Automating Website Registration for Studying GDPR Compliance, Authors' Extended Version*

Karel Kubicek
ETH Zurich
Zurich, Switzerland
karel.kubicek@inf.ethz.ch

Jakob Merane
ETH Zurich
Zurich, Switzerland
jakob.merane@gess.ethz.ch

Ahmed Bouhoula
ETH Zurich
Zurich, Switzerland
ahmed.bouhoula@inf.ethz.ch

David Basin
ETH Zurich
Zurich, Switzerland
basin@inf.ethz.ch

ABSTRACT

Investigating how websites use sensitive user data is an active research area. However, research based on automated measurements has been limited to those websites that do not require user authentication. To overcome this limitation, we developed a crawler that automates website registrations and newsletter subscriptions and detects both security and privacy threats at scale.

We demonstrate our crawler's capabilities by running it on 660k websites. We use this to identify security and privacy threats and to contextualize them within EU laws, namely the General Data Protection Regulation and ePrivacy Directive. Our methods detect private data collection over insecure HTTP connections and websites sending emails with user-provided passwords. We are also the first to apply machine learning to web forms, assessing violations of marketing consent collection requirements. Overall, we find that 37.2% of websites send marketing emails without proper user consent. This is mostly caused by websites failing both to verify and store consent adequately. Additionally, 1.8% of websites share users' email addresses with third parties without a transparent disclosure.

KEYWORDS

Crawling, Registration, Consent, GDPR, ePrivacy, Compliance

1 INTRODUCTION

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 [74]. 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 [38, 62] or e-commerce [22]), 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 publication is structured as follows: In Section 2, we describe our manual pilot study. Specifically, in Section 2.1 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 2.2 and emails in Section 2.3. After the pilot study, in Section 3 we report automating and scaling up our manual procedures. In Section 3.1, we describe the development of a crawler designed to automate the website sign-up process. In Section 3.2, 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 4. Finally, our automated findings are subjected to a manual inspection in Section 5.

2 PILOT STUDY

2.1 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

*This extended version combines the manual annotation study [53] with its follow-up that automated the registration and violation detection processes [52]. Additionally, this version includes further comparisons from PhD thesis by Kubicek [51]. If you utilize this work, kindly cite the individual publications according to your context rather than this author version. For additional information, please visit the following URL: https://karelkubicek.github.io/post/reg-www.

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 [64, p. 79]. While some member states like Germany adapted such a regime early on [21, p. 168], the ePrivacy Directive [31] has established the opt-in requirement in July 2002 at European level [20, 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*) [15], 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 [60]. Although controversially discussed, providing personal data as payment for "free" services is insufficient to generally trigger the exception ([76] in discussion of [48]). To protect customers from unsolicited commercial communications, legal scholars and German courts have tended to interpret the exception strictly [66]. 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 [30] or the German Telemedia Act (TMG) as the corresponding national implementation [16]. Furthermore, the EU's Directive on Unfair Commercial Practices (UCPD) [32] specifically bans persistent and unwanted solicitations by email [32, 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 [25, 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 [18, p. 10].

2.1.1 Valid consent under the GDPR. The interplay between the ePrivacy Directive, the UWG, and the GDPR is complex [26], 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 [29]. 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 [47]. The European Court of Justice also agreed with this view in the underlying preliminary ruling [42].

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) [28].

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 [40, 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 [45]. 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 [43]. The mere acceptance of general terms and conditions or privacy policies is also deemed insufficient [28, 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 [41, 42]. 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.

2.1.2 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.0.1 for more information about the pilot study), and refined the legal properties accordingly. In Section 4.2, we present a decision method that determines whether a website potentially violates the legal requirements based on an evaluation of these properties.

Let = fma• pp• tcg be a set of pre-/suf-fixes for marketing, privacy policy, and a terms and conditions checkbox, respectively. Also let *0* 2 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 (*0_pre_checked*): The corresponding checkbox is already ticked by default.
Forced checkbox (*0_forced*): It is required to tick the corresponding checkbox to successfully register. This is often indicated with asterisks on the registration forms.
#tying_*1*: There is only one checkbox asking for (tying) two or three consents together. Therefore, *1 2 f*ma_pp, ma_tc, pp_tc, ma_pp_tc*g*.
#forced_*2*: The website does not ask for consent to the privacy policy and/or terms and conditions, but assumes it through the registration process. Hence *2 2 f*pp, tc, pp_tc*g*.
#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.

2.2 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 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. 1):

(1) We collected a set of websites from Alexa's ranking (Section 2.2.1).
(2) We designed a website annotation procedure (Section 2.2.2).
(3) We had a group of six legally-trained annotators execute this procedure on the set of websites (Section 2.2.3).
(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 2.2.4).

2.2.1 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.

Table 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

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.

Based on the study by Chatzimpyrros et al. [10], 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 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 [50], which was working with the crawler used for the pre-filtering, provides similar analysis of the filtered websites. Its results are aligned with ours.

2.2.2 Annotation procedure. Every website was manually annotated with the legal properties described in Section 2.1.2. 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 they successfully registered to the

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

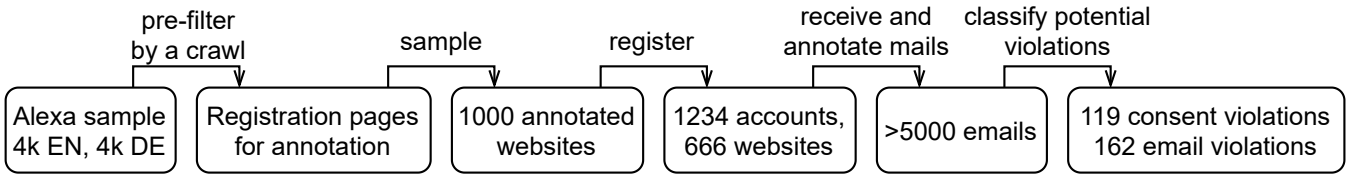

Figure 1: Overview of steps of our pilot study and results.

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 2.1.2. 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.

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.

2.2.3 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.

2.2.4 Resolving disagreements. Following empirical social science standards, every website was validated by a second independent annotator [24, 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 ^ [13]. Like a correlation, it takes values between -1 to 1, where ^ = 0 indicates the absence of agreement, ^ = 1 indicates perfect agreement, and ^ = 1 indicates perfect disagreement. For legal properties that were satisfied by at least 10% of the websites, the average ^ in our sample was 0.74. All the individual ^'s are given in Appendix A.2.

Our annotation procedure was more rigorous than those procedures used in most other related studies. For example, in Zimmeck et al. [78], 350 policies were labeled by two law students. Only 35 of them were doubly annotated and their Krippendorff's U 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 ^). In Bannihatti et al. [8], 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' ^, and its value was 0″7 (in this context, Fleiss' and Cohen's ^ are identical). To the best of our knowledge, the only other study with an annotation procedure as rigorous as ours is Wilson et al. [77], who used two law students to annotate 115 privacy policies with an average Krippendorff's U of 0″71, and had a third law student resolve any inconsistencies.

2.2.5 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. 16 in the Appendix. Fig. 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.

2.2.6 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.

Figure 2: The number of observed legal properties (de ned in Section 2.1.2) in the successful registrations.

2.3 Email annotation

We registered for websites using a real email address with a c-tional 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 con rming 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 di erent services, which serves as the baseline for this section. While each address received around ve 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 de ne 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, nding 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.

2.3.1 Marketing and servicing emails. In order to detect emails falling within the EU regulatory framework, we distinguish between marketing and servicing emails [64, 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 [33]. This case law was last affirmed by the BGH in 2018 [14]. 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 2.3.1. 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.

Figure 3: Email classification of the 5030 annotated emails, where we zoom into the marketing and servicing subclasses.

2.3.2 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 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.

2.3.3 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 [67].

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 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 [79]. 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.

Figure 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.

Due to the specific nature of these requirements in German law and the technical challenges associated with recognizing email components, we report violations related to missing email elements only within the dataset from the pilot study. Fig. 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 B.4.

3 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 3.1) and then the classification of legal properties (Section 3.2). We illustrate these steps and the associated statistics in Fig. 5.

3.1 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 email addresses. We therefore do not attempt to register using single sign-on, which was covered by other compliance studies [17].

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

3.1.1 Crawler implementation. To simulate users' browsing patterns, our crawler utilizes a real browser orchestrated by Selenium. Since existing frameworks such as OpenWPM [28] or webXray [58] 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 support Firefox as well.

To maximize the chances of successfully loading websites, we employ several techniques to evade bot detection, which we describe in Appendix C. 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. [55], 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.

3.1.2 Crawler navigation. After loading each website with a fresh cache, our crawler determines the page's language using the polyglot Python package. If 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 D).

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 [39] or lemmagen3 [49] 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 pages to visit 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

Figure 5: Overview of steps of our automated study and results.

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 type of a <form> tag. We apply the decision tree from Fig. 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 the Tranco 1M,[3] containing 426 forms. There were 12 contact, 32 login, 139 subscription, 163 registration, and 80 other forms. The procedure from Fig. 6 detected 74% of the registration forms and 94% of the subscription forms, with an overall accuracy of 82%.

3.1.3 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 3.1.2 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, ensuring that our fictitious credentials seem plausible. Most importantly, we generate a unique email address for every website.

[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 6: Crawler's form classification procedure.

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 is unchanged, usually indicating a `failed' registration; the form changes after a redirect, indicating a multi-step registration; and none of the above applies, which we denote as 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

[4]Checking all checkboxes hinders detecting the `marketing email despite user did not consent' violations.

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: 75% of them ReCaptcha v2, 20% ReCaptcha v3, 2% hCaptcha, and 3% image CAPTCHA.

Self-hosted mailserver. We self-host generated email addresses at sybilmail.de , con gured to only receive emails using the Mail Delivery Agent implemented with the Python Maildir library.

3.1.4 Registration confirmation. Once the crawler determines that the registration state is either `successful' or `unknown,' it waits for a con rmation email. As shown in [53], only 85% of 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 con rmation 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 ve 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, fteen minutes would only marginally improve this rate to 99.0%. The longer waiting time, however, comes at the expense of crawling time. Speci cally, waiting for ve minutes doubles the crawling time, while waiting for fteen minutes almost quadruples it. Emails with activation codes are typically prompt, so late email con rmation uses only activation links, not the codes.

Unfortunately, due to technical issues the independent con rmation 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 con rmation rates. This causes the presented results in Section 4 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.

3.1.5 Deployment. We evaluated our crawler by visiting the Tranco 1M list[5] [56], 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. [69] 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 with replacement, rather than crawling all the websites. For this reason, our results are only based on 660 202 unique domains, corresponding to the rst crawl.

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

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 12 IP addresses provided by the German Research Network, ensuring that the tra c originates in the EU. This, together with an EU address of our ctitious credentials, should indicate for the website that EU privacy laws may apply, which we further discuss in Section 6.

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.

3.1.6 Crawling results. Fig. 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 de ned in Section 3.1.3. The estimation indicates that 30.2% of form interactions were successful (51 290), 38.4% failed (65 220), and 31.4% resulted in an unde ned state (53 255).

The detection of the form submission's state is prone to false positives. Hence we manually investigated the correctness of the crawler determined registration state by inspecting 200 websites and testing the credentials used. The analysis revealed three newsletter subscriptions correctly deemed successful by the crawler and nine registrations, seven of which were correctly identi ed 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 as nearly half of the received emails resulted from newsletter subscriptions. Further observations from the manual analysis are presented in Appendix F.

We also analyze the results based on whether the websites are in the CrUX list. Note that Tranco 1M and CrUX have only a 51.9% overlap. The crawl was signi cantly more successful for the CrUX websites. Speci cally, 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, registration was detected as successful in 11.7% of cases (3.9% for non-CrUX websites). Our list choice supports a comparison with [19], relying on the DNS-based Alexa list with domains like WindowsUpdate.com without HTTP(S) endpoint. In the future, we recommend crawling the CrUX list to prevent unnecessary computations.

3.1.7 Ethical considerations. We have identi ed the following three risks of our study. 1) Legal risks arising from crawling: we reviewed various legal regimes and concluded that our research activities do not violate laws related to fraud, trespass, or breach of contract. This is underpinned by the fact that 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

Figure 7: Sankey plot of the crawler's intermediate results.

manual action by website owners, as the vast majority of emails are automated. In Section 4, 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 contracted with a third-party CAPTCHA solving service. Given the substantial prevalence of CAPTCHAs, accounting for one-third of our successful registrations, and their particular prevalence on higher-profit services, omitting CAPTCHA solving would introduce a significant bias. We carefully compared several providers, excluding those with evidently poor working conditions. Subsequently, we discussed the outsourcing with our university's legal department. Furthermore, we implemented multiple measures to avoid bot detection and, consequently, the need to solve CAPTCHAs. We try to reduce the use of CAPTCHA solving, by reusing the crawl results for subsequent studies and by transitioning to CAPTCHA solving by research assistants employed at our university for email confirmations.

3.2 Classifying legal properties

In this section, we automate the prediction of the legal properties defined in Section 2.1. Using the annotated datasets from Sections 2.2 and 2.3, 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.

3.2.1 Email features.

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 a TF-IDF encoding[6] that we fit to the training dataset, as well as a universal sentence encoder [9]. This pretrained transformer model maps sentences to an embedding in $\mathbb{R}^{512}$.

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 Training ML models for emails.

Given that our features correspond to tabular data, we use the XGBoost model [12]. 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,

[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 [1].

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 classi cation and `binary:logistic' for binary classi cation. 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 rst model predicts whether an email is a marketing email (i.e., newsletters, noti cations promoting service monetization, and surveys), a servicing double-opt-in email, or another kind of servicing email (con rmation 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. 8a, we present the confusion matrices of the mail-type model. This model achieves 97.7% balanced accuracy, while in the simpli ed task of deciding only whether email is marketing or servicing (aggregating double opt-ins with con rmations and legal updates), the balanced accuracy increases to 99.2%. The same balanced accuracy of 99.2% is achieved by the model predicting the presence of the unsubscribe options.

3.2.3 Form features. To transform forms of unlimited length to tabular features, we aggregate the form inputs by the crawler's keyword-based element classi cation. We group semantically similar inputs, such as the rst and last name, full name, and username, see Appendix E for details. We also reduce the complexity by excluding inputs irrelevant to legal classi cation, such as CAPTCHAs. From all inputs, we extract whether they are required or optional, and from checkboxes 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 process the form texts similarly as emails. Note that checkbox labels often consist of complex, nuanced statements, such as I don't want to receive special o ers 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 forms' input elds, 3154 TF-IDF columns, and ve sequences of $\mathbb{R}^{512}$ sentence embeddings.

3.2.4 Training ML models for forms. Similarly as with the email classi cation, we trained an XGBoost model for each of the 21 binary legal properties annotated by [53]. Note that the training dataset consists of only 666 annotated forms. To address this data scarcity, we also conducted experiments using the Tabnet model [8], 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 classi ed 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 2 presents the results of XGBoost with predictions based solely on the crawler's keyword-based classi cation 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 insu cient 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 rst place.

4 POTENTIAL VIOLATIONS

In this section, we describe our analysis of security threats and potential privacy violations concerning consent in forms and emails. For each method, we give context regarding related work and the EU privacy laws – 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 rst 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.

4.1 Security violations

Using our automated methods, we investigate websites' adherence to security best practices in private data protection as mandated by Article 32 of the GDPR. We focus on the personal information collection through user registration and newsletter sign-up processes. We present our ndings in Fig. 9.

4.1.1 Insecure registration form. GDPR Article 32(1)(a,b) mandates that the data controller implements appropriate technical measures to ensure the con dentiality of data processing. These measures should consider state-of-the-art methods that are economically viable. The widespread adoption of Let's Encrypt has signi cantly reduced the costs and technical hurdles associated with implementing encryption via HTTPS. Consequently, the adoption rate of this protocol has surpassed 95% [7]. To simplify our evaluation, we identify insecure registration forms by detecting forms that collect sensitive data over an 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.

(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 110k emails. More information on this dataset is in Appendix G.

Figure 8: Confusion matrices of mail type classi cation.

(a) Results from the manual pilot study.

(b) Results from the automated study.

Figure 9: Security threats of registration to websites.

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 con rmation code.

Our automated web crawler identi ed 5.2% of websites with forms collecting email addresses or passwords via unsecured HTTP connections. The higher prevalence compared to the manual study 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. [34]. 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. 24. 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. [75] found such violations on only 2.85% of websites. The disparity may arise from our more in-depth selection of forms for inspection and di erences in the crawling lists.

[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 6.

Table 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%	

4.1.2 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 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 [6].

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., [2, 35, 68]). 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.

4.2 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. 10a illustrates the opt-in requirements. Websites that engage in email marketing require this consent, which we have annotated using the `mail_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 2.1.1, 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. 10b.

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 2.1). 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.3.

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.

4.2.1 Manual pilot study results. We summarize the marketing consent violations in Fig. 11a. 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. 10a). 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. 10a). 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. 10b). 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. 10b).

4.2.2 Automated large-scale study results. In Fig. 11b, we present findings from the automated study using the decision trees outlined in Figs. 10a and 10b. Note that the baseline of reported incidence is 33 899 of websites that send any email.

Over 43% of registrations resulted in websites that never sent us any marketing emails, potentially influenced by issues with account activation (see Section 3.1.4), and up to 44% of the marketing emails we received resulted from newsletter subscriptions, reflecting the 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 then violate the GDPR consent requirements by not including a marketing checkbox, pre-checking the checkbox by default, or tying the checkbox with privacy policy or terms. In 2.0% cases, we received a marketing email despite rejecting consent, where the checkbox was neither pre-checked nor checked by the crawler. The crawler checked all checkboxes on additional 450 (1.3%) of websites.

It is important to acknowledge that the differences in violation statistics between the manual and automated studies. The statistically significant difference 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. The difference for `Email despite user did not consent' ($?$-value = 0.0086) can be attributed to the crawler checking all checkboxes on 1.3% of sites. While on many of these websites, the crawler truly consented to email marketing, on others may have been this detection a false positive, and hence directly reduced the observation of this violation.

4.3 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.

4.3.1 Double opt-in. In case of legal disputes, companies sending marketing emails must be able to demonstrate that recipients provided informed consent [40, 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 [57]. 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. However, there is a growing body of case law that considers proper double opt-in as a legal obligation. In a recent Austrian case [5], 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. 12a, 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 from Section 3.2.2, we classify the first email we receive from the website. The results presented in Fig. 12b show that 42.4% of websites adhere to the double-opt-in requirements and 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.

(a) Decision tree about opt-in validity based on legal properties.

(b) Decision tree about consent validity based on legal properties.

Figure 10: Decision trees about form interface violations based on legal properties. [1]

[1] Framing missing checkbox as `unspeci c' and tied checkbox as `unfree' consent is aligned to EDPB Guidelines 05/2020 [27] and rec. 43 of the GDPR. This context di ers from separation of processing purposes in (typically cookie notices) where tying multiple purposes to single checkbox is considered unspeci c [70, Sec. 5.3].

(a) Results from the manual pilot study.

(b) Results from the automated study.

Figure 11: Portion of senders that violate at least one marketing consent requirement.

(a) Annotated in the pilot study.

(b) Predicted in the automated study.

Figure 12: Classi cation of the rst email from the service.

Another reason likely includes the misclassi cation of our methods, as we discuss later in Section 5.

Although the double-opt-in procedure is not a legal requirement, its signi cance is ampli ed in the presence of registration crawlers like ours. Without this veri cation, our crawler could be exploited to subscribe arbitrarily email addresses to thousands of newsletters without the owners' consent, potentially leading to the Bomb attack [71].

4.3.2 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 2.4.1 [16]. Therefore, third parties sending marketing emails must be able to demonstrate that prior consent was obtained. This requires that users must be speci cally informed about whom their email address is shared with and for which marketing purposes [4]. Third parties must therefore be speci cally 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 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. 13a, 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 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 [7, 36, 63]. 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 the website declares this domain. We consider that domains match if the longest common subsequence between two domains is at least half of the shorter domain. This threshold of 0.5 was determined by empirical evaluation of a set of 200 domain matches, resulting in an accuracy of 91% with 2.5% of false negatives (wrongly predicting that domains are not similar) and 7.5% of 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. (2) The domain of the first received email. (3) Any common email host (e.g., gmail.com) 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 `Dis. email host.' (6) Domains in the privacy policy and terms and conditions, are marked as `In policy/terms.' (7) If all these checks fail, the domain is marked as `Undeclared.' We list other methods we considered for third-party sharing detection in Appendix H.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. 13b, 1.6% of our email addresses received emails from undeclared domains, including one website that shared our email address to 56 undeclared domains. Additionally, 0.1% of websites sent emails from domains that were only declared in the policy or terms, which are rarely read [71]. 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.

4.4 Summary and future work

In Fig. 14 we present the potential violations in emails from Section 2.3 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. [62] 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 more likely given that Englehardt et al. [22] 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. [71] 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.

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

(a) Annotated in the pilot study.

(b) Predicted in the automated study.

Figure 13: Observed types of email sharing to a third party.

Figure 14: 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.

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 B.2, we explore the violations depending on the website popularity. We present results only from the manual study, which have not found any statistically signi cant 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 ne website operators for such violations. Our crawler is able to collect contact emails addresses, which we employed in a noti cation study by Soldner [54]. Before we can start reporting the observations found by our automated methods, we have to establish the trustworthiness of them in the rst place, which we do in the next section.

5 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 misclassi cations. 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 lled out the form. Second, websites that sent us emails serve as a baseline for reporting violations. Among these 100 websites, our crawler submitted one contact, 54 subscription, and 45 registration forms. Our crawler misclassi ed six subscription forms as registration forms and one registration and contact form as subscription forms.

Out of the registrations or newsletter sign-ups, our crawler was unable to complete 25 double-opt-in procedures. Note that our evaluation of failed double-opt-ins is conservative since we classi ed any lack of email con rmation 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 con rmation should be improved in future work. Additionally, two registrations were incomplete, but the websites reminded us to nish the registration a behavior that was studied by Senol et al. [72]. Finally, the crawler successfully submitted the remaining 73 forms.

We examined the email opt-in violations and found that the rst emails from 83 websites were correctly classi ed. Unfortunately, the model misclassi ed that the rst email was for marketing rather than single- or double-opt-in in nine and ve cases, respectively. For a subsequent study described in Appendix G, we completed double-opt-ins manually, which allowed us to inspect 110k labeled emails, which we summarize in Fig. 8b. The comparison with Fig. 8a suggests that we tend to classify emails more rarely to be marketing compared to the annotators of the dataset we used for training [53].

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 misclassi cation of servicing emails for marketing, but the method correctly identi ed 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 di erent 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.

6 LIMITATIONS

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

6.1 Bias

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

The registration crawler introduces bias into the reported statistics of marketing consent violations. As explained in Section 5, our crawler exhibits greater success when signing up for simple websites and forms such as newsletters compared to complex registrations. However, it is possible that form complexity and website compliance are correlated. Hence, our results may not be representative of the entire population of websites visited by users.

To mitigate this limitation, we propose involving real users in part of the process. For example, semi-automated techniques can be employed for email con rmation, 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 similarly inspected.

6.2 Trustworthiness of violations

All of our ndings are prone to misclassi cation. Hence all violations should be regarded as potential violations. In particular, in cases where our methods exhibit low precision in identifying violations, caution should be exercised when using the results for enforcement purposes. We propose two complementary solutions to address this. First, one can carefully examine the evidence of the violation in the form of screenshots and website source code, similarly to our approach in Section 5. Moreover, a larger training dataset can be constructed by rectifying misclassi ed violations and adjusting the corresponding legal labels, thereby improving our models in the future. This is particularly crucial for properties with few positive samples, such as the pre-checked marketing checkbox.

Finally, our methods are not a complete audit as there may be additional unaddressed violations. Detecting email sharing might require a longer 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 3.1.6, the registration rate is favorable when crawling such lists. By utilizing these lists and considering additional factors, like 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.

6.3 Adversarial websites

ML models are susceptible to adversarial ML methods. Website operators could modify their forms, for example by including input elds or text labels invisible to users. We assume that websites do not engage in such practices, since we have not published our violation detection models, making it di cult for websites to exploit their weaknesses to evade detection. Moreover, our classi cation relies on both machine learning and the crawler's keyword-based form classi cation, 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. [11] to enhance the robustness of decision tree models against adversarial modi cations.

7 CONCLUSIONS AND FUTURE WORK

We have developed a crawler capable of conducting large-scale studies on the security and privacy of website registration. Our crawler more than doubles successful registrations of prior work, signing up to 5.9% of 660k websites. This led to the collection of over 2 million emails. Using this crawler, we were able to detect a wide range of security and privacy threats, fully automating previous manual studies and scaling them by orders of magnitude. To do so, we automated the prediction of complex legal properties of forms and emails using ML. We observed 12 605 websites, which is 37.2% of the websites sending us emails, containing at least one potential violation, or sending a marketing email as the rst email.

Our automation fosters various kinds of research. First, our crawler enables future work to analyze the security and privacy of authenticated sections, re ecting how real users browse websites. Second, the option to collect a large-scale dataset of forms and emails can foster research on communication practices. Examples include analyzing whether websites respect the unsubscribe action or studying whether tracking by third-parties is even more present in those parts of websites requiring authentication.

In future work, we will explore using our infrastructure for regulatory enforcement. Namely, by extending our training datasets, such as the annotation of emails of the subsequent crawl, we plan to enhance the predictive capabilities of our machine learning models in detecting violations. These enhanced methods can potentially help understa ed and under-resourced data protection authorities by pre- ltering non-compliant websites and collecting supporting evidence. This can foster e cient enforcement at scale and thereby improve security and privacy for users of the web.

ACKNOWLEDGMENTS

We thank Stefan Bechtold and Alexander Stremitzer for their guidance on the legal aspects of this work. Further, we thank David Bernhard, Patrice Kast[50], Luka Lodrant [59], and Vandit Sharma, for assistance developing the crawler and Joachim Posegga and the Deutsches Forschungsnetz for providing us with university VPN.

REFERENCES

[1] Stephen Akuma, Tyosar Lubem, and Isaac Terngu Adom. 2022. Comparing Bag of Words and TF-IDF with di erent models for hate speech detection from live tweets. International Journal of Information Technology 14, 7 (2022), 3629 3635. https://doi"org/10'1007/s41870-022-01096-4

[2] Fatma Al Maqbali and Chris J Mitchell. 2018. Web Password Recovery: A Necessary Evil?. In Proceedings of the Future Technologies Conference. Springer, Cham, 324 341. https://doi"org/10'1007/978-3-030-02683-7_23

[3] Sercan Ö. Arik and Tomas P ster. 2021. TabNet: Attentive Interpretable Tabular Learning. Proceedings of the AAAI Conference on Arti cial Intelligence 35, 8 (5 2021), 6679 6687. https://doi"org/10'1609/aaai"v35i8'16826

[4] Art. 29 Data Protection Working Party. 2004. Opinion 5/2004 on unsolicited communications for marketing purposes under Article 13 of Directive 2002/58/EC. https://ec"europa"eu/justice/article-29/documentation/opinion-recommendation/ les/2004/wp90_en"pdf

[5] Austrian Data Protection Authority (Datenschutzbehörde). 2019. DSB-D130.073/0008-DSB/2019. https://gdprhub"eu/index"php?title=DSB_-_DSB-D130'073/0008-DSB/2019.

[6] Baden-Württemberg Data Protection Authority (LfDI Baden-Württemberg). 2018. LfDI - O 1018/115. https://gdprhub"eu/index"php?title=LfDI_-_O_1018/115.

[7] Yannis Bakos, Florencia Marotta-Wurgler, and David R Trossen. 2014. Does anyone read the ne print? Consumer attention to standard-form contracts. The Journal of Legal Studies 43, 1 (2014), 1 35.

[8] Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, Florian Schaub, and Norman Sadeh. 2020. Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text. In Proceedings of The Web Conference 2020 (Taipei, Taiwan) (WWW '20). Association for Computing Machinery, New York, NY, USA, 1943 1954. https://doi"org/10'1145/3366423'3380262

[9] Daniel Cer, Yinfei Yang, Sheng yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. arXiv:1803.11175 [cs.CL]

[10] Manolis Chatzimpyrros, Konstantinos Solomos, and Sotiris Ioannidis. 2019. You Shall Not Register! Detecting Privacy Leaks Across Registration Forms. In Computer Security. Springer, Cham, 91 104.

[11] Hongge Chen, Huan Zhang, Duane Boning, and Cho-Jui Hsieh. 2019. Robust Decision Trees Against Adversarial Examples. In Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97). Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, Cambridge, MA, 1122 1131. https://proceedings"mlr"press/v97/chen19l"html

[12] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (San Francisco, California, USA) (KDD '16). Association for Computing Machinery, New York, NY, USA, 785 794. https://doi"org/10'1145/2939672'2939785

[13] Jacob Cohen. 1960. A coe cient of agreement for nominal scales. Educational and psychological measurement 20, 1 (1960), 37 46.

[14] Nurullah Demir, Matteo Groÿe-Kampmann, Tobias Urban, Christian Wressnegger, Thorsten Holz, and Norbert Pohlmann. 2022. Reproducibility and replicability of web measurement studies. In Proceedings of the ACM Web Conference 2022 (Virtual Event, Lyon, France) (WWW '22). Association for Computing Machinery, New York, NY, USA, 533 544. https://doi"org/10'1145/3485447'3512214

[15] Deutsche Bundestag. 2021. 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). https://www"gesetze-im-internet"de/uwg_2004/BJNR141400004"html

[16] Deutsche Bundestag. 2021. 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). https://www"gesetze-im-internet"de/tmg/BJNR017910007"html

[17] Yana Dimova, Tom Van Goethem, and Wouter Joosen. 2023. Everybody's Looking for SSOmething: A large-scale evaluation on the privacy of OAuth authentication on the web. Proceedings on Privacy Enhancing Technologies 2023, 3 (2023), 452 467. Issue 4. https://doi"org/10'56553/popets-2023-0119

[18] Directorate-General for the Information Society and Media (European Commission). 2015. ePrivacy Directive, assessment of transposition, e ectiveness and compatibility with the proposed data protection regulation. doi:10.2759/419180. https://op"europa"eu/en/publication-detail/-/publication/1529b684-d3d0-4445-bcdb-72f5cef237bc/language-en

[19] Kostas Drakonakis, Sotiris Ioannidis, and Jason Polakis. 2020. 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 (Virtual Event, USA) (CCS '20). Association for Computing Machinery, New York, NY, USA, 1953 1970. https://doi"org/10'1145/3372297'3417869

[20] Lilian Edwards. 2005. The New Legal Framework for E-Commerce in Europe. Hart Publishing, Portland, Oregon.

[21] Volker Emmerich and Knut Werner Lange. 2019. Unfair competition (Unlauterer Wettbewerb). C.H. Beck, Munich, DE.

[22] Steven Englehardt, Je rey Han, and Arvind Narayanan. 2018. I never signed up for this! Privacy implications of email tracking. Proceedings on Privacy Enhancing Technologies 2018 (2018), 109 126. Issue 1.

[23] Steven Englehardt and Arvind Narayanan. 2016. Online Tracking: A 1-Million-Site Measurement and Analysis. In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (Vienna, Austria) (CCS '16). Association for Computing Machinery, New York, NY, USA, 1388 1401. https://doi"org/10'1145/2976749'2978313

[24] Lee Epstein and Andrew D Martin. 2014. An introduction to empirical legal research. Oxford University Press, Oxford, UK.

[25] European Commission. 2016. Guidance on the implementation/application of Directive 2005/29/EC on Unfair Commercial Practices. https://eur-lex"europa"eu/legal-content/EN/TXT/HTML/?uri=CELEX:52016SC0163&from=EN

[26] European Data Protection Board. 2019. 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. https://edpb"europa"eu/our-work-tools/our-documents/styrelsens-yttrande-art-64/opinion-52019-interplay-between-eprivacy_en

[27] European Data Protection Board. 2020. 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.

[28] European Data Protection Board. 2020. Guidelines 05/2020 on consent under Regulation 2016/679 (GDPR). https://edpb"europa"eu/our-work-tools/our-documents/guidelines/guidelines-052020-consent-under-regulation-2016679_en

[29] European Parliament, Council of the European Union. 1995. 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. http://data"europa"eu/eli/dir/1995/46/oj

[30] European Parliament, Council of the European Union. 2000. 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'). http://data"europa"eu/eli/dir/2000/31/oj

[31] European Parliament, Council of the European Union. 2002. 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). , 31 pages. http://data"europa"eu/eli/dir/2002/58/oj

[32] European Parliament, Council of the European Union. 2005. 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'). http://data"europa"eu/eli/dir/2005/29/oj

[33] European Parliament, Council of the European Union. 2006. Directive 2006/114/EC of the European Parliament and of the Council of 12 December 2006 concerning misleading and comparative advertising. http://data"europa"eu/eli/dir/2006/114/oj

[34] Adrienne Porter Felt, Richard Barnes, April King, Chris Palmer, Chris Bentzel, and Parisa Tabriz. 2017. Measuring HTTPS adoption on the web. In 26th USENIX security symposium (USENIX security 17). USENIX Association, Vancouver, BC, Canada, 1323 1338.

[35] Nethanel Gelernter, Senia Kalma, Bar Magnezi, and Hen Porcilan. 2017. The Password Reset MitM Attack. In 2017 IEEE Symposium on Security and Privacy (SP). IEEE, San Jose, CA, USA, 251 267. https://doi.org/10.1109/SP.2017.79

[36] Joshua Gluck, Florian Schaub, Amy Friedman, Hana Habib, Norman Sadeh, Lorrie Faith Cranor, and Yuvraj Agarwal. 2016. How short is too short? Implications of length and framing on the e ectiveness of privacy notices. In Twelfth Symposium on Usable Privacy and Security (SOUPS 2016). USENIX Association, Denver, CO, USA, 321 340.

[37] Google Transparency Report Authors. 2023. HTTPS encryption on the web. https://transparencyreport.google.com/https/overview.

[38] Maia Hamin. 2018. "don't ignore this:" Automating the Collection and Analysis of Campaign Emails. Technical Report. Princeton University.

[39] Matthew Honnibal, Ines Montani, So e Van Landeghem, and Adriane Boyd. 2023. spaCy: Industrial-strength Natural Language Processing in Python. https://doi.org/10.5281/zenodo.7970450

[40] Dietmar Jahnel. 2021. Legal commentary on the General Data Protection Regulation (GDPR) (Kommentar zur Datenschutz-Grundverordnung (DSGVO)), Art. 7 Conditions for consent (Bedingungen für die Einwilligung). Jan Sramek Verlag, Vienna, AT.

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

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

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

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

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

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

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

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

[49] Matjaz Jur²ic, Igor Mozetic, Tomaz Erjavec, and Nada Lavrac. 2010. Lemmagen: Multilingual lemmatisation with induced ripple-down rules. Journal of Universal Computer Science 16, 9 (2010), 1190 1214.

[50] Patrice Kast. 2021. Automating website registration for GDPR compliance analysis. Bachelor's thesis. ETH Zurich.

[51] Karel Kubicek. 2024. Automated Analysis and Enforcement of Consent Compliance. Ph. D. Dissertation. ETH Zurich. https://doi.org/10.3929/ethz-b-000662039

[52] Karel Kubicek, Jakob Merane, Ahmed Bouhoula, and David Basin. 2024. Automating Website Registration for Studying GDPR Compliance. In Proceedings of the ACM Web Conference 2024 (WWW '24). ACM, New York, NY, USA, 12. https://doi.org/10.1145/3589334.3645709

[53] Karel Kubicek, Jakob Merane, Carlos Cotrini, Alexander Stremitzer, Stefan Bechtold, and David Basin. 2022. Checking Websites' GDPR Consent Compliance for Marketing Emails. Proceedings on Privacy Enhancing Technologies 2022 (2022), 282 303. Issue 2. https://doi.org/10.2478/popets-2022-0046

[54] Laura-Vanessa Laura-Vanessa. 2023. Quantifying Mechanisms behind Cookie Consent (Non-)Compliance: A Noti cation Study of Audit Tools. Bachelor's thesis. ETH Zurich.

[55] Victor Le Pochat, Tom Van Goethem, and Wouter Joosen. 2019. Evaluating the Long-Term E ects of Parameters on the Characteristics of the Tranco Top Sites Ranking. In Proceedings of the 12th USENIX Conference on Cyber Security Experimentation and Test (Santa Clara, CA, USA) (CSET'19). USENIX Association, USA, 10.

[56] Victor Le Pochat, Tom Van Goethem, Samaneh Tajalizadehkhoob, Maciej Korczy«ski, and Wouter Joosen. 2019. Tranco: A Research-Oriented Top Sites Ranking Hardened Against Manipulation. In Proceedings of the 26th Annual Network and Distributed System Security Symposium (NDSS 2019). Internet Society, San Diego, CA, USA, 1 15. https://doi.org/10.14722/ndss.2019.23386

[57] Legal team of the Certi ed Senders Alliance. 2017. DOI: if not now, then when?! https://certi ed-senders.org/blog/doi-if-not-now-then-when/.

[58] Timothy Libert. 2015. Exposing the Hidden Web: An Analysis of Third-Party HTTP Requests on One Million Websites. International Journal of Communication 9 (2015), 3544 3561.

[59] Luka Lodrant. 2022. Designing a generic web forms crawler to enable legal compliance analysis of authentication sections. Master's thesis. ETH Zurich. https://doi.org/10.3929/ethz-b-000534764

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

[61] Arunesh Mathur, Mihir Kshirsagar, and Jonathan Mayer. 2021. 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 (Yokohama, Japan) (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 360, 18 pages. https://doi.org/10.1145/3411764.3445610

[62] Arunesh Mathur, Angelina Wang, Carsten Schwemmer, Maia Hamin, Brandon M. Stewart, and Arvind Narayanan. 2020. Manipulative tactics are the norm in political emails: Evidence from 100K emails from the 2020 U.S. election cycle. https://electionemails2020.org.

[63] Aleecia M McDonald and Lorrie Faith Cranor. 2008. The cost of reading privacy policies. ISJLP 4 (2008), 543.

[64] Daniela Mederle. 2010. The regulation of spam and unsolicited commercial emails (Die Regulierung von Spam und unerbetenen kommerziellen E-Mails). Hanns, Cologne, DE.

[65] Abraham Mhaidli, Selin Fidan, An Doan, Gina Herakovic, Mukund Srinath, Lee Matheson, Shomir Wilson, and Florian Schaub. 2023. Researchers' Experiences in Analyzing Privacy Policies: Challenges and Opportunities. Proceedings on Privacy Enhancing Technologies 2023 (2023), 287 305. Issue 4. https://doi.org/10.56553/popets-2023-0111

[66] Hans Micklitz and Martin Schirmbacher. 2019. Legal commentary on the German Act against Unfair Competition (Kommentar zum Gesetz gegen den unlauteren Wettbewerb (UWG)), Ÿ 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. Au . 2019).

[67] Hans Micklitz and Martin Schirmbacher. 2019. Legal commentary on the German Telemedia Act (Kommentar zum Telemediengesetz (TMG)), Ÿ 4-6 TMG, in G. Spindler and F. Schuster, Electronic Media Law, 4th edition 2019, (Recht der elektronischen Medien, 4. Au . 2019).

[68] Caleb Routh, Brandon DeCrescenzo, and Swapnoneel Roy. 2018. Attacks and vulnerability analysis of e-mail as a password reset point. In 2018 Fourth International Conference on Mobile and Secure Services (MobiSecServ). IEEE, Miami Beach, FL, USA, 1 5.

[69] Kimberly Ruth, Deepak Kumar, Brandon Wang, Luke Valenta, and Zakir Durumeric. 2022. Toppling Top Lists: Evaluating the Accuracy of Popular Website Lists. In Proceedings of the 22nd ACM Internet Measurement Conference (Nice, France) (IMC '22). Association for Computing Machinery, New York, NY, USA, 374 387. https://doi.org/10.1145/3517745.3561444

[70] Cristiana Santos, Arianna Rossi, Lorena Sanchez Chamorro, Kerstin Bongard-Blanchy, and Ruba Abu-Salma. 2021. 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 (Virtual Event, Republic of Korea) (WPES '21). ACM, Virtual Event Republic of Korea, 187 194. https://doi.org/10.1145/3463676.3485611

[71] Markus Schneider, Haya Shulman, Adi Sidis, Ravid Sidis, and Michael Waidner. 2020. Diving into Email Bomb Attack. In 2020 50th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN). IEEE, Valencia, Spain, 286 293. https://doi.org/10.1109/DSN48063.2020.00045

[72] Asuman Senol, Gunes Acar, Mathias Humbert, and Frederik Zuiderveen Borgesius. 2022. Leaky Forms: A Study of Email and Password Ex ltration Before Form Submission. In 31st USENIX Security Symposium (USENIX Security 22). USENIX Association, Boston, MA, USA, 1813 1830. https://www.usenix.org/conference/usenixsecurity22/presentation/senol

[73] Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Physical therapy 85, 3 (2005), 257 268.

[74] The Direct Marketing Association. 2019. Marketing email tracker 2019. https://dma.org.uk/uploads/misc/marketers-email-tracker-2019.pdf.

[75] Christine Utz, Matthias Michels, Martin Degeling, Ninja Marnau, and Ben Stock. 2023. Comparing large-scale privacy and security noti cations. Proceedings on Privacy Enhancing Technologies 2023 (2023), 173 193. Issue 3. https://doi.org/10.56553/popets-2023-0076

[76] Jan Weiser. 2018. 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, 10 (2018), 291.

[77] Shomir Wilson, Florian Schaub, Aswarth Abhilash Dara, Frederick Liu, Sushain Cherivirala, Pedro Giovanni Leon, Mads Schaarup Andersen, Sebastian Zimmeck, Kanthashree Mysore Sathyendra, N. Cameron Russell, Thomas B. Norton, Eduard Hovy, Joel Reidenberg, and Norman Sadeh. 2016. The Creation and Analysis of a Website Privacy Policy Corpus. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Katrin Erk and Noah A. Smith (Eds.). Association for Computational Linguistics, Berlin, Germany, 1330 1340. https://doi.org/10.18653/v1/P16-1126

[78] Sebastian Zimmeck, Peter Story, Daniel Smullen, Abhilasha Ravichander, Ziqi Wang, Joel Reidenberg, N Cameron Russell, and Norman Sadeh. 2019. MAPS:

Scaling privacy compliance analysis to a million apps. Proceedings on Privacy Enhancing Technologies 2019, 3 (2019), 66–86.
[79] Kerstin A. Zscherpe. 2008. 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, 9 (2008), 327–322.

A 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.0.1 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 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 15. 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.

Figure 15: 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.2 Inter-annotator agreement

Sim et al. [73] 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 3. Therefore, we also present the contingency tables for every legal property in Tables 4 and 5.

A.3 Linkage to dark patterns

In this section, we compare our defined potential violation types to the taxonomy of dark patterns by Marthur et al. [61]. We refer to terms from [61] 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 *covert* and *information hiding*

Table 3: The individual Cohen's ^s of legal properties. Note that ^ = 1 implies full agreement, while ^ = 1 implies full disagreement.

Checkbox	^	Hashtag	^
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 4: Contingency tables of checkbox values. Rows represent the rst annotation, the second annotation is depicted by the column.

(a) mark_consent

	True	False
True	244	42
False	51	663

(b) ma_checkbox

	True	False
True	192	41
False	41	726

(c) ma_pre_checked

	True	False
True	32	10
False	8	950

(d) ma_forced

	True	False
True	10	7
False	10	970

(e) mark_purpose

	True	False
True	34	15
False	25	926

(f) pp_checkbox

	True	False
True	187	40
False	40	733

(g) pp_pre_checked

	True	False
True	2	4
False	1	993

(h) pp_forced

	True	False
True	169	42
False	43	746

(i) tc_checkbox

	True	False
True	165	37
False	34	764

(j) tc_pre_checked

	True	False
True	6	3
False	1	990

(k) tc_forced

	True	False
True	143	40
False	42	775

Table 5: Contingency tables of hashtag values. Rows represent the rst annotation, the second annotation is depicted by the column.

(a) #tying12

	True	False
True	1	7
False	7	985

(b) #tying13

	True	False
True	0	0
False	0	1000

(c) #tying23

	True	False
True	86	26
False	25	863

(d) #tying123

	True	False
True	0	4
False	1	995

(e) #forcedpp

	True	False
True	131	46
False	41	782

(f) #forcedtc

	True	False
True	22	18
False	14	946

(g) #forcedpptc

	True	False
True	100	30
False	25	845

(h) #hidden

	True	False
True	4	32
False	31	933

(i) #age

	True	False
True	63	26
False	41	870

(j) #settings

	True	False
True	0	4
False	2	994

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-noti cations, 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 noti cations, and hence the emails are not user-triggered, which is deceptive. When an email is missing the unsubscribe option, it is restrictive

A.4 Registered accounts

Annotators registered to the selected 1000 websites in both annotating rounds. Each of the rounds resulted in a di erent number of successful registrations, namely 576 in the rst 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 di erence is caused by 34 websites that were inaccessible during one of the rounds and di erences in how the annotators browse the website to nd the registration form.

Note that if we would have to split the registration and annotation processes, we would lose signi cant 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 be resolved by resolving annotation. Moreover, we would not have subscribed to many of the 701 websites.

A.5 Email address generation

We considered two options for generating emails: setting up a custom email server or using Gmail + su xes. 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 di ers from [38] that used Gmail su xes.

B DATASETS CONTENT

We now elaborate on our dataset described in Section 2.2, 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.

B.1 Successful form annotations

In Figure 16, 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 17 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.

B.2 Email classi cation

In Appendix B.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 ?-value of the two proportions

Figure 16: Registration state of the resolved annotations. For agreement, Cohen's ^ for distinction between successful and failed registrations is 0.64.

Figure 17: 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 rst row, second column, marks how many websites with marketing consent (row label) have the marketing purpose (column label).

Z-Test of the rankŸ 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?value of 0.054.

B.3 Third party email sharing

As we stated in Section 4.3.2, we classi ed four websites as `other.' In the rst 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 o ine 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.

Figure 18: Summary of all potential violations of this study and the split into popularity groups by rank.

B.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 in uencing 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 19), the observations suggest that trending topics are used to improve marketing campaigns, but they do not generate new newsletter tra c. This hypothesis is based on the fact that

IP address: As we study the impact of the EU's privacy regulations, we focused our data collection on tra c 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. [14], residential proxies are the least likely to be detected as bot tra c, 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 ($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.

Figure 19: Classi cation 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 (1"2 number of accounts), while number of emails linearly increases over time (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.

during the limited period of the study, we did not observe any spikes in the number of newsletters during these periods. However, to con rm this hypothesis, we would need a more longitudinal study.

C BOT-EVASION TECHNIQUES

Given the experience from bot detection with our annotation tool Appendix A.1, we implemented the following methods to further decrease the chance of our crawling being detected as a bot activity.

Browser: We use Undetected Chromedriver,[9] which extends the usual Chromedriver with numerous bot evasion techniques, such as removing ngerprints 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 ngerprinting. The proxy and browser di er in the ciphersuite, which is inspected by modern bot detection systems like Cloud are. 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 agged as a bot by Cloud are.

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

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

D SUPPORTED LANGUAGES

Our crawler supports 37 languages, with most of the keywords being translated by native or pro cient speakers of the language, whom we instructed in observing multiple registration or newsletter-subscription 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. Note that the LibreTranslate support is constantly improving, both in the terms of translation quality and the number of supported languages, which rose to 28 by the nal version of this thesis.

We are aware of the following limitation of the machine translation. First, nuances in the form or email text might be lost. Second, as the training data is in German and English, the models should re ect well the websites in these languages, yet their performance can drop on websites in a language absent in the training data. Finally, the ePrivacy Directive implementation is not absolutely consistent among EU countries, the impact of such inconsistencies remains a limitation. We believe that the generalization of our methods to so many languages, even if constrained by the machine translation quality, is an important contribution of our work in the context of understudied non-English websites [65, Sec. 4.4.2].

E CRAWLER FORM CLASSIFICATION

For form classi cation, we use aggregated form features, features for speci c input types that we order to provide stable tabular features' ordering, and features extracted from speci c parts of text in the form. The aggregated features include the number of inputs in total, the number of the <input>, <textarea>, <select>, and <button> tags. Our crawler distinguishes various form inputs by their semantics, which we aggregate into the following groups.

mail
password
phone
username
names: rst, middle, last or full name

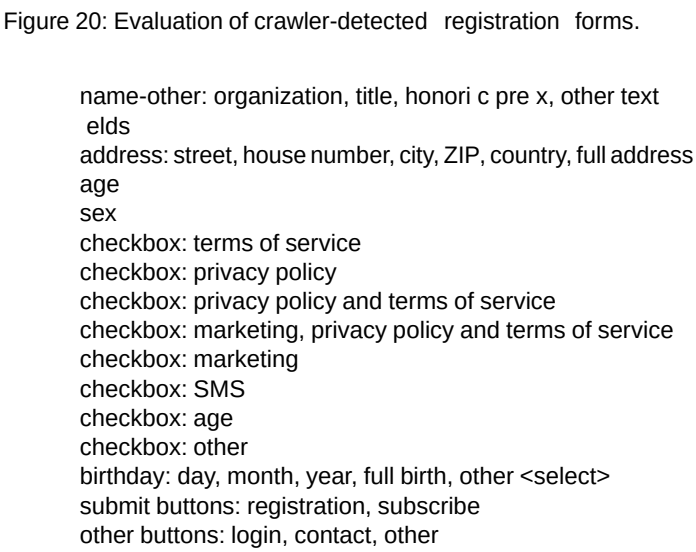

Figure 20: Evaluation of crawler-detected registration forms.

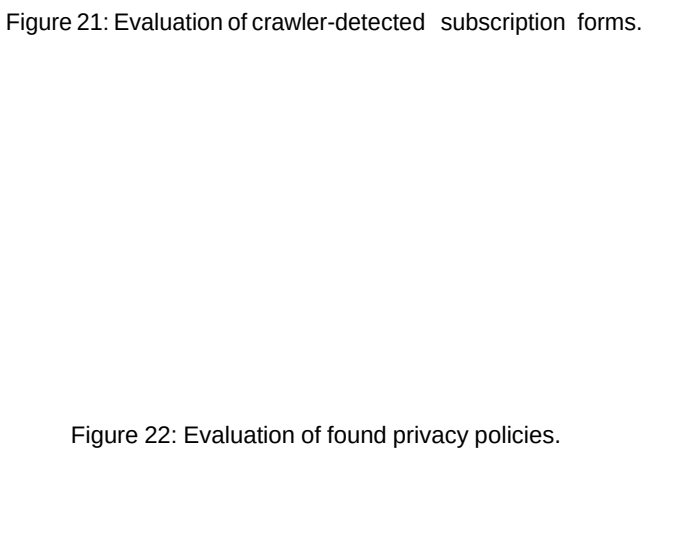

Figure 21: Evaluation of crawler-detected subscription forms.

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

For each of these groups, features correspond to the number of inputs in the group, whether any of these inputs is required, the default values, i.e., text for text input or Boolean for checkboxes and radio buttons, and the text of the closes label. The texts are then processed by the TD-IDF model, with a vocabulary size of 50. In the case of checkboxes, the submission button, and the entire aggregated text of the form, the text is processed by the TF-IDF model with a 500 words vocabulary and embeddings are extracted using the universal sentence encoder [9].

Figure 22: Evaluation of found privacy policies.

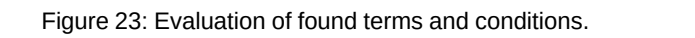

Figure 23: Evaluation of found terms and conditions.

F 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, the analysis presented below pertains to the remaining 181 websites.

In Fig. 20, 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. 21 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.

Figs. 22 and 23 illustrate the evaluation of the detected policies and terms and conditions on a list of 300 websites. Our manual evaluation showed that almost 75% and 65% of websites contain privacy policies 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.