dark patterns research, our focus is on measurement literature. For a literature overview covering dark patterns, please refer to Mathur et al. [152] or Gray et al. [153].

- We excluded work that focused on tracking methods other than cookies and on platforms other than websites (e.g., mobile, IoT).

Finally, we skimmed the works and filtered out measurement studies that did not involve the consent notice as part of the inspection (e.g., works that only measured present cookies without any connection to the notice). After this process, we identified 30 publications for further investigation.

6.2.1.1 *Attributes for evaluation*

Following a preliminary study of more than half of the selected publications, we decided to label the following attributes. The resulting Table 6.1 follows this scheme.

AUTOMATED/MANUAL. Indicates whether the primary measurements in the study were conducted using A=automated or M=manual methods.

SOURCE OF WEBSITES. Website selection largely influences the representativeness of the observations, as shown by Ruth et al. [80]. By & we denote intersection of sources, while + combines sources in an union.

POPULATION COMPARISON. Whether the study compared results depending on parameters such as location, website popularity, website category, present CMP, etc.

NOTICE DETECTION. The majority of selected works interact with consent notices, requiring their detection in the first place. Common methods (which can be combined, either to narrow down the selection (&) or to extend it (|)) include keyword-based detection, utilization of community constructed filter lists like Easylist Cookie, detection of a higher z-index, hard-coded detection of selected CMPs, etc.

NOTICE INTERACTION. Interacting with consent notices allows for the inspection of various granted consents (e.g., accept all, accept preselected, reject all, etc.). We document the methods used.

OBSERVING TRACKING. To detect non-essential data collection, studies tend to use various methods such as community constructed privacy filter lists, the presence of third-party cookies, data collected in the consent notice, or machine learning methods.

VIOLATIONS. We define categories of observed violations, dark patterns, or other privacy issues, including unconsented tracking (missing notice or implicit consent), tracking after (partially) rejected consent, interface violations/dark patterns, or incorrect information in notices (e.g., expiration, purposes).

TABLE 6.1: A literature overview according to criteria specified above.

Publication	Automated/manual	Source of websites	Population comparisons	Notice detection	Notice interaction	Observing tracking	Unconsented tracking	Tracking after rejected consent	Interface	Wrong information in notice	Summary of the study and clarifications of the documented criteria.
Alharbi et al. [154]	M	Governmental sites	✗	Manual	Manual	3rd-party cookies	✓	✓	✓	✗	Manual investigation of 243 e-gov websites.
Berens et al. [155]	M	German top 500 Alexa	✗	Manual	Manual	GA | 3rd-party tool*	✓	✓	✓	✗	Tracking decided by presence of Google Analytics and *cookieserve.com.
Bouhoula et al. [26]	A	CrUX	Rank, lists	Keywords & list & z-index & NLP	NLP	ML (CookieBlock+)	✓	✓	✓	✓	A generalization of violation and dark patterns detection.
Demir et al. [120]	A	Tranco	✗	Keywords	Keywords	Cookiepedia	✓	✓	✗	✗	A comparison of browser extensions that automate consent.
Fouad et al. [156]	A	Alexa	Category, country	✗	✗	Cookiepedia	✓	✗	✗	✗	A detection of cookie respawning.
Gotze et al. [157]	A	Governmental sites	✗	✗	✗	Lists	✓	✗	✗	✗	A detection of tracking on governmental websites.
Gunawan et al. [138]	M	Top sites & top apps	Category	Manual	Manual	✗	✗	✗	✓	✗	A comparison of dark patterns among websites and apps.
Gundelach et al. [158]	A	Tranco	✗	Keywords | List | NLP	Keywords	✗	✗	✓*	✓	✗	A comparison of different notice-detection methods. *Results only about dark patterns in notice design.
Habib et al. [159]	M	Tranco & 5 CMPs	✗	Manual	Manual	✗	✗	✗	✓	✗	A user study of 191 notices of 5 CMPs (found in 1k websites), followed by evaluation of dark patterns influencing users.
Hasan Mansur et al. [137]	A	Alexa & e-commerce	✗	ML*	✗	✗	✗	✗	✓	✗	A deep-learning detection of dark patterns in any interface, including *cookie notices.
Hils et al. [160]	A	Tranco & CMPs	CMPs	API & keywords	✗	✗	✗	✗	✗	✓	The evolution of privacy preference signals, with measurement of TCF adoption.
Jha et al. [161]	A	Similarweb + Tranco	countries	Keywords	Keywords	Lists & expiry	✓	✓	✗	✗	Comparing websites of different EU countries upon different interaction with cookie notices.
Jha et al. [162]	M	Single unnamed CMP	EU vs Latin America	1 CMP	1 CMP	✗	✗	✗	✓*	✗	A user study from 4M user interactions with a Latin American CMP. Reported behavior does not constitute violations locally, but would in the EU.
Khandelwal et al. [163]	A	Tranco	✗	z-index & NLP	NLP	✗	✗	✗	✗	✗	Proposed a similar extension to CookieBlock and show how many websites have notices with options. No violations.
Kirkman et al. [164]	A	Tranco	✗	List & z-index & keywords	Keywords	Entropy & expiry	✗	✓	✓	✓	A detection of 10 dark patterns types.
Kopmeiners [165]	A	Tranco	Countries, rank	Keywords	Keywords	✗	✗	✗	✗	✗	Only analysis of observed interactive elements. The crawler based on [161].
Krisam et al. [166]	M	German top 500 Alexa	✗	Manual	Manual	✗	✗	✗	✓	✗	A manual evaluation of options provided by consent notices.
Kyi et al. [167]	A	Tranco	✗	Keywords	Keywords	✗	✗	✗	✓	✓	The crawl detects dark patterns and "legitimate interest," which are then used in user study of perception of them.
Moti et al. [168]	A	Common Crawl & Child sites	EU vs US	List	Autoconsent	List (EasyList)	✓*	✗	✗	✗	*Using autoconsent to accept all, which is not valid parental consent, GDPR requires at least email verification.
Munir et al. [169]	A	Tranco	✗	✗	✗	ML (CookieBlock+)	✓	✗	✗	✗	An improved CookieBlock's purpose classification model by JS execution features.
Pedersen et al. [170]	A	Domcop	✗	Keywords & list	Keywords, only exploration	✗	✗	✗	✓*	✗	*An automated detection of dark patterns, without actual cookie usage. Hence no framing as violations.
Rasaii et al. [171]	A	Tranco	Location, rank	Keywords & z-index	Keywords & List (Fanboy)	Lists	✓	✓	✗	✗	A comparison of tracking between countries, including case study of CCPA and Brazilian privacy laws.
Santos et al. [172]	M	Tranco	✗	Manual	Manual	✗	✗	✗	✗	✓	An annotation of 407 notice texts.
Tang [173]	A	Tranco	✗	Keywords	Keywords, only exploration	✗	✗	✗	✓*	✓*	*An extraction consent options and information without actual cookie usage. Hence no framing as violations.
Utz et al. [87]	A	TheInternetBackup	✗	List | IAB's CMP	✗	List (WhoTracks.me)	✓	✗	✗	✗	A notification study, the use of tracking cookies without/prior consent is one of the observations.
van Eijk et al. [174]	A	Majestic	Country	List	✗	✗	✗	✗	✗	✗	Detecting cookie notices depending on location of the crawl and website.
Van Hofslot et al. [175]	A	Tranco	✗	Sample from [172]	✗	✗	✗	✗	✗	✓	An NLP classification of consent notice content. No new measurements.
Van Nortwick et al. [176]	A	Tranco & California	✗	Keywords*	Keywords*	List (EasyPrivacy)	✓	✗	✗	✗	*The investigation limited to opt-out regime in CCPA, namely the presence of Do Not Sell My Personal Information links.
Wesselkamp et al. [177]	M	FR+DE health websites	✗	Manual	Manual	Differential analysis	✓	✓	✗	✗	A manual study of 176 German and French health websites.
Zimmeck et al. [178]	A	BuiltWith & Tranco & USPAPI	✗	✗	GPC signal	Self declaration	✗	✓	✗	✗	CCPA. Relied on self declaration of compliance, which can be wrong.

6.3 OTHER

WEBSITE COMPLIANCE ANALYSIS Numerous studies have analyzed website compliance with privacy regulations that are complementary to our analysis.

Linden et al. [179] and Degeling et al. [180] analyzed how privacy policies changed with the GDPR coming into legal force. They observed an increase in the length and the number of policies and an improvement in GDPR compliance. Amos et al. [181] observed similar results in their longitudinal study of privacy policies. Liepina et al. [182] present Claudette, a scanner for GDPR violations in privacy policies. Harkous et al. [183] propose Polisis, a privacy policy scanner that summarizes policies' content. We used Polisis in our manual pilot study to analyze whether websites disclose sharing email addresses with third parties. A semantic text analysis of policies by Bui et al. [184] can further improve the automation by extracting the names of the third parties defined in the privacy policy. However, this work was published after we finished our privacy policies analysis using only Polisis and in our large-scale study, we employed a simple search for the domain name in the complete text of privacy policy and terms and conditions.

Urban et al. [114] studied loading of third-party dependencies and report that 93% of websites that embed third parties potentially violate regulations. Some et al. [185] studied how websites follow browser security policies (namely Content Security Policy and Same Origin Policy) and found that 31.1% of websites are potentially vulnerable. Englehardt summarized in his dissertation [186] the extensive work on the OpenWPM crawling platform, used mainly to detect tracking or data exfiltration from web forms. Fietkau et al. [187] observed browser fingerprinting on 19% of websites and show how common privacy-enhancing technologies fail to prevent it. All of these studies complement our research, as they focus on other privacy issues and they do not frame them as violations of the privacy regulations.

7 FINAL WORDS

This chapter concludes the thesis. We begin by summarizing the key findings from our research and discussing their implications. Subsequently, we delve into avenues for future research, considering emerging technologies and evolving privacy regulations. Finally, we discuss the challenge of enforcing privacy regulations.

7.1 SUMMARY OF FINDINGS

Investigating how websites handle sensitive user data is an active research area. With the establishment of privacy regulations, such as the General Data Protection Regulation (*GDPR*) and the ePrivacy Directive (*ePD*) studied in this thesis, measuring these invasive practices has been framed as violations of these regulations. Our work contributes to this body of empirical research by providing measurements related to the unauthorized sending of marketing emails and the unconsented collection of private data through cookies. Notably, our research goes beyond identifying violations during the consent process, as we also detect the actual misuse of data that directly harms users.

Regarding consent for sending marketing emails, we have presented both manual and automated methods for detecting potential violations. Our methods reflect legal requirements for the registration process. Our findings revealed that in the pilot study, 21.9%, and in the large-scale study, 37.2% of websites engage in practices such as sending marketing emails without proper consent in the form, sending first a marketing email immediately after the subscription, or sharing email addresses with third parties without a proper disclosure of this practice. While these figures are significant, a comparison with similar studies conducted in the United States (*US*) suggests that the prevalence of such practices in the European Union (*EU*) may be comparatively lower. Drawing definitive conclusions about the protective effect of EU privacy regulations requires a comprehensive study employing consistent methodologies across both the EU and the US. Nevertheless, such a study would only uncover correlations, leaving room for multiple causal factors, including market differences between the EU and the US.

Regarding cookies, our research focused on Consent Management Platforms (*CMPs*) that disclose individual cookie purposes in their notices. This, along with other detailed information, allowed us to identify previously unexamined privacy violations. We found that 94.7% of the websites we inspected violated at least one requirement stemming from the GDPR and ePD. These violations included the use of cookies with unconsented purposes, incorrect expiration dates, or entirely missing cookies in the consent notice. Furthermore, these websites created non-essential cookies before obtaining user consent or despite receiving negative consent. Although these violations are widespread, they are not entirely unexpected, as numerous other studies have reported a high prevalence of cookie consent violations. The higher prevalence observed in our study can be attributed to our methods' ability to access programmatically detailed information in these notices. Subsequent studies building upon our methods generalize detection to any website, thereby expanding the scope of such analysis.

We automated the detection violations in both of the studied areas. However, to extract deeper insights from our results, interdisciplinary collaboration, such as with law or economics experts, is essential. In our ongoing collaboration with legal scholars, we plan to establish correlations between observed violations and measurement attributes, such as website category, popularity, economic aspects of companies running the websites, CMPs and other technologies employed by websites, language, location, and crawl location. Additionally, we have initiated a project aimed at uncovering the reasons behind non-compliance of cookie notices. Our hypothesis for non-compliance includes neglect by website operators, either due to a lack of legal knowledge or technical expertise in configuring cookie notices. We will explore these hypotheses through a notification study in which we provide website operators with reports generated by our cookie crawler or CookieAudit, facilitating an evaluation of the respective hypotheses. Identifying other causes for non-compliance, such as malicious intent, will require an in-depth examination of the reasoning behind website operators' and CMP companies' actions.

7.2 FUTURE TECHNOLOGIES AND REGULATIONS

In an ideal world, CookieBlock would be unnecessary. Proposed technologies, such as Advanced Data Protection Control (*ADPC*) [188] and Global Privacy Control (*GPC*) [189], promise to standardize the protocol of the consent process, enabling automation through web browsers or browser extensions or unify the interface to ensure compliance. However, these ideas are not novel, with projects like the Platform for Privacy Preferences (*P3P*) and Do Not Track (*DNT*) having previously attempted similar goals, only to fail due to websites' lack of adherence. The success of these new proposals hinges on regulatory mandates, such as those established by GDPR and the California Consumer Privacy Act (*CCPA*). Similarly, CookieBlock's functionality could be mandated if the World Wide Web Consortium (*W3C*) extends cookie headers to include a "purpose" flag as a new attribute, thereby integrating cookie consent directly into web browsers. An advantage of this approach over ADPC or P3P is that our classifier could facilitate this transition by predicting the purpose for any cookie that lacks a specified purpose. This could facilitate the web's transition from the status quo to a future with transparent cookie declarations. In such a future, our classification could detect violations, such as cookies assigned the wrong purpose. Until major browser vendors take action, CookieBlock can assist users in enforcing their cookie policies on any website, regardless of their location within or outside the EU.

Although some new technologies can improve privacy, others can increase privacy risks. For instance, Google's proposal for private yet targeted advertising [190], presented as an alternative to third-party cookies, failed to achieve its privacy claims [27, 191]. Similarly, the rise of privacy-enhancing technologies has prompted privacy-invasive countermeasures, such as CNAME-cloaking. Given that most research only examines privacy compliance at a single point in time, measuring such changes remains complicated. We propose conducting evaluations over extended periods, as demonstrated by our re-evaluation of cookie violations after one year. Ideally, more frequent experiments would allow for evaluating the impact of new technologies, enforcement actions, and evolving regulations, such as the upcoming Digital Markets Act, Digital Services Act, and ePrivacy Regulation. Such long-term studies would facilitate the discovery of causal links, providing policymakers with a feedback loop enabling them to refine future regulations more effectively.

7.3 ENFORCEMENT

Our findings underscore the critical importance of enforcement. The majority of our discoveries violate the ePD as of its 2009 amendment. However, enforcement of the ePD has been infrequent, and fines have remained low until the enforcement powers of the GDPR came into play. Nevertheless, the enforcement lacks behind the strength of the regulations. This is where our automated methods can be applied. By identifying potential violations in the wild, we can reduce the time-consuming manual investigations typically carried out by regulatory authorities. Our methods can identify websites that are likely to be in violation of regulations. Moreover, our methods can generate evidence assisting overloaded and underfunded regulatory agencies in policing the Internet more efficiently and increasing compliance with legal requirements. To enhance the practicality of such applications, we must further increase the precision of our models, which can be achieved by labeling a more balanced training dataset.

Finally, when legal enforcement mechanisms are not swiftly addressing privacy-invasive practices, we propose client-side enforcement as an alternative. With our privacy extension, CookieBlock, we have demonstrated that machine learning predictions of cookie purposes can achieve accuracy levels comparable to those of experts, enabling the blocking of cookie-based tracking directly in web browsers. This browser extension is available for the majority of web browsers, reaches over 13k installations, and continues to be extended by subsequent studies. Nevertheless, not all tracking methods can be effectively countered through client-side enforcement. For instance, once a user submits their email address via a form to a server, it is impossible to limit the usage of that email address. An alternative client-side method combines our violation detection methods with a browser extension designed to raise privacy awareness. This extension could warn users before they engage with forms on websites that engage in privacy-abusive practices. Alternatively, the extension could reorder search results to penalize such websites, reducing the likelihood of users visiting them. If such behavior becomes part of Search Engine Optimization, it can serve as an additional incentive for maintaining user privacy alongside regulatory enforcement.

BIBLIOGRAPHY

1. Laperdrix, P., Bielova, N., Baudry, B. & Avoine, G. Browser fingerprinting: A survey. *ACM Transactions on the Web (TWEB)* **14**, 1 (2020).
2. Roesner, F., Kohno, T. & Wetherall, D. *Detecting and Defending Against Third-Party Tracking on the Web* in *9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)* (USENIX Association, San Jose, CA, 2012), 155.
3. Solomos, K., Ilia, P., Ioannidis, S. & Kourtellis, N. Clash of the trackers: Measuring the evolution of the online tracking ecosystem. *arXiv preprint arXiv:1907.12860* (2019).
4. Acquisti, A., Brandimarte, L. & Loewenstein, G. Privacy and human behavior in the age of information. *Science* **347**. Publisher: American Association for the Advancement of Science, 509 (2015).
5. Boldyreva, E. *Cambridge Analytica: Ethics And Online Manipulation With Decision-Making Process* in (2018), 91.
6. Englehardt, S., Han, J. & Narayanan, A. I never signed up for this! Privacy implications of email tracking. *Proceedings on Privacy Enhancing Technologies* **2018**, 109 (2018).
7. Mathur, A., Wang, A., Schwemmer, C., Hamin, M., Stewart, B. M. & Narayanan, A. *Manipulative tactics are the norm in political emails: Evidence from 100K emails from the 2020 U.S. election cycle* `https://electionemails2020.org`. 2020.
8. Senol, A., Acar, G., Humbert, M. & Borgesius, F. Z. *Leaky Forms: A Study of Email and Password Exfiltration Before Form Submission* in *31st USENIX Security Symposium (USENIX Security 22)* (2022), 1813.
9. Oh, J., Hong, J., Lee, C., Lee, J. J., Woo, S. S. & Lee, K. Will EU's GDPR Act as an Effective Enforcer to Gain Consent? *IEEE Access* (2021).

10. Drakonakis, K., Ioannidis, S. & Polakis, J. *The Cookie Hunter: Automated Black-box Auditing for Web Authentication and Authorization Flaws* in *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security* (2020), 1953.

11. Kubicek, K., Merane, J., Cotrini, C., Stremitzer, A., Bechtold, S. & Basin, D. Checking Websites' GDPR Consent Compliance for Marketing Emails. *Proceedings on Privacy Enhancing Technologies* **2022**, 282 (2022).

12. Burgess, M. *We need to fix GDPR's biggest failure: broken cookie notices* `https://www.wired.co.uk/article/gdpr-cookie-consent-eprivacy`. 2020.

13. Kampanos, G. & Shahandashti, S. F. *Accept All: The Landscape of Cookie Banners in Greece and the UK* in *IFIP International Conference on ICT Systems Security and Privacy Protection* (2021), to appear.

14. Matte, C., Bielova, N. & Santos, C. *Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework* in *2020 IEEE Symposium on Security and Privacy (SP)* (2020), 791.

15. Trevisan, M., Traverso, S., Bassi, E. & Mellia, M. 4 Years of EU Cookie Law: Results and Lessons Learned. *Proceedings on Privacy Enhancing Technologies* **2019**, 126 (2019).

16. Nouwens, M., Liccardi, I., Veale, M., Karger, D. & Kagal, L. Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence. *CoRR* **abs/2001.02479** (2020).

17. Utz, C., Degeling, M., Fahl, S., Schaub, F. & Holz, T. (Un)informed Consent: Studying GDPR Consent Notices in the Field. *CoRR* **abs/1909.02638** (2019).

18. Sanchez-Rola, I., Dell'Amico, M., Kotzias, P., Balzarotti, D., Bilge, L., Vervier, P.-A. & Santos, I. *"Can I Opt Out Yet?": GDPR and the Global Illusion of Cookie Control* in *Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security* (Association for Computing Machinery, Auckland, New Zealand, 2019), 340.

19. Hausner, P. & Gertz, M. Dark Patterns in the Interaction with Cookie Banners. *arXiv preprint arXiv:2103.14956* (2021).

20. Bösch, C., Erb, B., Kargl, F., Kopp, H. & Pfattheicher, S. Tales from the Dark Side: Privacy Dark Strategies and Privacy Dark Patterns. *Proceedings on Privacy Enhancing Technologies* **2016**, 237 (2016).

21. Grassl, P., Schraffenberger, H., Borgesius, F. Z. & Buijzen, M. Dark and bright patterns in cookie consent requests. *PsyArXiv* (2020).

22. Soe, T. H., Nordberg, O. E., Guribye, F. & Slavkovik, M. *Circumvention by design-dark patterns in cookie consent for online news outlets* in *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (2020), 1.

23. Cranor, L. F. P3P: Making Privacy Policies More Useful. *IEEE Security and Privacy* **1**. `https://www.w3.org/TR/P3P11/`, 50 (2003).

24. Bollinger, D., Kubicek, K., Cotrini, C. & Basin, D. *Automating Cookie Consent and GDPR Violation Detection* in *31st USENIX Security Symposium (USENIX Security 22)* (USENIX Association, Boston, MA, 2022), 2893.

25. Bollinger, D. *Analyzing Cookies Compliance with the GDPR* MA thesis (ETH Zurich, 2021).

26. Bouhoula, A., Kubicek, K., Zac, A., Cotrini, C. & Basin, D. *Automated, Large-Scale Analysis of Cookie Notice Compliance* in *33st USENIX Security Symposium (USENIX Security 24)* (USENIX Association, Philadelphia, PA, 2024).

27. Turati, F., Kubicek, K., Cotrini, C. & Basin, D. Locality-Sensitive Hashing Does Not Guarantee Privacy! Attacks on Google's FLoC and the MinHash Hierarchy System. *Proceedings on Privacy Enhancing Technologies* **2023**, 117 (2023).

28. Schöni, L., Kubicek, K. & Zimmermann, V. Block Cookies, Not Websites: Analysing Mental Models and Usability of the Privacy-Preserving Browser Extension CookieBlock. *Proceedings on Privacy Enhancing Technologies* **2024** (2024).

29. European Parliament, Council of the European Union. *Directive 2002/58/EC of the European Parliament and of the Council of 12 July 2002 concerning the processing of personal data and the protection of privacy in the electronic communications sector (Directive on privacy and electronic communications)* 2002.

30. European Parliament, Council of the European Union. *Regulation (EU) 2016/679 Of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*

http://data.europa.eu/eli/reg/2016/679/2016-05-04; Last accessed on: 2023-10-04. 2016.

31. European Data Protection Board. *Opinion 5/2019 on the interplay between the ePrivacy Directive and the GDPR, in particular regarding the competence, tasks and powers of data protection authorities* 2019.

32. Judgement of the Court (Grand Chamber). *Case C-673/17 Planet49 GmbH v Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband e.V. ECLI:EU:C:2019:246* http://curia.europa.eu/juris/document/document.jsf?docid=218462&doclang=EN; Last accessed on: 2023-10-04. 2019.

33. European Data Protection Board. *Facebook and Instagram decisions: "Important impact on use of personal data for behavioural advertising"* https://edpb.europa.eu/news/news/2023/facebook-and-instagram-decisions-important-impact-use-personal-data-behavioural_en; Last accessed on: 2023-10-04. 2023.

34. European Data Protection Board. *Guidelines 2/2019 on the processing of personal data under Article 6(1)(b) GDPR in the context of the provision of online services to data subjects* https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-22019-processing-personal-data-under-article-61b_en; Last accessed on: 2023-10-04. 2019.

35. European Data Protection Board. *Report of the work undertaken by the Cookie Banner Taskforce* https://edpb.europa.eu/our-work-tools/our-documents/report/report-work-undertaken-cookie-banner-taskforce_en; Last accessed on: 2023-10-04. 2023.

36. Rechtbank Amsterdam (Netherlands). *Case C-621/22* 2022.

37. Bakos, Y., Marotta-Wurgler, F. & Trossen, D. R. Does anyone read the fine print? Consumer attention to standard-form contracts. *The Journal of Legal Studies* **43**, 1 (2014).

38. European Data Protection Board. *Guidelines 05/2020 on consent under Regulation 2016/679* https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-052020-consent-under-regulation-2016679_en; Last accessed on: 2023-10-04. 2020.

39. *Marketing email tracker 2019* https://dma.org.uk/uploads/misc/marketers-email-tracker-2019.pdf.

40. Hamin, M. *"don't ignore this:" Automating the Collection and Analysis of Campaign Emails* tech. rep. (Princeton University, 2018).

41. Mederle, D. *The regulation of spam and unsolicited commercial emails (Die Regulierung von Spam und unerbetenen kommerziellen E-Mails)* (Heymanns, 2010).

42. Emmerich, V. & Lange, K. W. *Unfair competition (Unlauterer Wettbewerb)* (ISBN 978-3-406-72639-2, C.H. Beck, 2019).

43. Edwards, L. *The New Legal Framework for E-Commerce in Europe* (Hart Publishing, 2005).

44. Deutsche Bundestag. *German Act against Unfair Competition (Gesetz gegen den unlauteren Wettbewerb) in the version published on 3 March 2010 (Federal Law Gazette I p. 254), as last amended by Article 1 of the Act of 10 August 2021 (Federal Law Gazette I, p. 3504)* 2021.

45. Mankowski, P. *Legal commentary on the German Act against Unfair Competition (Kommentar zum Gesetz gegen den unlauteren Wettbewerb (UWG)), § 7 UWG Unacceptable nuisance (Unzumutbare Belästigungen), Par. 238, in K. Fezer, W. Büscher and E. Obergfell. Unfair competition law (Lauterkeitsrecht)* 2016.

46. Weiser, J. The possibility of using a partnership exchange can be "selling a service" in the sense of the UWG (Nutzungsmöglichkeit einer Partnerschaftsbörse kann "Verkauf einer Dienstleistung" im Sinne des UWG sein). *GRUR-Prax, (Gewerblicher Rechtsschutz und Urheberrecht, Praxis im Immaterialgüter- und Wettbewerbsrecht)* **2018**, 291 (2018).

47. Judgement of the Higher Regional Court of Munich (OLG München) from February 15, 2018. *29 U 2799/17* 2018.

48. Micklitz, H. & Schirmbacher, M. *Legal commentary on the German Act against Unfair Competition (Kommentar zum Gesetz gegen den unlauteren Wettbewerb (UWG)), § 7 UWG Unacceptable nuisance (Unzumutbare Belästigungen), Par. 203 in G. Spindler and F. Schuster, Electronic Media Law, 4th edition 2019, (Recht der elektronischen Medien, 4. Aufl. 2019)* 2019.

49. European Parliament, Council of the European Union. *Directive 2000/31/EC of the European Parliament and of the Council of 8 June 2000 on certain legal aspects of information society services, in particular electronic commerce, in the Internal Market ('Directive on electronic commerce')* 2000.

50. Deutsche Bundestag. *German Telemedia Act (Telemediengesetz) in the version published on 26 February 2007 (Federal Law Gazette I p. 179, 251), as last amended by Article 3 of the Act of 12 August 2021 (Federal Law Gazette I, p. 3544)* 2021.

51. European Parliament, Council of the European Union. *Directive 2005/29/EC of the European Parliament and of the Council of 11 May 2005 concerning unfair business-to-consumer commercial practices in the Internal Market and amending Council Directive 84/450/EEC, Directives 97/7/EC, 98/27/EC and 2002/65/EC of the European Parliament and of the Council and Regulation (EC) No 2006/2004 of the European Parliament and of the Council ('Unfair Commercial Practices Directive')* 2005.

52. European Commission. *Guidance on the implementation/application of Directive 2005/29/EC on Unfair Commercial Practices* 2016.

53. Directorate-General for the Information Society and Media (European Commission). *ePrivacy Directive, assessment of transposition, effectiveness and compatibility with the proposed data protection regulation* doi:10.2759/419180. 2015.

54. European Parliament, Council of the European Union. *Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data* 1995.

55. Judgement of the Federal Court of Justice (BHG) from May 28, 2020. *I ZR 7/16* 2020.

56. Judgement of the Court of Justice of the European Union from October 1, 2019. *C-673/17, EU:C:2019:801* 2019.

57. European Data Protection Board. *Guidelines 05/2020 on consent under Regulation 2016/679 (GDPR)* 2020.

58. Jahnel, D. *Legal commentary on the General Data Protection Regulation (GDPR) (Kommentar zur Datenschutz-Grundverordnung (DS-GVO)), Art. 7 Conditions for consent (Bedingungen für die Einwilligung)* (ISBN 978-3-709-70178-2, Jan Sramek Verlag, 2021).

59. Judgement of the Federal Court of Justice (BHG) from July 16, 2008. *VIII ZR 348/06* 2008.

60. Judgement of the Federal Court of Justice (BHG) from February 1, 2018. *III ZR 196/17* 2018.

61. Judgement of the Court of Justice of the European Union from November 11, 2020. *C-61/19, EU:C:2020:901* 2020.

62. Chatzimpyrros, M., Solomos, K. & Ioannidis, S. in *Computer Security* 91 (Springer, 2019).

63. Kast, P. *Automating website registration for GDPR compliance analysis, Bachelor's thesis, ETH Zurich* Bachelor's Thesis. 2021.

64. Epstein, L. & Martin, A. D. *An introduction to empirical legal research* (Oxford University Press, 2014).

65. Cohen, J. A coefficient of agreement for nominal scales. *Educational and psychological measurement* **20**, 37 (1960).

66. Zimmeck, S., Story, P., Smullen, D., Ravichander, A., Wang, Z., Reidenberg, J., Russell, N. C. & Sadeh, N. MAPS: Scaling privacy compliance analysis to a million apps. *Proceedings on Privacy Enhancing Technologies* **2019**, 66 (2019).

67. Kumar, V. B., Iyengar, R., Nisal, N., Feng, Y., Habib, H., Story, P., Cherivirala, S., Hagan, M., Cranor, L., Wilson, S., *et al. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text* in *Proceedings of The Web Conference 2020* (2020), 1943.

68. Wilson, S., Schaub, F., Dara, A. A., Liu, F., Cherivirala, S., Leon, P. G., Andersen, M. S., Zimmeck, S., Sathyendra, K. M., Russell, N. C., *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)* (2016), 1330.

69. European Parliament, Council of the European Union. *Directive 2006/114/EC of the European Parliament and of the Council of 12 December 2006 concerning misleading and comparative advertising* 2006.

70. Judgement of the Federal Court of Justice (BHG) from July 10, 2018. *VI ZR 225/17* 2018.

71. Micklitz, H. & Schirmbacher, M. *Legal commentary on the German Telemedia Act (Kommentar zum Telemediengesetz (TMG)), § 4-6 TMG, in G. Spindler and F. Schuster, Electronic Media Law, 4th edition 2019, (Recht der elektronischen Medien, 4. Aufl. 2019)* 2019.

72. Zscherpe, K. A. Direct marketing by e-mail – How can companies proceed legally? (Direktmarketing per E-Mail – Wie können Unternehmen rechtlich einwandfrei vorgehen?) *Journal of Business and Consumer Law, (Zeitschrift für Wirtschafts- und Verbraucherrecht)* **2008**, 327 (2008).

73. Dimova, Y., Van Goethem, T. & Joosen, W. Everybody's Looking for SSOmething: A large-scale evaluation on the privacy of OAuth authentication on the web. *Proceedings on Privacy Enhancing Technologies* **4**, 452 (2023).

74. Englehardt, S. & Narayanan, A. *Online Tracking: A 1-Million-Site Measurement and Analysis* in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery, Vienna, Austria, 2016), 1388.

75. Libert, T. *Exposing the hidden web: An analysis of third-party HTTP requests on 1 million websites* 2015.

76. Le Pochat, V., Van Goethem, T. & Joosen, W. *Evaluating the Long-Term Effects of Parameters on the Characteristics of the Tranco Top Sites Ranking* in *Proceedings of the 12th USENIX Conference on Cyber Security Experimentation and Test* (USENIX Association, Santa Clara, CA, USA, 2019), 10.

77. Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. *spaCy: Industrial-strength Natural Language Processing in Python* version v3.2.6. 2023.

78. Juršic, M., Mozetic, I., Erjavec, T. & Lavrac, N. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. *Journal of Universal Computer Science* **16**, 1190 (2010).

79. Le Pochat, V., Van Goethem, T., Tajalizadehkhoob, S., Korczyński, M. & Joosen, W. *Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation* in *Proceedings of the 26th Annual Network and Distributed System Security Symposium* (2019).

80. Ruth, K., Kumar, D., Wang, B., Valenta, L. & Durumeric, Z. *Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists* in *Proceedings of the 22nd ACM Internet Measurement Conference* (Association for Computing Machinery, Nice, France, 2022), 374.

81. Akuma, S., Lubem, T. & Adom, I. T. Comparing Bag of Words and TF-IDF with different models for hate speech detection from live tweets. *International Journal of Information Technology*, 1 (2022).

82. Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B. & Kurzweil, R. *Universal Sentence Encoder* 2018.

83. Chen, T. & Guestrin, C. *XGBoost: A scalable tree boosting system* in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (2016), 785.

84. Arik, S. Ö. & Pfister, T. *Tabnet: Attentive interpretable tabular learning* in *Proceedings of the AAAI Conference on Artificial Intelligence* **35** (2021), 6679.

85. Google Transparency Report Authors. *HTTPS encryption on the web* `https://transparencyreport.google.com/https/overview`. 2023.

86. Felt, A. P., Barnes, R., King, A., Palmer, C., Bentzel, C. & Tabriz, P. *Measuring HTTPS adoption on the web* in *26th USENIX security symposium (USENIX security 17)* (2017), 1323.

87. Utz, C., Michels, M., Degeling, M., Marnau, N. & Stock, B. Comparing large-scale privacy and security notifications. *Proceedings on Privacy Enhancing Technologies* (2023).

88. Baden-Württemberg Data Protection Authority (LfDI Baden-Württemberg). *LfDI - O 1018/115* `https://gdprhub.eu/index.php?title=LfDI_-_O_1018/115`. 2018.

89. Gelernter, N., Kalma, S., Magnezi, B. & Porcilan, H. *The password reset MitM attack* in *2017 IEEE Symposium on Security and Privacy (SP)* (2017), 251.

90. Al Maqbali, F. & Mitchell, C. J. *Web Password Recovery: A Necessary Evil?* in *Proceedings of the Future Technologies Conference* (2018), 324.

91. Routh, C., DeCrescenzo, B. & Roy, S. *Attacks and vulnerability analysis of e-mail as a password reset point* in *2018 Fourth International Conference on Mobile and Secure Services (MobiSecServ)* (2018), 1.

92. Legal team of the Certified Senders Alliance. *DOI: if not now, then when?!* `https://certified-senders.org/blog/doi-if-not-now-then-when/`. 2017.

93. Austrian Data Protection Authority (Datenschutzbehörde). *DSB-D130.073/0008-DSB/2019* `https://gdprhub.eu/index.php?title=DSB_-_DSB-D130.073/0008-DSB/2019`. 2019.

94. Schneider, M., Shulman, H., Sidis, A., Sidis, R. & Waidner, M. *Diving into Email Bomb Attack* in *2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)* (2020), 286.

95. Judgement of the Federal Court of Justice (BHG) from March 14, 2017. *VI ZR 721/15* 2017.

96. Art. 29 Data Protection Working Party. *Opinion 5/2004 on unsolicited communications for marketing purposes under Article 13 of Directive 2002/58/EC* 2004.

97. Gluck, J., Schaub, F., Friedman, A., Habib, H., Sadeh, N., Cranor, L. F. & Agarwal, Y. *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 2016)* (2016), 321.

98. McDonald, A. M. & Cranor, L. F. The cost of reading privacy policies. *ISJLP* **4**, 543 (2008).

99. Laura-Vanessa, L.-V. *Quantifying Mechanisms behind Cookie Consent (Non-)Compliance: A Notification Study of Audit Tools* Bachelor's thesis (ETH Zurich, 2023).

100. Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F. & Preneel, B. *FPDetective: dusting the web for fingerprinters* in *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security* (2013), 1129.

101. Woods, D. W. & Böhme, R. *The Commodification of Consent* in *20th Annual Workshop on the Economics of Information Security, WEIS* (2020), 25.

102. Hils, M., Woods, D. W. & Böhme, R. *Measuring the Emergence of Consent Management on the Web* in *Proceedings of the ACM Internet Measurement Conference* (Association for Computing Machinery, Virtual Event, USA, 2020), 317.

103. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *CoRR* **abs/1603.02754** (2016).

104. CookiePro. *Legacy Article – Categorizing Cookies* `https://web.archive.org/web/20210208155826/https://community.cookiepro.com/s/article/UUID-6f01b88c-0440-0642-3610-819c6ca0f7c4`. Accessed on: 2023-10-04. 2021.

105. Support, C. *Unclassified cookies - how do I classify them manually?* `https://web.archive.org/web/20201111204915/https://support.cookiebot.com/hc/en-us/articles/360003735214-Unclassified-cookies-how-do-I-classify-them-manually-`. Accessed on: 2023-10-04. 2018.

106. Kulyk, O., Hilt, A., Gerber, N. & Volkamer, M. *"This Website Uses Cookies": Users' Perceptions and Reactions to the Cookie Disclaimer* in *European Workshop on Usable Security (EuroUSEC)* (2018).

107. BuiltWith.com. *Privacy Compliance Usage Distribution in the Top 1 Million Sites* `https://web.archive.org/web/20201021075918/https://trends.builtwith.com/widgets/privacy-compliance/`. 2020.

108. Europe, I. *IAB Europe Transparency and Consent Framework Policies* `https://web.archive.org/web/20210520213158/https://iabeurope.eu/iab-europe-transparency-consent-framework-policies/`. 2021.

109. International Chamber of Commerce UK. *ICC UK Cookie guide* `https://www.cookielaw.org/wp-content/uploads/2019/12/icc_uk_cookiesguide_revnov.pdf`; Accessed on 2023-10-04. 2012.

110. Dabrowski, A., Merzdovnik, G., Ullrich, J., Sendera, G. & Weippl, E. *Measuring Cookies and Web Privacy in a Post-GDPR World: Methods and Protocols* in *International Conference on Passive and Active Network Measurement* (2019), 258.

111. Eijk, R. V., Asghari, H., Winter, P. & Narayanan, A. *The Impact of User Location on Cookie Notices (Inside and Outside of the European Union)* in *Workshop on Technology and Consumer Protection (ConPro'19). IEEE* (2019).

112. Mozilla. *OpenWPM – A web privacy measurement framework* `https://github.com/mozilla/OpenWPM`. Version used: 0.12.0. 2020.

113. Janus Bager Kristensen, R. B. *Consent-O-Matic* `https://github.com/cavi-au/Consent-O-Matic`. 2020.

114. Urban, T., Degeling, M., Holz, T. & Pohlmann, N. *Beyond the Front Page: Measuring Third Party Dynamics in the Field* in *Proceedings of The Web Conference 2020* (Association for Computing Machinery, New York, NY, USA, 2020), 1275.

115. Papadopoulos, P., Kourtellis, N. & Markatos, E. P. Cookie Synchronization: Everything You Always Wanted to Know But Were Afraid to Ask. *CoRR* **abs/1805.10505** (2018).

116. OneTrust. *Cookiepedia* `https://cookiepedia.co.uk/`. Accessed on 2023-10-04.

117. Friedman, J., Hastie, T., Tibshirani, R., *et al. The elements of statistical learning* **10** (Springer series in statistics New York, 2001).

118. Wyner, A. J., Olson, M., Bleich, J. & Mease, D. Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research* **18**, 1558 (2017).

119. Schapire, R. E. & Freund, Y. Boosting: Foundations and algorithms. *Kybernetes* (2013).

120. Demir, N. & Urban, T. A Large-Scale Study of Cookie Banner Interaction Tools and Their Impact on Users' Privacy. *Proceedings on Privacy Enhancing Technologies* (2024).

121. Acquisti, A. & Grossklags, J. Privacy and rationality in individual decision making. *Security & Privacy, IEEE* **3**, 26 (2005).

122. Jensen, C. & Potts, C. *Privacy policies as decision-making tools: An evaluation of online privacy notices* in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (2004), 471.

123. Brooke, J. SUS - A quick and dirty usability scale. *Usability evaluation in industry* **189**, 189 (1996).

124. Corner, M., Dogan, H., Mylonas, A. & Djabri, F. in *Design, User Experience, and Usability. Practice and Case Studies* Series Title: Lecture Notes in Computer Science, 442 (Springer International Publishing, Cham, 2019).

125. Kubicek, K., Bollinger, D., Zanga, A., Cotrini, C. & Basin, D. *CookieBlock & CookieAudit: Fixing Cookie Consent with ML* in (USENIX Association, Boston, MA, 2022).

126. Tramèr, F., Dupré, P., Rusak, G., Pellegrino, G. & Boneh, D. *AdVersarial: Perceptual Ad Blocking Meets Adversarial Machine Learning* in *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (Association for Computing Machinery, London, United Kingdom, 2019), 2005.

127. Dimova, Y., Acar, G., Olejnik, L., Joosen, W. & Van Goethem, T. The CNAME of the Game: Large-scale Analysis of DNS-based Tracking Evasion. *Proceedings on Privacy Enhancing Technologies* **3**, 394 (2021).

128. Ganz, R. *Understanding GDPR compliance of tracking pixel declarations using privacy filter lists* Bachelor's thesis (ETH Zurich, 2022).

129. Siby, S., Iqbal, U., Englehardt, S., Shafiq, Z. & Troncoso, C. *WebGraph: Capturing advertising and tracking information flows for robust blocking* in *31st USENIX Security Symposium (USENIX Security 22)* (2022), 2875.

130. Chen, H., Zhang, H., Boning, D. & Hsieh, C.-J. *Robust decision trees against adversarial examples* in *International Conference on Machine Learning* (2019), 1122.

131. Zhou, Y. & Evans, D. *SSOScan: Automated testing of web applications for Single Sign-On vulnerabilities* in *23rd USENIX Security Symposium (USENIX Security 14)* (2014), 495.

132. András, B. *Bugzilla: Stealing Firefox saved passwords* `https://bugzilla.mozilla.org/show_bug.cgi?id=1107422`. 2014.

133. Georgi. *Bugzilla: password manager + XSS = disaster* `https://bugzilla.mozilla.org/show_bug.cgi?id=408531`. 2007.

134. Acar, G., Englehardt, S. & Narayanan, A. No boundaries: data exfiltration by third parties embedded on web pages. *Proceedings on Privacy Enhancing Technologies* **2020**, 220 (2020).

135. Xin, R., Lin, S. & Yang, X. *Quantifying User Password Exposure to Third-Party CDNs* in *International Conference on Passive and Active Network Measurement* (2023), 652.

136. Jonker, H., Karsch, S., Krumnow, B. & Sleegers, M. Shepherd: a Generic Approach to Automating Website Login. *Workshop on Measurements, Attacks, and Defenses for the Web (MADWeb) 2020* (2020).

137. Hasan Mansur, S. M., Salma, S., Awofisayo, D. & Moran, K. *AidUI: Toward Automated Recognition of Dark Patterns in User Interfaces* in *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)* (IEEE, Melbourne, Australia, 2023), 1958.

138. Gunawan, J., Pradeep, A., Choffnes, D., Hartzog, W. & Wilson, C. A Comparative Study of Dark Patterns Across Web and Mobile Modalities. *Proceedings of the ACM on Human-Computer Interaction* **5**, 377:1 (2021).

139. Hu, X., Sastry, N. & Mondal, M. *CCCC: Corralling Cookies into Categories with CookieMonster* in *13th ACM Web Science Conference 2021* (Association for Computing Machinery, Virtual Event, United Kingdom, 2021), 234.

140. Calzavara, S., Tolomei, G., Casini, A., Bugliesi, M. & Orlando, S. A Supervised Learning Approach to Protect Client Authentication on the Web. *ACM Trans. Web* **9** (2015).

141. Brodkin, J. *Google tricks Internet Explorer into accepting tracking cookies, Microsoft claims* `https://arstechnica.com/tech-policy/2012/02/google-tricks-internet-explorer-into-accepting-tracking-cookies-microsoft-claims/`; Accessed on 2023-10-04. 2012.

142. Singer, D. & Fielding, R. *Tracking Preference Expression (DNT) W3C Working Group Note* `https://www.w3.org/TR/tracking-dnt/`. 2019.

143. Hegaret, P. L. *"WG closed – w3c/dnt@5d85d6c"* `https://github.com/w3c/dnt/commit/5d85d6c3`; Accessed on 2023-10-04.

144. Mozilla. *Enhanced Tracking Protection in Firefox for desktop* `https://support.mozilla.org/en-US/kb/enhanced-tracking-protection-firefox-desktop`. 2021.

145. Mozilla. *Firefox 86 Introduces Total Cookie Protection* `https://blog.mozilla.org/security/2021/02/23/total-cookie-protection/`. 2021.

146. Developers, C. *Chromium Projects – The Privacy Sandbox* `https://www.chromium.org/Home/chromium-privacy/privacy-sandbox`. 2021.

147. Macbeth, S. *Cliqz Autoconsent* `https://github.com/cliqz-oss/autoconsent`. 2020.

148. Electronic Frontier Foundation. *Privacy Badger* `https://privacybadger.org/`. 2019.

149. Arrieta, A., Cyphers, B., Miagkov, A. & Barnett, D. *Privacy Badger Is Changing to Protect You Better* `https://www.eff.org/deeplinks/2020/10/privacy-badger-changing-protect-you-better`; Accessed on 2023-10-04.

150. Santos, C., Bielova, N. & Matte, C. Are cookie banners indeed compliant with the law? Deciphering EU legal requirements on consent and technical means to verify compliance of cookie banners. *CoRR* **abs/1912.07144** (2019).

151. Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F. & Holz, T. We Value Your Privacy... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy. *CoRR* **abs/1808.05096** (2018).

152. Mathur, A., Kshirsagar, M. & Mayer, J. *What Makes a Dark Pattern... Dark?: Design Attributes, Normative Considerations, and Measurement Methods* in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (ACM, Yokohama Japan, 2021), 1.

153. Gray, C. M., Sanchez Chamorro, L., Obi, I. & Duane, J.-N. *Mapping the Landscape of Dark Patterns Scholarship: A Systematic Literature Review* in *Companion Publication of the 2023 ACM Designing Interactive Systems Conference* (2023), 188.

154. Alharbi, J. A., Albesher, A. S. & Wahsheh, H. A. An Empirical Analysis of E-Governments' Cookie Interfaces in 50 Countries. *Sustainability* **15**. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute, 1231 (2023).

155. Berens, B. M., Bohlender, M., Dietmann, H., Krisam, C., Kulyk, O. & Volkamer, M. Cookie Disclaimers: Dark Patterns and Lack of Transparency. *Computers & Security*, 103507 (2023).

156. Fouad, I., Santos, C., Legout, A. & Bielova, N. *My Cookie is a phoenix: detection, measurement, and lawfulness of cookie respawning with browser fingerprinting* in *PETS 2022-22nd Privacy Enhancing Technologies Symposium* (2022).

157. Gotze, M., Matic, S., Iordanou, C., Smaragdakis, G. & Laoutaris, N. *Measuring Web Cookies in Governmental Websites* in *14th ACM Web Science Conference 2022* (ACM, Barcelona Spain, 2022), 44.

158. Gundelach, R. & Herrmann, D. *Cookiescanner: An Automated Tool for Detecting and Evaluating GDPR Consent Notices on Websites* in *Proceedings of the 18th International Conference on Availability, Reliability and Security* arXiv:2309.06196 [cs] (2023), 1.

159. Habib, H., Li, M., Young, E. & Cranor, L. *"Okay, whatever": An Evaluation of Cookie Consent Interfaces* in *CHI Conference on Human Factors in Computing Systems* (ACM, New Orleans LA USA, 2022), 1.

160. Hils, M., Woods, D. W. & Böhme, R. Privacy Preference Signals: Past, Present and Future. *Proceedings on Privacy Enhancing Technologies* **2021**. arXiv:2106.02283 [cs], 249 (2021).

161. Jha, N., Trevisan, M., Vassio, L. & Mellia, M. The Internet with Privacy Policies: Measuring The Web Upon Consent. *ACM Trans. Web* **16** (2022).

162. Jha, N., Trevisan, M., Mellia, M., Irarrazaval, R. & Fernandez, D. *I Refuse if You Let Me: Studying User Behavior with Privacy Banners at Scale* in *2023 7th Network Traffic Measurement and Analysis Conference (TMA)* (IEEE, Naples, Italy, 2023), 1.

163. Khandelwal, R., Nayak, A., Harkous, H. & Fawaz, K. *Automated Cookie Notice Analysis and Enforcement* in *32nd USENIX Security Symposium (USENIX Security 23)* (2023), 1109.

164. Kirkman, D., Vaniea, K. & Woods, D. W. *DarkDialogs: Automated detection of 10 dark patterns on cookie dialogs* in *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)* (IEEE, Delft, Netherlands, 2023), 847.

165. Kopmeiners, G. The landscape of (un)balanced choices in cookie consent dialogues in Europe (2023).

166. Krisam, C., Dietmann, H., Volkamer, M. & Kulyk, O. *Dark Patterns in the Wild: Review of Cookie Disclaimer Designs on Top 500 German Websites* in *Proceedings of the 2021 European Symposium on Usable Security* (ACM, Karlsruhe Germany, 2021), 1.

167. Kyi, L., Ammanaghatta Shivakumar, S., Santos, C. T., Roesner, F., Zufall, F. & Biega, A. J. *Investigating Deceptive Design in GDPR's Legitimate Interest* in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York, NY, USA, 2023), 1.

168. Moti, Z., Senol, A., Bostani, H., Borgesius, F. Z., Moonsamy, V., Mathur, A. & Acar, G. *Targeted and Troublesome: Tracking and Advertising on Children's Websites* arXiv:2308.04887 [cs]. 2023.

169. Munir, S., Siby, S., Iqbal, U., Englehardt, S., Shafiq, Z. & Troncoso, C. CookieGraph: Understanding and Detecting First-Party Tracking Cookies. *ACM Conference on Computer and Communications Security (CCS)* (2023).

170. Pedersen, M., Guribye, F. & Slavkovik, M. Automatic detection of manipulative Consent Management Platforms and the journey into the patterns of darkness (2023).

171. Rasaii, A., Singh, S., Gosain, D. & Gasser, O. *Exploring the Cookieverse: A Multi-Perspective Analysis of Web Cookies* in *International Conference on Passive and Active Network Measurement* (2023), 623.

172. Santos, C., Rossi, A., Sanchez Chamorro, L., Bongard-Blanchy, K. & Abu-Salma, R. *Cookie Banners, What's the Purpose?: Analyzing Cookie Banner Text Through a Legal Lens* in *Proceedings of the 20th Workshop on Workshop on Privacy in the Electronic Society* (ACM, Virtual Event, Republic of Korea, 2021), 187.

173. Tang, B. J. Scraping Cookie and Activity Declarations from Privacy Policies.

174. Van Eijk, R., Asghari, H., Winter, P. & Narayanan, A. *The Impact of User Location on Cookie Notices (Inside and Outside of the European Union)* arXiv:2110.09832 [cs]. 2021.

175. Van Hofslot, M., Akdag Salah, A., Gatt, A. & Santos, C. *Automatic Classification of Legal Violations in Cookie Banner Texts* in *Proceedings of the Natural Legal Language Processing Workshop 2022* (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid), 2022), 287.

176. Van Nortwick, M. & Wilson, C. Setting the Bar Low: Are Websites Complying With the Minimum Requirements of the CCPA? *Proc. Priv. Enhancing Technol.* **2022**, 608 (2022).

177. Wesselkamp, V., Fouad, I., Santos, C., Boussad, Y., Bielova, N. & Legout, A. In-depth technical and legal analysis of Web tracking on health related websites with Ernie extension (2021).

178. Zimmeck, S., Wang, O., Alicki, K., Wang, J. & Eng, S. Usability and Enforceability of Global Privacy Control. *Proceedings on Privacy Enhancing Technologies* **2023**, 265 (2023).

179. Linden, T., Khandelwal, R., Harkous, H. & Fawaz, K. The privacy policy landscape after the GDPR. *Proceedings on Privacy Enhancing Technologies* **2020**, 47 (2020).

180. Degeling, M., Utz, C., Lentzsch, C., Hosseini, H., Schaub, F. & Holz, T. *We Value Your Privacy... Now Take Some Cookies: Measuring the GDPR's Impact on Web Privacy* in *Network and Distributed Systems Security (NDSS) Symposium* (2019).

181. Amos, R., Acar, G., Lucherini, E., Kshirsagar, M., Narayanan, A. & Mayer, J. *Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset* in *Proceedings of The Web Conference 2021* (Association for Computing Machinery, Ljubljana, Slovenia, 2021), 22.

182. Liepin, R., Contissa, G., Drazewski, K., Lagioia, F., Lippi, M., Micklitz, H.-W., Palka, P., Sartor, G. & Torroni, P. *GDPR privacy policies in CLAUDETTE: Challenges of omission, context and multilingualism* in *3rd Workshop on Automated Semantic Analysis of Information in Legal Texts, ASAIL 2019* **2385** (2019).

183. Harkous, H., Fawaz, K., Lebret, R., Schaub, F., Shin, K. G. & Aberer, K. *Polisis: Automated analysis and presentation of privacy policies using deep learning* in *27th USENIX Security Symposium (USENIX Security 18)* (2018), 531.

184. Bui, D., Shin, K. G., Choi, J.-M. & Shin, J. Automated Extraction and Presentation of Data Practices in Privacy Policies. *Proceedings on Privacy Enhancing Technologies* **2021**, 88 (2021).

185. Some, D. F., Bielova, N. & Rezk, T. *On the content security policy violations due to the same-origin policy* in *Proceedings of the 26th International Conference on World Wide Web* (2017), 877.

186. Englehardt, S. *et al. Automated discovery of privacy violations on the web* PhD thesis (PhD thesis, 2018).

187. Fietkau, J., Thimmaraju, K., Kybranz, F., Neef, S. & Seifert, J.-P. *The Elephant in the Background: A Quantitative Approach to Empower Users Against Web Browser Fingerprinting* tech. rep. (EasyChair, 2020).

188. Human, S., Schrems, M., Toner, A., Gerben & Wagner, B. *Advanced Data Protection Control (ADPC)* WorkingPaper (WU Vienna University of Economics and Business, 2021).

189. GPC Group. *Global Privacy Control (GPC)* `https://globalprivacycontrol.org`. 2022.

190. Ravichandran, D. & Vassilvitskii, S. *Evaluation of Cohort Algorithms for the FLoC API* `https://github.com/google/ads-privacy/blob/master/proposals/FLoC/FLOC-Whitepaper-Google.pdf`. 2021.

191. Berke, A. & Calacci, D. *Privacy limitations of interest-based advertising on the web: A post-mortem empirical analysis of Google's FLoC* in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security* (2022), 337.

192. Sim, J. & Wright, C. C. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy* **85**, 257 (2005).

193. Mathur, A., Kshirsagar, M. & Mayer, J. *What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods* in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (2021), 1.

194. Demir, N., Große-Kampmann, M., Urban, T., Wressnegger, C., Holz, T. & Pohlmann, N. *Reproducibility and replicability of web measurement studies* in *Proceedings of the ACM Web Conference 2022* (2022), 533.

A

APPENDIX

A.1 PRIVACY OF REGISTRATION

A.1.1 *Annotation process*

The dataset, legal instructions, and supplementary materials are available on request at `https://forms.gle/dTGpfs5vKqdLz8sQ7`. In this section, we provide additional information to the annotation process.

A.1.1.1 *Pilot study of annotation process*

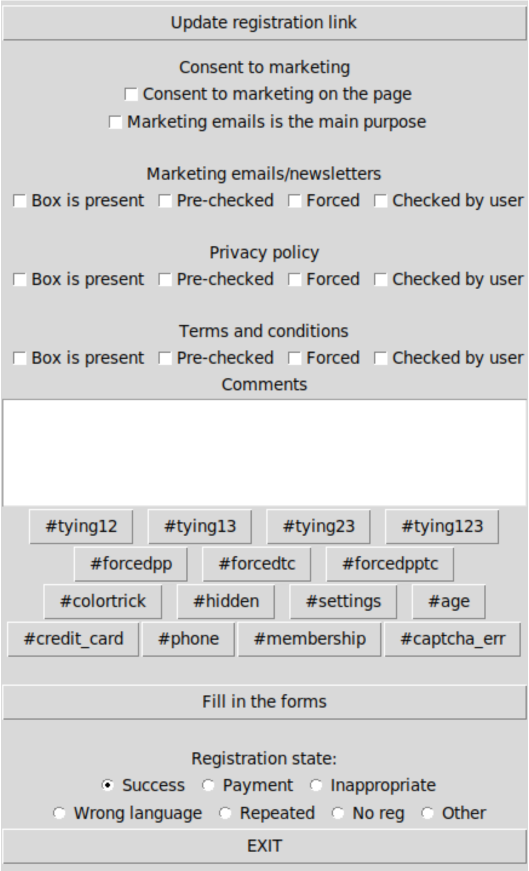
The exploratory pilot study of the annotation process aimed to test the clarity of our instructions and the completeness of our legal properties. Two legal research assistants each registered for 50 websites, selected by a similar website selection process without any pre-filtering. These annotators worked with an Excel spreadsheet to record their annotations using Boolean values and textual comments. After this pilot, we designed the annotation tool, significantly improved the annotator's instructions by reducing ambiguities and increasing the readability of the documentations. Moreover, we created a set of examples of 22 annotated websites with explanations for the annotations. Finally, we added labels to track the most common reasons for unsuccessful registrations.

A.1.1.2 *Annotating tool*

For easy deployment by various OSes, we package the whole annotating tool as a VirtualBox image based on Ubuntu 20.04. All the traffic of the system is routed via German proxy endpoint. We noticed that publicly available VPN and proxy endpoints are blocked by bot detection suites as Cloudflare, and even when the service is not blocked, the registration with such an IP address requires much longer reCAPTCHA solving time.

The system contains scripts for both registration and resolving annotation rounds. In the registration round, the annotator is provided a Firefox browser that is partially automated using Selenium library. This program automatically loads the registration page and the annotation interface illustrated in

Figure A.1. The annotators do not have to fill the credentials. Instead, they fill only keywords to required input fields and click *Fill in the forms* and the annotating tool substitutes these keywords by credentials generated for this website. For the resolving round, the annotator is provided with screenshots from the first two annotators with the difference among them highlighted, the two replicas of the annotation interface (again with a highlighted difference that he has to resolve), and a browser for checking something not visible in the screenshots.



Consent to marketing

□ Consent to marketing on the page

□ Marketing emails is the main purpose

Marketing emails/newsletters

□ Box is present □ Pre-checked □ Forced □ Checked by user

Privacy policy

□ Box is present □ Pre-checked □ Forced □ Checked by user

Terms and conditions

□ Box is present □ Pre-checked □ Forced □ Checked by user

Comments

Registration state:

● Success ○ Payment ○ Inappropriate

○ Wrong language ○ Repeated ○ No reg ○ Other

FIGURE A.1: Annotation tool interface. Both checkboxes and hashtags cover binary decisions. Their distinction is that, for hashtags, annotators often provide additional information as a note in the comment section. The registration state option captures if the registration was successful or why it failed. The second window of the tool is Firefox, controlled by the Selenium library, which loads the registration page in the first place and auto-fills the forms.

A.1.1.3 *Inter-annotator agreement*

Sim et al. [192] describe that Cohen's κ is not a proper statistics for highly imbalanced variables (high *prevalence*) or biased variables, which is our case for several of the legal properties, notably those with very low κ in Table A.1. Therefore, we also present the contingency tables for every legal property in Tables A.2 and A.3.

A.1.1.4 *Linkage to dark patterns*

In this section, we compare our defined potential violation types to the taxonomy of dark patterns by Marthur et al. [193]. We refer to terms from [193] in *italics.*

Both "Email despite no opt-in" and "Email despite user did not consent" are potential violations of consent, so they are *restrictive* dark patterns. "Email after invalid consent" in all four cases constitutes a dark pattern. Namely, unspecific and unfree forms are *restrictive*, ambiguous forms are *asymmetric*, and forms that use nudging are instances of *convert* and *information hiding.*

Of the potential violations in the email content, there were marketing emails trying to resemble servicing emails. Most of these emails were annotated as marketing-notifications, because their appearance suggests that they are triggered by user's activity. By checking both accounts for the service, we found that both of our addresses were receiving the same notifications, and hence the emails are not user-triggered, which is *deceptive.* When an email is missing the unsubscribe option, it is *restrictive.*

A.1.1.5 *Registered accounts*

Annotators registered to the selected 1000 websites in both annotating rounds. Each of the rounds resulted in a different number of successful registrations, namely 576 in the first round, and 582 in the second round. The intersection of successful registration is 500 websites and the union is 701 websites, which is the number of websites that we assume can send us emails. The difference is caused by 34 websites that were inaccessible during one of the rounds and differences in how the annotators browse the website to find the registration form.

Note that if we would have to split the registration and annotation processes, we would lose significant information. The annotators need to see the whole registration to determine all the legal properties. In addition, the annotators would be provided a potentially wrong form, which by our approach would not

Table A.1: The individual Cohen's κs of legal properties. Note that $\kappa = 1$ implies full agreement, while $\kappa = -1$ implies full disagreement.

Checkbox	κ	**Hashtag**	κ
mark_consent	0.77	#tying12	0.12
mark_purpose	0.61	#tying13	1.00
ma_checkbox	0.77	#tying23	0.74
ma_pre_checked	0.77	#tying123	0.00
ma_forced	0.53	#forcedpp	0.70
pp_checkbox	0.77	#forcedtc	0.56
pp_pre_checked	0.44	#forcedpptc	0.75
pp_forced	0.75	#hidden	0.08
tc_checkbox	0.78	#settings	0.00
tc_pre_checked	0.75	#age	0.62
tc_forced	0.73		

Table A.2: Contingency tables of checkbox values. Rows represent the first annotation, the second annotation is depicted by the column.

	True	**False**
True	244	42
False	51	663

(a) mark_consent

	True	**False**
True	192	41
False	41	726

(b) ma_checkbox

	True	**False**
True	32	10
False	8	950

(c) ma_pre_checked

	True	**False**
True	10	7
False	10	970

(d) ma_forced

	True	**False**
True	34	15
False	25	926

(e) mark_purpose

	True	**False**
True	187	40
False	40	733

(f) pp_checkbox

	True	**False**
True	2	4
False	1	993

(g) pp_pre_checked

	True	**False**
True	169	42
False	43	746

(h) pp_forced

	True	**False**
True	165	37
False	34	764

(i) tc_checkbox

	True	**False**
True	6	3
False	1	990

(j) tc_pre_checked

	True	**False**
True	143	40
False	42	775

(k) tc_forced

TABLE A.3: Contingency tables of hashtag values. Rows represent the first annotation, the second annotation is depicted by the column.

	True	**False**
True	1	7
False	7	985

(A) #tying12

	True	**False**
True	0	0
False	0	1000

(B) #tying13

	True	**False**
True	86	26
False	25	863

(C) #tying23

	True	**False**
True	0	4
False	1	995

(D) #tying123

	True	**False**
True	131	46
False	41	782

(E) #forcedpp

	True	**False**
True	22	18
False	14	946

(F) #forcedtc

	True	**False**
True	100	30
False	25	845

(G) #forcedpptc

	True	**False**
True	4	32
False	31	933

(H) #hidden

	True	**False**
True	63	26
False	41	870

(I) #age

	True	**False**
True	0	4
False	2	994

(J) #settings

be resolved by resolving annotation. Moreover, we would not have subscribed to many of the 701 websites.

A.1.1.6 *Email address generation*

We considered two options for generating emails: setting up a custom email server or using Gmail "+ suffixes." An appended + sign and any combination of alphanumeric characters are ignored for resolving the recipient for Gmail addresses. This way, `john@gmail.com` also receives emails from `john+friends@gmail.com`. We chose the custom email server as it cannot be detected and exploited by marketing services. This differs from [40] that used Gmail suffixes.

A.1.2 *Datasets content*

We now elaborate on our dataset described in Section 3.1.2, showing insights that help to understand the content and to illustrate other potential applications of the dataset.

Note that following ethical principles, we had to redact our datasets. We removed all the URLs and credentials within both the email and website datasets. The redacted datasets suit the goals of automated potential violation detection as well as the full dataset.

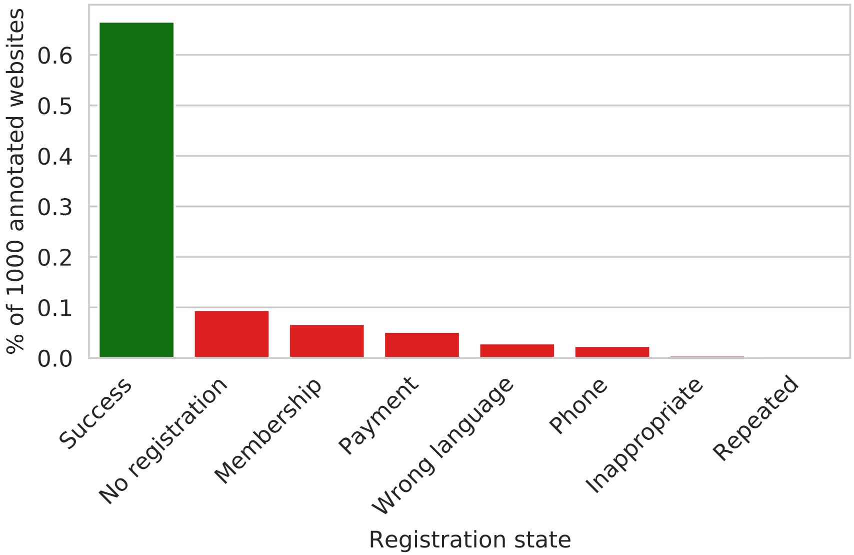
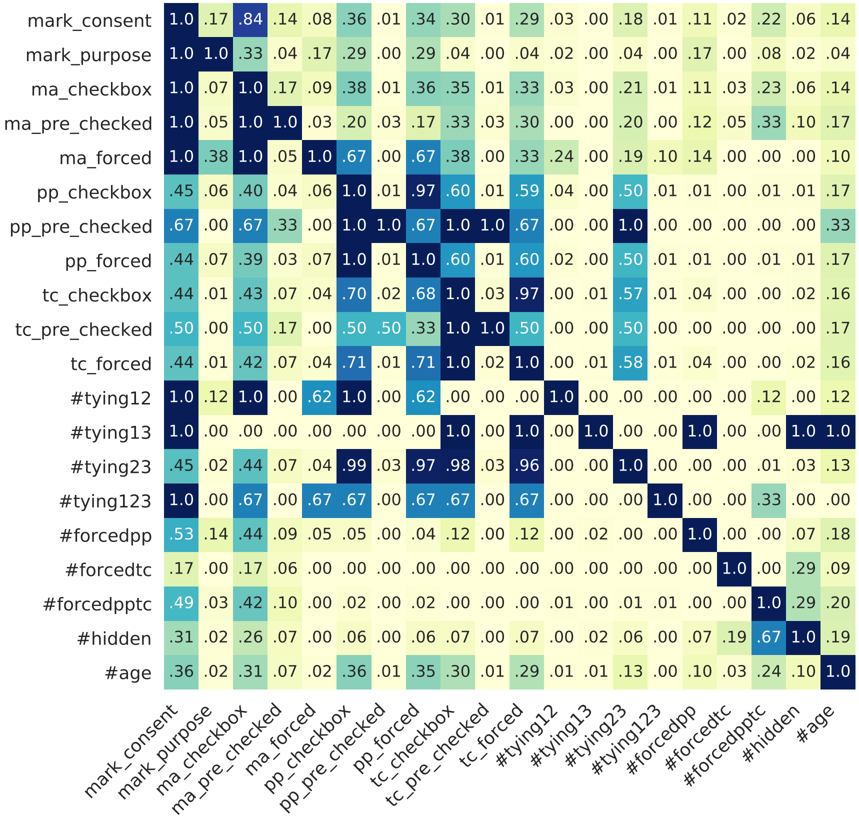


Figure A.2: Registration state of the resolved annotations. For agreement, Cohen's κ for distinction between successful and failed registrations is 0.64.

A.1.2.1 *Successful form annotations*

In Figure A.2, we present the outcomes of the registration process, showing that 70% of registrations were successful, and listing how often and why the registration failed.

Figure A.3 shows interdependence between legal properties of successful annotations. It illustrates that 97% of the privacy policy and term and conditions checkboxes are pre-checked. Another observation is that websites with pre-checked marketing checkbox more likely pre-check other checkboxes, or force the acceptance of terms and conditions and privacy policy.



	mark_consent	mark_purpose	ma_checkbox	ma_pre_checked	ma_forced	pp_checkbox	pp_pre_checked	pp_forced	tc_checkbox	tc_pre_checked	tc_forced	#tying12	#tying13	#tying23	#tying123	#forcedpp	#forcedtc	#forcedpptc	#hidden	#age
mark_consent	1.0	.17	.84	.14	.08	.36	.01	.34	.30	.01	.29	.03	.00	.18	.01	.11	.02	.22	.06	.14
mark_purpose	1.0	1.0	.33	.04	.17	.29	.00	.29	.04	.00	.04	.02	.00	.04	.00	.17	.00	.08	.02	.04
ma_checkbox	1.0	.07	1.0	.17	.09	.38	.01	.36	.35	.01	.33	.03	.00	.21	.01	.11	.03	.23	.06	.14
ma_pre_checked	1.0	.05	1.0	1.0	.03	.20	.03	.17	.33	.03	.30	.00	.00	.20	.00	.12	.05	.33	.10	.17
ma_forced	1.0	.38	1.0	.05	1.0	.67	.00	.67	.38	.00	.33	.24	.00	.19	.10	.14	.00	.00	.00	.10
pp_checkbox	.45	.06	.40	.04	.06	1.0	.01	.97	.60	.01	.59	.04	.00	.50	.01	.01	.00	.01	.01	.17
pp_pre_checked	.67	.00	.67	.33	.00	1.0	1.0	.67	1.0	1.0	.67	.00	.00	1.0	.00	.00	.00	.00	.00	.33
pp_forced	.44	.07	.39	.03	.07	1.0	.01	1.0	.60	.01	.60	.02	.00	.50	.01	.01	.00	.01	.01	.17
tc_checkbox	.44	.01	.43	.07	.04	.70	.02	.68	1.0	.03	.97	.00	.01	.57	.01	.04	.00	.00	.02	.16
tc_pre_checked	.50	.00	.50	.17	.00	.50	.50	.33	1.0	1.0	.50	.00	.00	.50	.00	.00	.00	.00	.00	.17
tc_forced	.44	.01	.42	.07	.04	.71	.01	.71	1.0	.02	1.0	.00	.01	.58	.01	.04	.00	.00	.02	.16
#tying12	1.0	.12	1.0	.00	.62	1.0	.00	.62	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.12	.00	.12
#tying13	1.0	.00	.00	.00	.00	.00	.00	.00	1.0	.00	1.0	.00	1.0	.00	.00	1.0	.00	.00	1.0	1.0
#tying23	.45	.02	.44	.07	.04	.99	.03	.97	.98	.03	.96	.00	.00	1.0	.00	.00	.00	.01	.03	.13
#tying123	1.0	.00	.67	.00	.67	.67	.00	.67	.67	.00	.67	.00	.00	.00	1.0	.00	.00	.33	.00	.00
#forcedpp	.53	.14	.44	.09	.05	.05	.00	.04	.12	.00	.12	.00	.02	.00	.00	1.0	.00	.00	.07	.18
#forcedtc	.17	.00	.17	.06	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.29	.09
#forcedpptc	.49	.03	.42	.10	.00	.02	.00	.02	.00	.00	.00	.01	.00	.01	.01	.00	.00	1.0	.29	.20
#hidden	.31	.02	.26	.07	.00	.06	.00	.06	.07	.00	.07	.00	.02	.06	.00	.07	.19	.67	1.0	.19
#age	.36	.02	.31	.07	.02	.36	.01	.35	.30	.01	.29	.01	.01	.13	.00	.10	.03	.24	.10	1.0

FIGURE A.3: Interdependence of legal properties as a ratio of annotations with the property of the row that has also the property of the column. A cell in the first row, second column, marks how many websites with marketing consent (row label) have the marketing purpose (column label).

A.1.2.2 *Email classification*

In appendix A.1.2.2 we summarize the presence of all potential violations discussed in this study. In addition, we split the graph into groups by website's ranking according to their Alexa rank. Note that more popular websites are not more compliant than lower ranked websites. Moreover, for the potential violation "Email despite no opt-in," the websites with high rank show more potential violations than those with low rank (*p*-value of the two proportions Z-Test of the rank < 1k against data of all other ranks is 0.156 after adjustment for multiple measurements by Holm–Bonferroni method). The number of websites of rank 1k-10k not sending legal notices is far larger than the websites of other ranks (including high-rank websites). This observation has a *p*-value of 0.054.

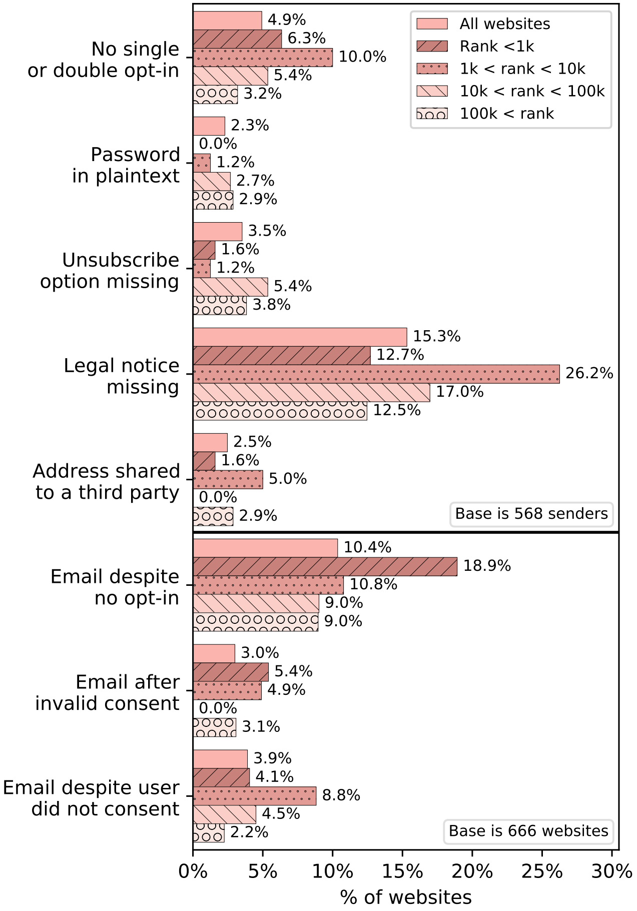


Figure A.4: Summary of all potential violations of this study and the split into popularity groups by rank.

A.1.2.3 *Third party email sharing*

As we stated in Section 3.3.3.2, we classified four websites as "other." In the first case, apart from the same physical address, we did not have enough indications that the two Chinese companies were part of the same group. In the second case, the third party was maintaining a reward system on behalf of the website. The third website was offline and could no longer be analyzed. The last service's data likely breached (reported by other users), which led to us receiving fraudulent emails. The service did not notify its users about any breach.

A.1.2.4 *Marketing trends in newsletters*

For our study, we annotated emails during the period starting in September 2020 and ending in February 2021, so we were able to observe several marketing trends influencing the email content. We observed that 5.8%, 11.7%, and 4.2% of marketing emails were related to Black Friday, Christmas, and New Year, respectively. These topics become relevant during autumn and winter, but we did not observe an overall increase in the number of marketing emails. Also, 17.2% of all processed emails were related to the Covid pandemic. As the frequency of marketing emails did not change during these periods (see Figure A.5), the observations suggest that trending topics are used to improve marketing campaigns, but they do not generate new newsletter traffic. This hypothesis is based on the fact that during the limited period of the study, we did not observe any spikes in the number of newsletters during these periods. However, to confirm this hypothesis, we would need a more longitudinal study.

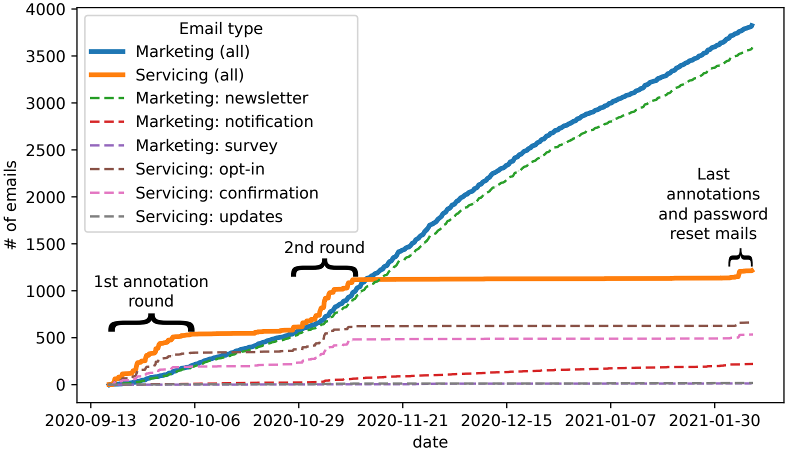


FIGURE A.5: Classification of the manually annotated emails, where reported marketing and servicing numbers are the sum of the number of emails of each subtype. The x-axis is continuous over the period of our study. We can see that the number of servicing emails is constant function in number of registrations ($\approx 1.2 \cdot$ number of accounts), while number of emails linearly increases over time ($\approx$ 2 emails per day per 100 accounts). The decrease in the email frequency by the end of our study may be caused by services removing us from their recipient list due to a long inactivity.

A.1.3 *Bot-evasion techniques*

Given the experience from bot detection with our annotation tool appendix A.1.1.2, we implemented the following methods to further decrease the chance of our crawling being detected as a bot activity.

BROWSER: We use Undetected Chromedriver,[1] which extends the usual Chromedriver with numerous bot evasion techniques, such as removing fingerprints unique to Selenium. Unfortunately, there is no equivalent driver available for Firefox.

FINGERPRINTING EVASION: For each page load, the crawler checks the load status. This functionality is not directly implemented by Selenium, so we use Chrome DevTools Protocol for Chrome and Selenium Wire for Firefox. The use of Selenium Wire is however prone to TLS fingerprinting. The proxy and browser differ in the ciphersuite, which is inspected by modern bot detection systems like Cloudflare. While the Firefox-based crawler is prone to this detection, the Chrome implementation does not use any proxy. Additionally we must run Chrome with a non-root user. Chrome disables sandboxing protections when run as root, making it flagged as a bot by Cloudflare.

INTERACTION SPEEDS: Interactions with the website cannot occur instantaneously, as humans are limited in their reading and writing speeds. Our crawler introduces random time delays before each click and during typing to mimic human behavior.

IP ADDRESS: As we study the impact of the EU's privacy regulations, we focused our data collection on traffic originating from within the EU. We considered using commercial VPNs, datacenter or residential proxies, or a university VPN located in the EU. According to a study by Demir et al. [194], residential proxies are the least likely to be detected as bot traffic, closely followed by university VPNs, while datacenters and commercial VPNs are blocked more frequently. Since purchasing a large number of residential IP addresses from services like Bright Data is expensive ($\geq$\$10k for our crawl), we used a VPN provided by a university in Germany, which gave us access to a block of 12 IP addresses.

[1] `https://github.com/ultrafunkamsterdam/undetected-chromedriver`

A.1.4 *Supported languages*

Our crawler supports 36 languages, with most of the keywords being translated by native or fluent speakers of the language, whom we instructed in collecting multiple example websites prior to the translation. These languages are: Bulgarian, Bosnian, Catalan, **Czech**, Welsh, **Danish**, **German**, **Greek**, **English**, **Spanish**, Estonian, Basque, **Finnish**, **French**, Galician, Croatian, **Hungarian**, Icelandic, **Italian**, Luxembourgish, Lithuanian, Latvian, Macedonian, Maltese, **Dutch**, Norwegian, **Polish**, **Portuguese**, Romanian, **Russian**, **Slovak**, Slovenian, Albanian, Serbian, **Swedish**, **Turkish**, and **Ukrainian**. From these languages, only 18 of them are supported by LibreTranslate and therefore are suitable for detection of all the violations. We highlighted these languages in bold.

A.1.5 *Crawler form classification*

Our crawler distinguishes various form fields, which we aggregate to the following groups for the form feature processing. This fixed structure allows us to process differently ordered forms using the same tabular pattern.

- mail
- password
- phone
- username
- names: first, middle, last or full name
- name-other: organization, title, honorific prefix, other text fields
- address: street, house number, city, ZIP, country, full address
- age
- sex
- checkbox: terms of service
- checkbox: privacy policy
- checkbox: privacy policy and terms of service
- checkbox: marketing, privacy policy and terms of service
- checkbox: marketing
- checkbox: SMS
- checkbox: age
- checkbox: other
- birthday: day, month, year, full birth, other <select>
- submit buttons: registration, subscribe
- other buttons: login, contact, other

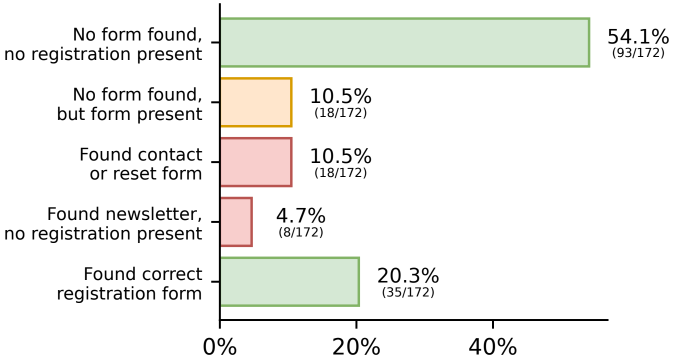


FIGURE A.6: Evaluation of crawler-detected registration forms.

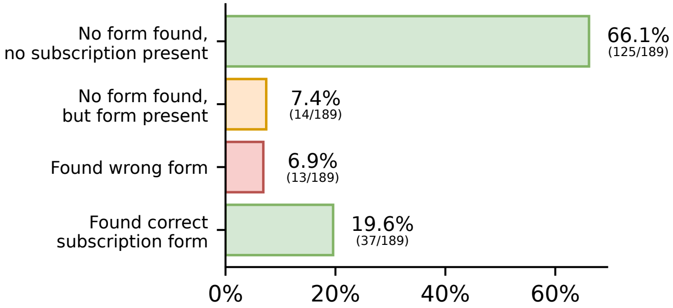


FIGURE A.7: Evaluation of crawler-detected subscription forms.

A.1.6 *Manual analysis of the crawler*

We conducted a manual investigation of 200 crawled websites to evaluate form detection. Out of the 200 pages, 19 failed to load, and thus the analysis presented below pertains to the remaining 181 websites.

In Fig. A.6, we present the evaluation of registration form detection. Among the sampled websites, 55 had a registration form, of which our crawler successfully detected two-thirds. Additionally, the crawler identified a wrong form (e.g., a contact form or password reset form) in 10.5% of the evaluated websites. Furthermore, in 4.7% of the websites, the crawler misclassified a subscription form as a registration form.

Fig. A.7 illustrates the evaluation of discovered subscription forms. Our findings reveal that 73.0% of websites do not have a subscription form (although note that many websites contain both a subscription form and a registration form). The crawler accurately determined the absence of this form on two-thirds of the websites, and on 19.6% of the websites, it correctly identified the existing form. However, the crawler failed to detect the subscription form on 7.4% of the analyzed websites, and in 6.9% of websites, it found an incorrect form.

We also inspected the detected privacy policies and terms and conditions on a list of 300 websites. Our manual evaluation showed that almost 80% and 70% of websites contain privacy policy and terms and conditions, respectively. Our crawler can then detect the correct privacy policy on 51% of websites and correctly conclude that there is no policy on 21% of websites. On 19% of websites it fails to find the policy and in the remaining 9% of cases it finds a wrong document. The crawler is correct in finding the terms and detects the absence of terms on 37% and 21% of websites, respectively. It failed to detect terms on 13% of websites and in the remaining 29% of cases, it detects a wrong document.

A.1.7 *Other results*

A.1.7.1 *Alternatives for detecting 3rd-party sharing*

In addition to the described methods in Section 3.3.3.2, we explored the following methodologies minimizing false positives and negatives in our third-party sharing violation detection.

TLS CERTIFICATES. We considered the extraction of company information from TLS certificates. However, note that only a minority, less than 30%, of websites include company names within their TLS certificates. This practice is predominantly observed among highly popular websites, whereas our automated crawling and classification methods perform the best on websites of medium popularity. Furthermore, our observations revealed that websites associated with the same parent companies commonly employ different company names in their certificates, dismissing the usefulness of this approach.

CO-OCCURRENCES. We investigated the co-occurrence of senders who send emails to multiple addresses registered by our crawler. This analysis unveiled two distinct scenarios. First, email hosting providers such as

TABLE A.4: Resulting p-values of Fisher's exact test with Holm-Bonferroni correction testing the hypothesis that the observations of manual pilot study and automated large-scale study are sampled from the same distribution. We reject the hypothesis when p-value < 0.001.

Observation	p-value
Insecure registration	2e-66,
Password in plaintext	0.86
Email despite no opt-in	9e-14,
Email after invalid consent	0.40
Email despite user did not consent	0.45
Marketing email first	2e-58,
Single-opt-in first	1e-08,
Double-opt-in first	3e-14,
Email-sharing violation	0.83

Gmail were observed to send emails to multiple accounts, suggesting that co-occurrence could be indicative of websites that are compliant with privacy regulations. Conversely, we identified clusters of websites that shared email addresses among themselves without belonging to the same corporate group and without obtaining proper user consent, which strongly indicated privacy violations.

COMPANY DATABASES. We explored the use of databases such as Whois, Crunchbase, or Orbis to discover connections between domains owned by the same companies. However, Whois data has become increasingly sparse due to privacy concerns. Moreover, both Crunchbase and Orbis feature inconsistent company name records, leading to false positive violation reports and occasionally attributing incorrect company names, resulting in false negative violation reports. We also considered the webXray dataset curated by Libert [75],[2] but it primarily targets third parties within the tracking industry, which seldom overlap with email senders.

[2] https://github.com/agilemobiledev/webXray/blob/master/webxray/resources/org_domains/org_domains.json

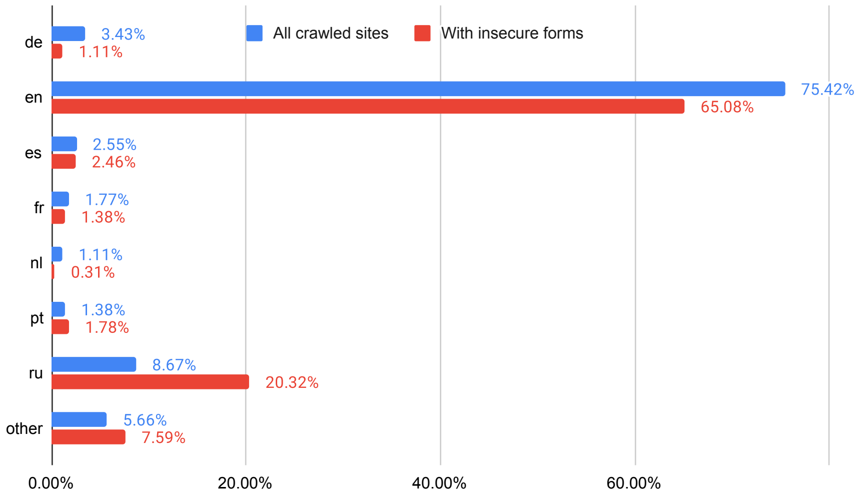


FIGURE A.8: Language inspection of insecure forms, with blue bars representing language distribution for reference.

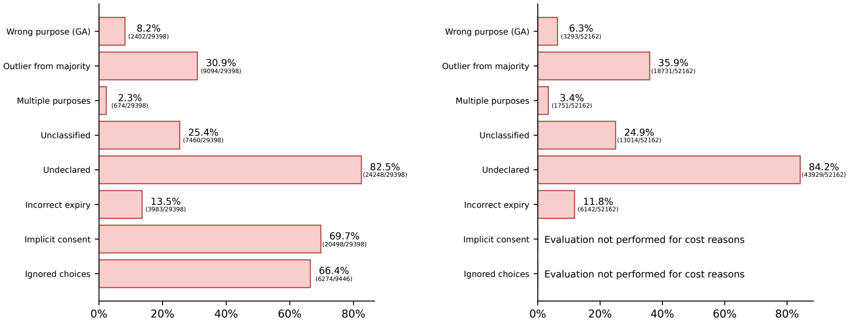
A.2 PRIVACY OF COOKIES

A.2.1 *CMP data versus Cookiepedia*

A.2.1.1 *Rationale for cookie scraping*

We considered collecting the ground truth for the training dataset by querying Cookiepedia, but we decided to scrape CMPs instead for multiple reasons, which we list below.

- The CMP descriptions are a primary source with the purpose either assigned or confirmed by the website administrator. Cookiepedia is a third-party, and despite the purposes being assigned by experts, they do not have complete information about the intentions the web administrators had.
- Scraping CMPs also allows us to analyze their compliance, which motivates client-side cookie policy enforcement.
- Cookiepedia identifies cookies by their name and not by the more specific identifier of the name and domain. This means that cookies of



(A) Violations observed in May 2021.

(B) Violations observed in July 2022.

FIGURE A.9: Observed violations based on methods from Section 4.2.

TABLE A.5: Performance of XGBoost when applied on our reduced cookie dataset labeled using Cookiepedia.

XGBoost	Necessary	Functional	Analytics	Advertising
F1 score	86.2% ±1.1%	59.3% ±4.7%	95.2% ±1.2%	89.0% ±1.1%
Accuracy: 89.2% ± 1.3%				

the same name used by different domains for different purposes would cause noise for training.

- For a long period during the course of our study, Cookiepedia was not accessible, and as such, it would have been a single point of failure for our data collection. Individual sites with CMPs can also be inaccessible, but their distributed nature ensures that we can always collect sufficient dataset for training.

A.2.1.2 *Classification using Cookiepedia labels*

To better compare our approach with the work of Hu et al. from [139], we applied a sequence of transformations to bring our model assumptions closer to theirs. Namely, we applied the following changes:

1. We replace the ground truth labels of our cookie dataset with labels queried from Cookiepedia. We discard all cookies for which Cookiepedia does not have a category, thus reducing the size of our dataset by 21%.
2. We reduce the number of our training samples further by randomly sampling a single cookie for each unique name. This is necessary because Cookiepedia always assigns the same label to the same name, while our dataset from CMPs could contain cookies of the same name with different purposes. Having many duplicate names would falsify the validation score.
3. We train an XGBoost model on the new dataset, and report the per-class F1 score, and overall micro F1 score.

The resulting values are presented in Table A.5. Notice that our micro F1 score, which in this setting is equivalent to the accuracy, has increased from 87.2% to 89.2%. Furthermore, this F1 score is better than the F1 score of 86.7% from [139], which can be recomputed from the reported confusion matrix, but lower than their stated micro F1 score of 94.6%.

A.2.2 *Feature extraction and hyperparameters*

In the following, we provide more details about the feature extraction and the model's hyperparameters. For the most detailed overview, please refer to the extended report [25] and project documentation.[3]

Feature types. We extract three major types of features from the cookies. First, from each cookie, we extract features from its attributes, presented in Table A.8. Second, with each cookie update, we store the updated features listed in Table A.9. The number of updates used for the feature extraction is configurable. By default we use two, so that the classification does not require longer observations of the cookie, which is a trade-off for model performance. Finally, starting with the first update, we compute the difference to the previous version of the cookie, which are the per-difference features we show in Table A.7.

Classifier hyperparameters. In Table A.6 we show the parameters we selected for training the XGBoost model. They were selected through the use of a randomized grid-search and 5-fold cross-validation.

[3] The feature documentation and classifier are available at: `https://github.com/dibollinger/CookieBlock-Consent-Classifier`.

Table A.6: Set of hyperparameters used for training the model with XGBoost, listed here for reproducibility.

Parameter name	**Value**
Booster type	`‘gbtree’`
Tree method	`‘hist’`
Learning objective	`‘multi:softprob’`,
Evaluation metrics	`‘merror’` and `‘mlogloss’`
Learning rate	0.25
Maximum tree depth	32
Minimum split loss	1
Minimum child weight	3
Maximum delta step	0 (no limit)
Subsample ratio	1.0
Alpha (L1 regularizer)	2
Lambda (L2 regularizer)	1
Tree growth policy	`‘depth-wise’`
Maximum bins	256

TABLE A.7: Per-difference features overview: All features that are extracted as comparisons between two contiguous updates, sorted by timestamp.

Feature name	**Description**
Expiry difference (1)	Expiration time difference in seconds between two updates.
“Difflib” similarity (1)	Similarity ratio between cookies, as measured by “difflib.”
Levenshtein distance (1)	Levenshtein distance between two cookie updates.

Table A.8: Per-cookie features overview: All features that are extracted once per unique cookie. Entries marked with a * may cause issues when used within the context of a browser extension. In the parentheses after the name, we show the number of vector entries the feature takes.

Feature name	**Description**
Top names (500)	One-hot vector of the most common cookie names.
Top domains (500)	One-hot vector of the most common domains.
Pattern names (50)	One-hot vector of the most common name patterns.
Name tokens (500)	Binary indicator of English tokens in the name.
IAB vendor (1)	Binary indicator, true if domain is an IAB vendor.
Domain period (1)	Indicates whether the domain starts with a period char.
Third-party* (1)	Whether the cookie originates from a third-party.
Non-root path (1)	Whether the cookie path is not the root path.
Update count* (1)	Total number of updates encountered for this cookie.
“HostOnly” flag (1)	Whether the “HostOnly” flag is set.
“HttpOnly” changed (1)	Whether the “HttpOnly” flag changed in any update.
“Secure” changed (1)	Whether the “secure” flag changed in any update.
“Same-Site” changed (1)	Whether the “same-site” flag changed in any update.
“Session” changed (1)	Whether the “session” flag changed in any update.
“Expiry” changed (1)	Whether the expiry changed by 1+ days between updates.
Content changed (1)	Whether the cookie content changed between updates.
Levenshtein total (2)	Mean and StdDev of Levenshtein dist. between updates.
Difflib total (2)	Mean and StdDev of Difflib similarity between updates.
Length total (2)	Mean and StdDev of the cookie value length in bytes.
Compressed total (2)	Mean and StdDev of the compressed cookie value length.
Entropy total (2)	Mean and StdDev of the Shannon Entropy of values.

Table A.9: Per-update feature overview: All features that are extracted once per cookie update. The number of updates used for extraction can be specified separately.

Feature name	**Description**
“HttpOnly” flag (1)	Binary indicator of whether the “HttpOnly” flag is set.
“Secure” flag (1)	Binary indicator of whether the “secure” flag is set.
“Session” flag (1)	Whether the cookie is a session cookie or not.
“Same-Site” flag (3)	Whether “None,” “Lax,” or “Strict” is set.
Expiration time (1)	Ordinal feature, contains the expiry in seconds.
Expiration intervals (8)	Interval checks on expiry, e.g., $>$ 1 day, $<$ 1 week.
Content length (1)	Total size of the cookie’s value in bytes.
Compressed length (1)	Size of the cookie value after *zlib* compression.
Compression rate (1)	Reduction of the size after *zlib* compression.
Shannon entropy (1)	Shannon entropy of the cookie update’s value.
URL encoding (1)	Indicates whether the cookie value is URL encoded.
Base64 encoding (1)	Indicates that the value is potentially Base64 encoded.
Delimiter separation (9)	Delimiter (CSV) separation type and #separators.
Contains JSON (1)	Whether the value contains a JSON object.
Content terms (50)	Binary indicator of English tokens in the value.
CSV contents (5)	Try to split as CSV and detect value types within.
JS contents (11)	Try to split as JSON and detect value types within.
Numerical content (1)	Whether the value consists entirely of digits.
Hexadecimal content (1)	Whether the value represents a hexadecimal number.
Alphabetical content (1)	Whether the value is entirely alphabetical.
Identifier content (1)	Whether the value is a valid code identifier.
All uppercase (1)	Whether the cookie value has all uppercase letters.
All lowercase (1)	Whether the cookie value has all lowercase letters.
Empty content (1)	Whether the value of the cookie is empty.
Boolean content (1)	Whether the cookie value is a boolean of some form.
Locale content (1)	Whether the value includes a country identifier.
Timestamp content (1)	Whether a UNIX timestamp is in the cookie value.
Date content (1)	Whether the value contains a date term or identifier.
URL content (1)	Whether the value contains a URL of some form.
UUID content (6)	Which UUID variant, if present in the value.

Karel Kubíček

Curriculum Vitae

Education

2018–Present **Doctoral student**, *Department of Computer Science, ETH Zurich*, Zurich, CH
Information Security Group.

2015–2017 **Master's Studies**, *Faculty of Informatics, Masaryk University (FI MU)*, Brno, CZ
Information Technology Security. Exchange semester at NTNU, Trondheim, NO.

2012–2015 **Bachelor's Studies**, *Faculty of Informatics, Masaryk University (FI MU)*, Brno, CZ
Computer Systems and Data Processing.

Publications

2024 *Automated, Large-Scale Analysis of Cookie Notice Compliance*, USENIX Security

2024 *Block Cookies, Not Websites: Analysing Mental Models and Usability of the Privacy-Preserving Browser Extension CookieBlock*, Proc. on Privacy Enhancing Technologies

2023 *Locality-Sensitive Hashing Does Not Guarantee Privacy! Attacks on Google's FLoC and the MinHash Hierarchy System*, Proceedings on Privacy Enhancing Technologies

2022 *Checking Websites' GDPR Consent Compliance for Marketing Emails*, Proceedings on Privacy Enhancing Technologies

2022 *Automating Cookie Consent and GDPR Violation Detection*, USENIX Security, **best artifact award**

2022 *Large-scale Randomness Study of Security Margins for 100+ Cryptographic Functions*, SECRYPT

2019 *BoolTest: The Fast Randomness Testing Strategy Based on Boolean Functions with Application to DES, 3-DES, MD5, MD6 and SHA-256*, E-Business and Telecommunications, Springer International Publishing

2017 *New results on reduced-round Tiny Encryption Algorithm using genetic programming*, Infocommunications journal

Awards

2022 1st place in CSAW'22 Europe Applied Research Competition for our USENIX paper *Automating Cookie Consent and GDPR Violation Detection*.

2017 Awarded the second place in the contest for the best thesis in the field of IT Security.

2013–2017 Various scholarships for contribution in student projects (Czech Science Foundation, university foundation), merit scholarships.

Experience

2018–Present **Doctoral student at ETH Zurich**, Information Security Group, Zurich

- Automated studies of websites' compliance with the EU privacy laws (GDPR, ePD).
- Teaching Information security, Algorithms; supervision of 11 MSc and BSc student theses.
- Board member of Academic staff organisation VMI.

2014–2018 **Development of randomness testing framework EACirc for analysis of cryptographic primitives**, Centre for Research on Cryptography and Security, FI MU, Brno

- Implementation and comparison of metaheuristics in EACirc (randomness testing framework).
- Analysis of Tiny Encryption Algorithm (TEA) using EACirc framework.

2017 Jan–Sep **Network security researcher**, NEXA Technologies CZ, Brno

- Working on R&D project in the area of cryptography, privacy, and machine learning.
- References: Jaroslav Šeděnka ✉ , Martin Stehlík ✉

2013–2017 **Seminar tutor of Algorithms and Automata's theory courses**, FI MU, Brno

- 2013–2017 Algorithms and data structures (I and II) course (BSc and MSc levels).
- 2015–2016 Automata, grammars, and complexity course.
- Composing an exercise book: 160 pages book with exercises and their solutions.
- Preparing and correcting assignments and final programming tasks.

2013–2017 **Contribution on organizing informatics seminar, competitions and puzzle hunts for both secondary-school and university students**, FI MU, Brno

- 2015 – Leading the organization of competition InterSob (30 team members, four months).

Featured Skills

Advanced	Python, privacy regulations, ML and data science, cryptography
Intermediate	LaTeX, C, C++, algorithm design, optimisation methods, DevOps
Basic	Haskell, Java, R, automata's theory, secure coding

Languages

Czech	Mothertongue	
English	Full professional proficiency	*C1*
German	Limited working proficiency	*B1*
Norwegian	Basic words and phrases only	*A1*

Interests

- Paragliding, mountaineering
- Work in education system
- Outdoor sports
- Puzzlehunting