Karel Kubíček*, Jakob Merane*, Carlos Cotrini, Alexander Stremitzer, Stefan Bechtold, and David Basin

Checking Websites' GDPR Consent Compliance for Marketing Emails

Abstract: The sending of marketing emails is regulated to protect users from unsolicited emails. For instance, the European Union's ePrivacy Directive states that marketers must obtain users' prior consent, and the General Data Protection Regulation (GDPR) specifies further that such consent must be freely given, specific, informed, and unambiguous.

Based on these requirements, we design a labeling of legal characteristics for websites and emails. This leads to a simple decision procedure that detects potential legal violations. Using our procedure, we evaluated 1000 websites and the 5000 emails resulting from registering to these websites. Both datasets and evaluations are available upon request. We find that 21.9% of the websites contain potential violations of privacy and unfair competition rules, either in the registration process (17.3%) or email communication (17.7%). We demonstrate with a statistical analysis the possibility of automatically detecting such potential violations.

Keywords: marketing email, website registration, ePrivacy Directive, GDPR, consent, compliance, privacy

DOI 10.2478/popets-2022-0046
Received 2021-08-31; revised 2021-12-15; accepted 2021-12-16.

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 advertising their products and services [51]. This misuse, along with the sheer number of users' online accounts, leaves users regularly overwhelmed with unsolicited marketing emails. Users often have no idea why they received a particular marketing email and from where the sender obtained their email addresses. This is both tiresome and upsetting. Indeed, one of the most common reasons for users unsubscribing from marketing emails is that the recipients do not remember ever registering for this service [50].

To counteract unsolicited email advertising, regulations on privacy and unfair competition have come into force. As early as 2002, the European Union adopted the ePrivacy Directive that established the requirement of users' prior consent for sending marketing emails. Furthermore, the General Data Protection Regulation (GDPR), another landmark European privacy law, was adopted in 2016. It provides a precise notion of consent.

In this paper, we analyze how well websites sending marketing emails comply with legal requirements. We conduct the first wide-scale study of website registration forms whose target includes EU citizens. We focus on the following three legal aspects of email marketing. First, we study how registration forms and emails ask for consent to marketing emails. We observe 17.3% of websites sending marketing emails after potentially violating at least one of the GDPR's consent requirements. In addition, only 59% of websites confirm that the address is correct by sending an activation email (*double opt-in*). Second, we analyze the content of the emails sent by these websites. We find that 16% of websites do not provide the user an unsubscribe method or legal notice. Moreover, 2.3% of websites disclose user-provided passwords directly in the email, risking the security of users who reuse passwords. Lastly, we detect that 4.1% of websites share users' email addresses with third parties. We elaborate on this by analyzing whether websites disclose this practice.

For our analysis, we manually annotated 1000 websites, documenting 21 legal properties, for example, "The registration form contains a checkbox for consenting to marketing emails" (see Section 2.2). We successfully registered for 666 of these websites. After registering for these websites, we received over 5000 emails from them that we also annotated with purpose, namely

***Corresponding Author: Karel Kubíček:** ETH Zurich, E-mail: karel.kubicek@inf.ethz.ch
***Corresponding Author: Jakob Merane:** ETH Zurich, E-mail: jakob.merane@gess.ethz.ch
Carlos Cotrini: ETH Zurich, E-mail: ccarlos@inf.ethz.ch
Alexander Stremitzer: ETH Zurich, E-mail: astremitzer@ethz.ch
Stefan Bechtold: ETH Zurich, E-mail: sbechtold@ethz.ch
David Basin: ETH Zurich, E-mail: basin@inf.ethz.ch

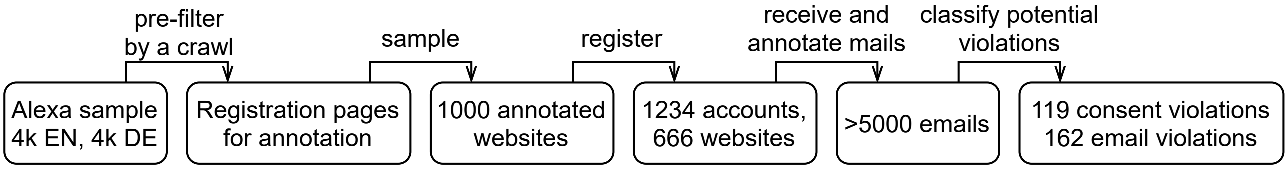

pre-filter by a crawl
sample
register
receive and annotate mails
classify potential violations
Alexa sample 4k EN, 4k DE
Registration pages for annotation
1000 annotated websites
1234 accounts, 666 websites
>5000 emails
119 consent violations 162 email violations

Fig. 1. *Overview of the process involved in our study and the (intermediate) results.*

"marketing" or "servicing." Based on the legal properties and presence of marketing emails, we defined a decision procedure for detecting potential violations. We observed at least one potential violation in 148 (22.2%) of these websites.

Previous studies examined the privacy threats posed by specific kinds of marketing campaigns. Hamin and Mathur et al. [31, 54] studied US political campaign emails. They observed malicious practices with regard to both email content and the handling of personal information, which was often shared with third parties. Engelhart et al. [18] analyzed user tracking by marketing emails and observed that over 70% of emails contain trackers. Other studies [52, 60] reported how websites use dark patterns to trick users into consenting to cookie policies. In contrast, this is the first study to systematically analyze the extent to which companies sending marketing emails comply with the GDPR's notion of consent.

Terminology

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

Contributions

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

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

Annotated datasets. We offer the privacy research community a dataset with annotations of the legal properties of registration forms for 1000 websites. We release these annotations, the registration page source code, and post-processed features of the registration form upon request. We also release a dataset of 5000 emails labeled with their purpose. Both datasets are suitable for other studies, such as email or registration form tracking analysis [8, 18], or marketing email content analysis.[1]

Feature analysis. We conduct a statistical analysis to identify which features are most influential when deciding potential violations. We illustrate how these features can simplify manual compliance analysis.

Organization

We review the legal requirements for email marketing in Section 2. We then describe the registration process and the content of the dataset of annotated websites in Section 3. In Section 4, we present legal requirements on email content and we report on potential violations of these requirements. Afterward, we undertake a legal analysis of both datasets in Section 5 and we present the first steps towards automating such analysis in Section 6. Finally, we consider related work, draw conclusions, and propose future steps.

[1] The datasets, intermediate results, and other materials are available on request at https://forms.gle/dTGpfs5vKqdLz8sQ7. A page with an overview of this study is at https://karelkubicek.github.io/post/reg-pets.

2 Legal requirements for email marketing in the EU

Privacy legislation has been recently introduced in many parts of the world aiming to strengthen consumer rights and privacy in the digital era. In Europe, the specific rules that relate to marketing emails consist of a complex interplay of European and national law. However, an essential pillar of legislative efforts against unsolicited marketing emails was the adoption of an *opt-in* requirement, whereby marketing emails are prohibited in the absence of prior consent [57, p. 79]. While some member states like Germany adapted such a regime early on [17, p. 168], the ePrivacy Directive [25] has established the opt-in requirement in July 2002 at European level [16, p. 46]. In particular, Art. 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) [11], 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 Art. 13(2) ePrivacy Directive and § 7(3) UWG. The exception implies that a product or service was provided for money [49]. Although controversially discussed, providing personal data as payment for "free" services is insufficient to generally trigger the exception ([66] in discussion of [42]). To protect customers from unsolicited commercial communications, legal scholars and German courts have tended to interpret the exception strictly [58]. 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 [24] or the German Telemedia Act (TMG) as the corresponding national implementation [12]. Furthermore, the EU's Directive on Unfair Commercial Practices (UCPD) [26] specifically bans persistent and unwanted solicitations by email [26, 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 Art. 13 of the ePrivacy Directive because these sector-specific provisions prevail over the UCPD [20, p. 90].

We selected the German implementation of the ePrivacy Directive as Germany is the largest economy in Europe. It is worth noting that Art. 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 [14, p. 10].

2.1 Valid consent under the GDPR

The interplay between the ePrivacy Directive, the UWG, and the GDPR is complex [21], 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 [23]. 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 [41]. The European Court of Justice also agreed with this view in the underlying preliminary ruling [36].

In general, Art. 4(11) and Art. 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, Art. 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) [22].

2.1.1 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 Art. 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 [33, par. 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 [39]. 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 [37]. The mere acceptance of general terms and conditions or privacy policies is also deemed insufficient [22, par. 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 [35, 36]. 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 fulfil this requirement.

2.2 Legal taxonomy

To operationalize the legal requirements of free, specific, and unambiguous consent, we have developed a legal taxonomy. We have tested the taxonomy in an exploratory pilot (see Appendix A.1.1 for more information about the pilot study), and refined the legal properties accordingly. In Section 5, we present a decision method that determines whether a website potentially violates the legal requirements based on an evaluation of these properties.

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

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

Marketing purpose (ma_purpose): Registering with the website is only, or mainly, for receiving marketing emails.

Marketing checkbox (ma_checkbox): There is a checkbox that the user must tick to give consent for marketing emails.

Privacy policy checkbox (pp_checkbox): There is a checkbox for consent for the website's privacy policy.

Terms and conditions checkbox (tc_checkbox): There is a checkbox for consent for the website's terms and conditions.

Pre-checked checkbox (*a*_pre_checked): The corresponding checkbox is already ticked by default.

Forced checkbox (*a*_forced): It is required to tick the corresponding checkbox to successfully register. This is often indicated with asterisks on the registration forms.

#tying_*b*: There is only one checkbox asking for (tying) two or three consents together. Therefore, $b \in \{\mathrm{ma\_pp}, \mathrm{ma\_tc}, \mathrm{pp\_tc}, \mathrm{ma\_pp\_tc}\}$.

#forced_*c*: The website does not ask for consent to the privacy policy and/or terms and conditions, but assumes it through the registration process. Hence $c \in \{\mathrm{pp}, \mathrm{tc}, \mathrm{pp\_tc}\}$.

#settings: Refusing consent requires more clicks, therefore the consent is assumed by default.

#age: The user's age or the date of birth are required for registration.

#colortrick: The colors on the website nudge the user to consent. For example, giving consent is highlighted with green, while refusing it is red.

#hidden: The declaration of consent can be easily missed by users.

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

3 Website dataset

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 Figure 1):

1. We collected a set of websites from Alexa's ranking (Section 3.1).
2. We designed a website annotation procedure (Section 3.2).
3. We had a group of six legally-trained annotators execute this procedure on the set of websites (Section 3.3).
4. We had each website annotated a second time by a second annotator. This allowed us to measure the annotators' consistency. Any conflicts were subsequently resolved by a third annotator. (Section 3.4).

3.1 Website collection

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

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

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

Based on the study by Chatzimpyrros et al. [8], 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.

Table 1. Website selection process.

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

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),[2] 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 thesis of Kast [43], which was working with the crawler used for the pre-filtering, provides similar analysis of the filtered websites. Its results are aligned with ours.

3.2 Annotation procedure

Every website was manually annotated with the legal properties described in Section 2.2. 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 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

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

a graphical interface for recording the legal properties, according to the legal taxonomy defined in Section 2.2. 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.1.2.

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

3.3 Annotators

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

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

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

3.4 Resolving disagreements

Following empirical social science standards, every website was validated by a second independent annotator [19, p. 114]. The second annotator was randomly chosen for every website and was different from the first annotator, but from the same group of six annotators. We observed only a single website that changed the registration form by the time the second annotator annotated the website, so website modifications were not a significant source of inter-annotator disagreement.

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

We measured the agreement between annotators with Cohen's κ [9]. Like a correlation, it takes values between -1 to 1, where $\kappa = 0$ indicates the absence of agreement, $\kappa = 1$ indicates perfect agreement, and $\kappa = -1$ indicates perfect disagreement. For legal properties that were satisfied by at least 10% of the websites, the average κ in our sample was 0.74. All the individual κ's are given in Appendix A.1.3.

Our annotation procedure was more rigorous than those procedures used in most other related studies. For example, in Zimmeck et al. [68], 350 policies were labeled by two law students. Only 35 of them were doubly annotated and their Krippendorff's α was 0.78 (text labeling requires this metric for inter-annotator agreement, but it has the same range and a similar interpretation as Cohen's κ). In Bannihatti et al. [44], a law student labeled 2692 opt-out statements from privacy policies. Only a subsample (50) was labeled independently by two additional annotators. The inter-annotator agreement was measured with Fleiss' κ, and its value was 0.7 (in this context, Fleiss' and Cohen's κ are identical). To the best of our knowledge, the only other study with an annotation procedure as rigorous as ours is Wilson et al. [67], who used two law students to annotate 115 privacy policies with an average Krippendorff's α of 0.71, and had a third law student resolve any inconsistencies.

3.5 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. 11 in the Appendix. Figure 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.

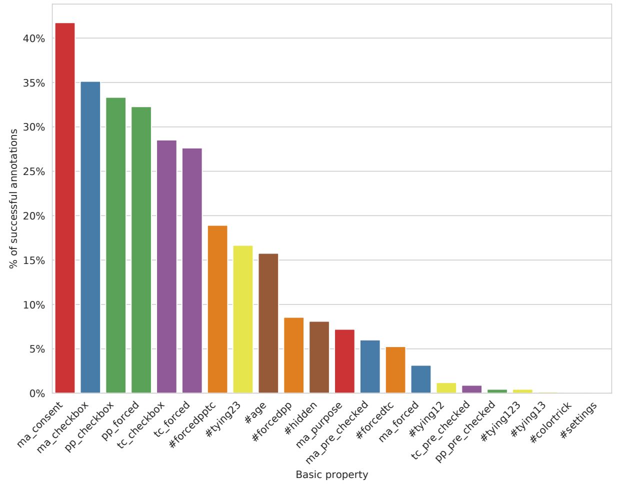


Fig. 2. *The number of observed legal properties (defined in Section 2.2) in the successful registrations.*

3.6 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.

4 Potential violations in the email dataset

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

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

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

1. We define marketing and servicing emails and show their distribution in our dataset.
2. We present the *double opt-in* procedure and report that fewer than 60% of websites follow this best practice.
3. We check the content of servicing emails for passwords in plaintext, finding 2.3% of websites send the user-provided password in plaintext via email.
4. We check the content of marketing emails for unsubscribe options and legal notices, observing that 16% of websites do not meet at least one requirement.
5. We check whether companies share the registered email address to third parties, finding that 4.1% of our addresses receive emails from multiple senders.

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

4.1 Marketing and servicing emails

In order to detect emails falling within the EU regulatory framework, we distinguish between marketing and servicing emails [57, 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 Art. 2(a) of the EU's Directive on misleading and comparative advertising [27]. This case law was last affirmed by the BGH in 2018 [38]. 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 Figure 3. 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.

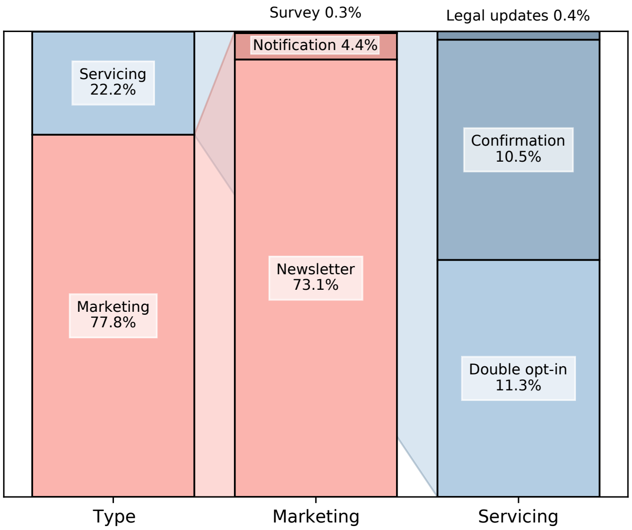


Fig. 3. *Email classification of the 5030 annotated emails, where we zoom into the marketing and servicing subclasses.*

4.2 Double opt-in

In case of legal disputes, the company that sends marketing emails must be able to demonstrate that the recipient knowingly consented [33, par. 6]. For this purpose, the *double opt-in* has been established as a best practice procedure that is not legally obligatory, but highly recommended by legal scholars and the marketing industry [45]. Alternative procedures, such as requiring users to send the service an email before registration, are not widely used. Such procedures can only be partially automated by mailto links, which would harm the usability of the registration procedure.

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 email address and can be implemented in various ways. The email contains either unique information (an activation link or a one-time password or code), or requires the user to ask for account activation by sending an email, which is used by less than 0.5% of the websites where we registered. Marketing emails can only be sent after consent is obtained using the previous actions. In contrast to a *single opt-in*, this procedure prevents users from registering, accidentally or maliciously, with an email address for an account 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.

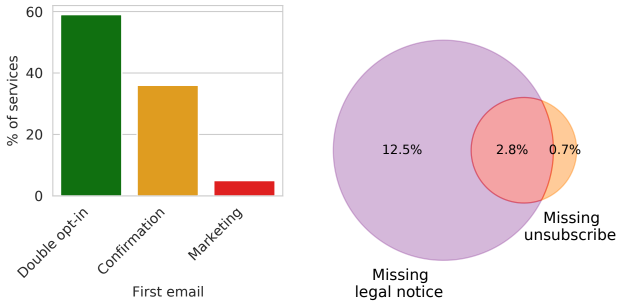
For the purpose of this study, we conservatively classify services that only provide single opt-in as GDPR compliant, even through they fail to follow best practices. In contrast, we classify services that directly send marketing emails without any confirmation email as potential GDPR violations. However, there is increasing case law requiring a proper double opt-in as a legal obligation. In a recent Austrian case [4], a minor was registered for a dating website by others. This registration caused the website to send him targeted marketing emails without confirming the email address beforehand. The Austrian Data Protection Authority decided that such a sign up procedure did not satisfy the requirements under Art. 32 GDPR.

For double opt-in registrations, we developed a script that classifies confirmation emails and automatically completes the registration. The script classifies the email by keyword-search in the subject, body, and other email headers (e.g., Reply-To or X-Headers). A manual inspection of 1000 emails shows that the classification works correctly in 96.8% of the cases (see Appendix A.2.2). The script extracts the link or confirmation code also by pattern matching. The extracted link or code is then used to complete the registration. We inspected all registrations, and those that the confirmation script could not finish were completed manually.

Figure 4a presents the first email sent by each service. Only 59% of websites that sent us at least one email first sent us a double opt-in email. Moreover, 5.5% of services sent us an unsolicited marketing email without any confirmation or double opt-in email.

4.3 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



(a) *Classification of the first email from the service.*

(b) *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.*

Fig. 4. *Email content legal analysis.*

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 Art. 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 [5].

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., [1, 28, 62]). Also, a list of websites that send passwords in plaintext is curated at `https://plaintextoffenders.com`, although it did not contain any of 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.

4.4 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. We provide the German legal background, but again, both provisions are derived from EU Community legislation (Art. 13(4) ePrivacy Directive and Art. 5 and 6 e-Commerce Directive) [59].

Marketing emails must contain a method for users to unsubscribe from subsequent emails. According to § 7(2) No. 4 (c) UWG, the method must be clear, unambiguous, and free of costs other than the transmission costs under the basic rates. Moreover, the GDPR clarifies in Art. 7(3) that it shall be as easy to opt-out as to opt-in. In addition, according to § 7(2) No. 4 UWG and with reference to § 6 of the German Telemedia Act (TMG), marketers must not disguise or conceal their identity. Companies that send marketing emails must include some company details (*legal notice*) in their emails based on § 5(1) TMG [69]. 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 are not all the requirements, but these are the requirements that apply most generally and are present in other jurisdictions.

We inspect both the presence of a method to unsubscribe from the emails and the existence of a legal notice. We combine both pattern matching and manual inspection to detect missing email content. The most common unsubscribe method is by an unsubscribe link placed either in the email body or in the X-Headers. An example of an alternative unsubscribe method requires the user to send the service an email to unsubscribe. We find the legal notices usually as a footnote to the email, containing the company name and service domain, based on which the legal notice can be detected.

Figure 4b show the portion of the email dataset missing an unsubscribe method and/or a legal notice. In total, 84.0% of emails properly contained both an unsubscribe method and a legal notice, whereas 2.8% were missing both. Note that the reported numbers are conservative because we used the number of services sending us any emails as the baseline, whereas we check the design requirements only in marketing emails.

Finally, the email dataset can reveal additional information about marketing emails, such as insight into the marketing trends, which are out of the scope of this work. We present examples in Appendix A.2.4.

4.5 Third party email sharing

The collection of email addresses and the sharing of these addresses to third parties for marketing purposes is subject to the same legal restrictions mentioned in Section 2.1 [40]. Hence, third parties that send marketing emails must be able to demonstrate that prior consent was obtained. This requires that a user must be specifically informed about whom their email address is shared with and for which marketing purposes [3]. Third parties must therefore be specifically named.

We check if all emails come from addresses whose first- and second-level domains match with that of the visited domain. We consider the combined top-level domains such as `co.uk` as the first-level domains. However, checking only the domain difference (as is the term "third party" used in CS) for our violation decision procedure is insufficient, since the domain name does not reflect the legal entity. Many websites have a dedicated domain name for sending emails, for example, `facebook.com` sends all the emails from `facebookmail.com`.

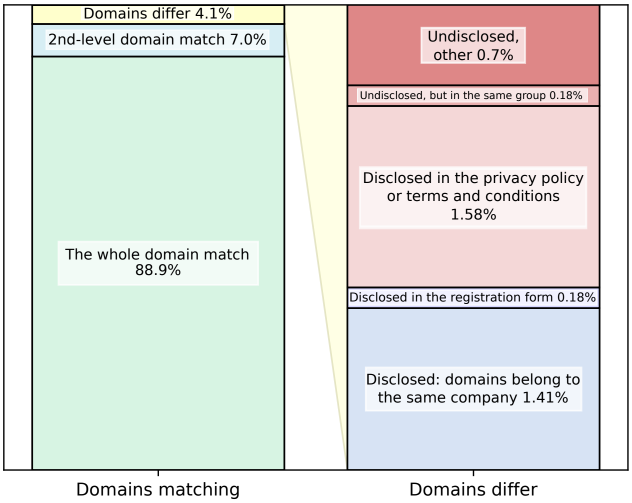


Fig. 5. *The first bar represents the matching of domains of the sender address and registration page. The second bar provides details how website disclose third-party sharing. We explain the three websites that remained in other group in Appendix A.2.3.*

We distinguish three scenarios based on sender's domains: (i) all the sender's domains match exactly, (ii) only their second-level domains match, and (iii) their domains differ completely.[3] The first bar in Figure 5 reports how often we encountered each scenario in the dataset. In the second bar in Figure 5, we focus on the senders whose domains are entirely different. We inspect how the website discloses how third parties can use the user's email address for sending marketing emails. In particular, we manually check the registration form content, the website's privacy policy, and the terms and conditions. If none of these inform the user about third parties, then we check if all sender domains are operated by the same group of companies based on publicly available sources such as corporate annual reports, Crunchbase, or the WHOIS database.

We conclude that services share email addresses of their users mostly within the same corporation, although very few of them disclose the practice of sharing email addresses with subsidiaries openly in their 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 the users [6, 29, 56]. During the fourteen months of our study, we observed that one of our email addresses received emails from nine different domains. Some of these domains were not stated in the registration form or in the terms and conditions. From another service, we received fraudulent emails without being notified about potential data breaches by the service.

5 Potential violations of the consent procedure

In the previous sections, we described the datasets of emails and websites. In this section, we combine these datasets using the unique email address as an identifier, and we report on the overall compliance. Using the combination of annotated legal properties, we propose a decision procedure for detecting potential violations of the ePrivacy Directive's (ePD) opt-in requirement and the GDPR's notion of consent.

[3] We inspected matching using `tldextract` Python package.

5.1 Opt-in violations of ePD

Under the ePrivacy Directive, marketers must obtain an individual's consent (opt-in) before they can send marketing emails. Figure 6 reports the adherence to the opt-in requirement. The leaf "No marketing" shows that 80% of websites never sent us a marketing email in the first place, and hence our violation decision procedure is not relevant for them. From the remaining websites in our analysis, 52.3% of them sent marketing emails despite their registration forms not mentioning marketing emails ("Email despite no consent" in Fig. 6). This constitutes a potential violation of Art. 13(1) of the ePD.

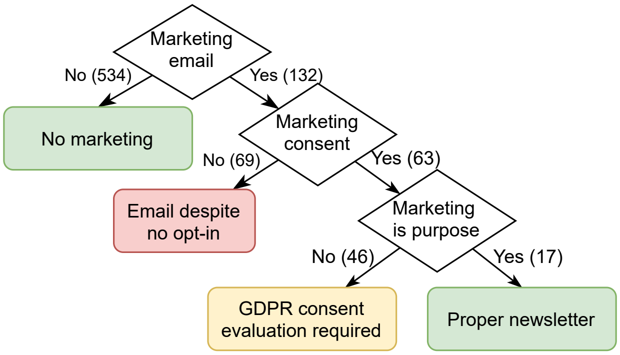


Fig. 6. *Decision procedure about opt-in validity based on legal properties.*

For 12.9% of websites that sent marketing emails, a newsletter subscription was the main purpose of the registration ("Proper newsletter" in Fig. 6). The remaining 34.8% had to be further assessed for consent requirements under the GDPR, as we explain next.

5.2 GDPR consent violations

As mentioned in Section 2.1, consent must be freely given, unambiguous, and specific. Based on this, we present selected potential violations of the GDPR's consent requirements. We describe the combination of legal properties that leads to a potential violation in Fig. 7.

Initially, we defined that obtaining consent without providing a specific marketing email checkbox is unspecific. Also, in line with case law, we classify the bundling of the marketing email consent with other purposes such as terms and conditions as unfreely obtained. In addition, we classify the practices of pre-checked marketing checkboxes and the nudging with visual features as ambiguous consent (see Section 2). Nudging is a typical example of a dark pattern; we summarize the similarities of potential violations from our study to dark patterns in Appendix A.1.4.

At least 43.5% of websites that sent marketing emails did not meet one of these requirements on consent ("Email after invalid consent" in Fig. 7). Surprisingly, we received marketing emails even from websites that did not violate any of our selected consent requirements. As we instructed annotators not to provide consent during the registration, such marketing emails most likely lack valid consent ("Email despite user did not opt-in" in Fig. 7).

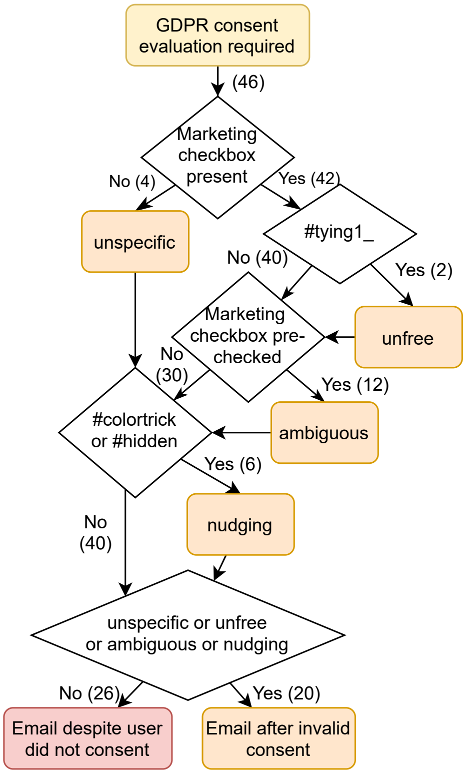

GDPR consent evaluation required
(46)
Marketing checkbox present
No (4)
Yes (42)
unspecific
#tying1_
No (40)
Yes (2)
unfree
Marketing checkbox pre-checked
No (30)
Yes (12)
ambiguous
#colortrick or #hidden
Yes (6)
nudging
No (40)
unspecific or unfree or ambiguous or nudging
No (26)
Yes (20)
Email despite user did not consent
Email after invalid consent

Fig. 7. *Decision procedure about consent validity based on legal properties.*

Our decision procedure detects a selection of potential violations. Note that when our procedure identifies no potential violations, a website may still fail to comply with consent requirements. For example, our procedure does not analyze the specific wording of consent declarations. Nevertheless, our procedure detects a substantial number of potential violations. Indeed, it finds that 17.3% of websites have at least one potential violation.

5.3 Summary

In Fig. 8 we present the potential violations in emails from Section 4 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.

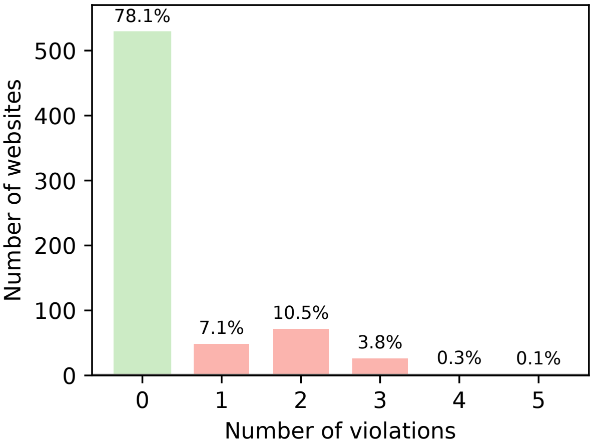


Fig. 8. *Histogram showing number of websites with the given number of potential violations. We report the potential violations from 676 websites, i.e., the union of websites where the annotation resulted in successful registration and websites that sent us an email. Note that we are conservative in determining potential violations, so the reported number does not imply that 78.1% of websites are fully compliant.*

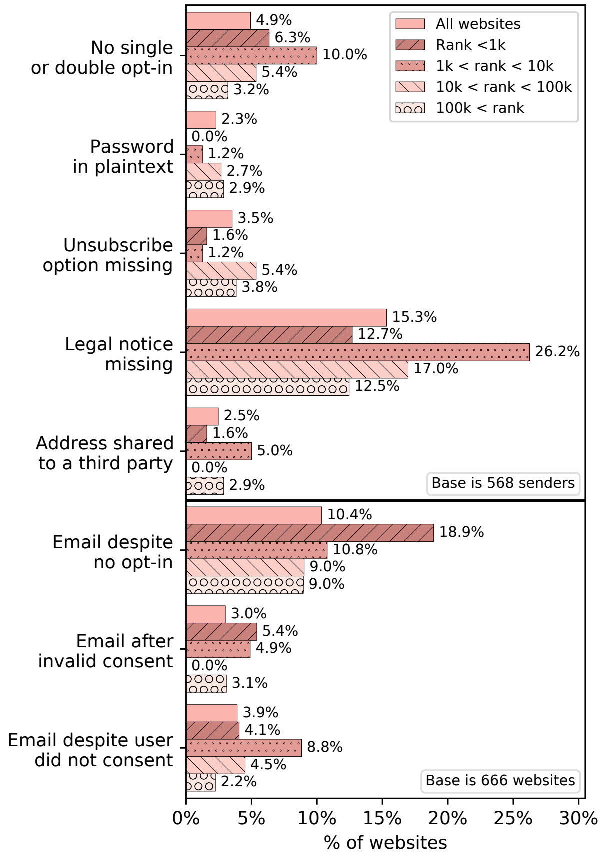


Fig. 9. *Summary of all potential violations of this study and the split into popularity groups by rank.*

In Fig. 9 we summarize the presence of all potential violations discussed in this study. In addition, we split the graph into groups by website's ranking according to their Alexa rank. Note that more popular websites are not more compliant than lower ranked websites. Moreover, for the potential violation "Email despite no opt-in," the websites with high rank show more potential violations than those with low rank (p-value of the two proportions Z-Test of the rank < 1k against data of all other ranks is 0.156 after adjustment for multiple measurements by Holm–Bonferroni method). The number of websites of rank 1k-10k not sending legal notices is far larger than the websites of other ranks (including high-rank websites). This observation has a p-value of 0.054.

6 Potential for automation

We found that 22% of websites have potential violations. Given this lack of compliance, regulators should take note and might wish to step in. However, our procedure still relies on many manual steps. Regulators would therefore benefit from an automated tool that scales up the detection of violations from sending marketing emails. In this section, we offer a statistical analysis that speaks to the feasibility of such automation.

We study the statistical properties of our annotated datasets. We define features that we extract from raw HTML, and afterwards, we train logistic models and compute which of these features are the most influential for deciding if a website satisfies a legal property. We also show how these features facilitate detecting potential violations.

6.1 Registration form and email features

During the website annotation process, we collect from each successfully registered website the registration form HTML subtree. A classification of this entire form is not possible, so we instead extract first-level features, described in Table 2.

Table 2. Features that we extract from each input, select, and button tag of the registration form.

Feature type	Individual features
HTML tags	tag name, accompanying label text
HTML attributes	class, type, attribute, placeholder, value
Is required?	HTML required attribute, asterisk in the label
Text processing	tag purpose extracted by keyword matching

The textual features are further processed by a feature-specific bag of the top 100 words after lemmatization and stop-words removal. As the number of input fields of a registration form is not limited, we select a fixed number of inputs of each type and remove the trailing ones (the last ones in the form), or fill missing values by empty values (NaN). We order the input fields according to the purpose from Table 2. So, for example, marketing checkboxes or email text field always have a fixed position. Our experiments all run in under 2 minutes, even with a large number of features, so we can use over 20 form inputs with over 10k features in total.

We apply a similar feature processing to email dataset. We extract a bag of words from the subject and body which is similarly as the registration form in HTML. In addition, we extract the number of links and images in the email.

6.2 Feature analysis

We evaluate the usefulness of our datasets by training a simple logistic regression model for each of the legal properties and for the classification of emails as marketing or servicing. We report the precision, recall, and F1 score for a subset of the most important properties in Table 3. Clearly, the precision is affected by the number of positive samples in the training dataset.

Table 3. Results of logistic regression for legal properties and for whether an email is marketing. The last column represents the percentage of positive samples (ps). The confidence intervals are based on five-fold cross-validation. The results are from EN dataset; the DE dataset is reported in Table 7 in the Appendix.

Property	Precision	Recall	F1	ps
ma_consent	82.5% ± 3.4%	70.8% ± 6.6%	76.0% ± 3.7%	38%
ma_purpose	19.6% ± 5.3%	51.7% ± 26.0%	27.4% ± 8.3%	7%
ma_checkbox	79.8% ± 11.6%	73.3% ± 7.3%	75.5% ± 3.0%	31%
ma_pre_checked	25.1% ± 13.8%	58.3% ± 33.3%	34.7% ± 18.7%	7%
ma_forced	1.8% ± 3.6%	10.0% ± 20.0%	3.1% ± 6.2%	2%
pp_checkbox	62.8% ± 6.7%	77.6% ± 3.9%	69.3% ± 5.2%	21%
pp_forced	61.3% ± 20.4%	71.8% ± 14.9%	64.6% ± 14.6%	20%
tc_checkbox	89.4% ± 7.6%	84.9% ± 7.0%	87.0% ± 6.4%	28%
tc_forced	77.5% ± 6.8%	76.3% ± 10.8%	76.2% ± 6.3%	26%
#hidden	18.5% ± 2.7%	73.3% ± 17.0%	29.5% ± 4.7%	12%
#forced	52.4% ± 4.4%	68.9% ± 8.3%	59.0% ± 2.2%	37%
#tying1	0.0% ± 0.0%	0.0% ± 0.0%	0.0% ± 0.0%	1%
Marketing email	97.4% ± 1.2%	98.0% ± 0.5%	97.7% ± 0.7%	80.6%

We observe vast differences among models for legal properties and emails, both in performance and features utility. While legal property models used at most 5% of all the features, the email model used over 95% of features, likely due to longer texts and denser bags of words for emails than for forms. We identify the useful features by a non-zero coefficient in the logistic regression model. We report in Tables 8 and 9 in the Appendix the most important features for a decision on whether the registration's purpose is to receive marketing emails (ma_purpose property) and if an email is marketing. For example, the ma_purpose model used a binary feature whether a password input is in the form. If there is a password input field, then the form likely does not satisfy ma_purpose. For emails, the presence of keywords like "account" or "confirm" in the email body indicate that the email is likely servicing, whereas a high number of links signalizes that email is marketing. Testing these two keywords alone already achieves 76% accuracy.

This preliminary analysis illustrates one of the applications of our datasets. Namely, combining the insights above with the automated registration procedure developed by Drakonakis et al. [15] or by our team [43], regulatory agencies could automatically detect potential violations. Furthermore, this analysis can be extended in future work with machine learning to fully automate this detection. Finally, our datasets can be used in the future as a source to analyze marketing trends, tracking in marketing emails, and how websites ask for consent.

Possibility of adversarial modification

Features for our classification are both the text of the form and numerical properties extracted from the HTML code. Both of these can potentially be manipulated to cause misclassifications by our models. We discuss the possibility of this below.

The number of password input fields is an important feature for the property ma_purpose (see Table 8). On WebKit-based browsers, it is possible to style a text input field such that it resembles a password field, which might fool our model. The textual features used by the models are based on a bag-of-words model, which is unable to represent word relations. The word selection and placement of invisible text might lead to both false positives and false negatives.

There are multiple ways of preventing these adversarial modifications. The feature extraction can use CSS, visual representation, and more advanced text models as BERT [13]. We can add artificial adversarial samples to our training dataset and force the model to use more reliable features. Lastly, Goodfellow et al. [30] and Javanmard et al. [34] proposed defense mechanisms against adversarial manipulation for logistic regression.

7 Related work

Newsletter analysis

Studies analyzing the content of emails either depend on a publicly available email dataset or their authors must collect an email dataset by signing up for services similar to our approach. The research closest to our study are the following three publications. Englehardt et al. [18] subscribed to 902 newsletters by crawling 15 700 shopping and news websites. They analyze how loading the email or following links in it causes information leakage, and they show that 30% of emails leak the recipient's email address to a third party. They also study the tracking protection of email servers and clients and propose new privacy measures. Our study focuses on the legal aspects of sending marketing emails, mostly from websites where the registration serves other purposes than only subscription to newsletters. Englehardt et al. subscribed to emails exclusively at those websites that we annotate as ma_purpose.

In the second study, Hamin [31] analyzes the content of election campaigns. She crawled 4487 campaign websites, and successfully subscribed to 1778 newsletters. A follow-up study of 2020 US elections by Mathur et al. [54] observed that 348 out of 2800 email campaigns shared the email address with a third party, while only 25% of those campaigns disclosed their email sharing practice. Both of these studies also analyze the email content, but their focus is on manipulative tactics and political implications. As in the previous paragraph, these two studies target a narrow group of email senders who send emails to a) subscribed users, b) who are interested in elections, and c) located in the US. Our study is generic, with a subscription to marketing emails (ma_purpose) corresponding to only 10% of the registrations. Even from these subscriptions, we did not observe as much email sharing as Mathur et al. The difference could be a result of EU privacy regulations protecting user's more than the US, or due to the political campaigns sending emails more aggressively than websites that mainly advertise their products, which are present in our study.

Consent compliance analysis

We study consent with marketing emails, but websites need to obtain consent for other processing purposes. Oh et al. [61] state four conditions on consent according to the GDPR. They inspect these conditions both manually on 500 websites and by crawling 10 000 websites. They show that their crawler is 96% aligned with the human decision. Their study partially overlaps with our inspection of GDPR consent violations in Section 5.2. However, our study is focused on marketing emails and goes legally more in-depth, while their study is related to privacy policies and is more generic. Our decision procedure requires observing the data misuse (receiving unsolicited marketing email), so we must complete a registration, which is challenging to automate. In contrast, their crawler detects violations solely by observing the registration form without any interaction and before the act of data misuse.

Other researchers focus on consent for cookie usage. A user study by Machuletz et al. [48] inspects how misleading cookie consent dialogs are. They confirm that users are nudged into less favorable choices by making these choices more accessible. Our work quantifies nudging with legal properties #hidden, #settings, and #colortrick, showing it is not as common in registration forms as in cookie popups. Matte et al. [55] detected cookie banner privacy violations on 53% of websites. Santos et al. [63] define 22 legal requirements on cookie banners and describe how to verify them. A similar study by Trevisan et al. [65] summarizes EU requirements on cookie consents that they can check automatically. Most notably, they report that 49% of websites activate cookies before the user gives consent. Nouwens et al. [60] study dark patterns of consent pop-ups, finding that only 11.8% of websites comply with the GDPR. These studies are complementary to our work as we do not analyze cookie consents.

Website compliance analysis

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

Linden et al. [47] and Degeling et al. [10] analyze how privacy policies changed with the GDPR coming into legal force. They observe an increase in the length and the number of policies and an improvement in GDPR compliance. Amos et al. [2] observed similar results in their longitudinal study of privacy policies. Liepina et al. [46] present Claudette, a scanner for GDPR violations in privacy policies. Harkous et al. [32] propose Polisis, a privacy policy scanner that summarizes policies' content. We used Polisis to analyze whether websites disclose sharing email addresses with third parties. A semantic text analysis of policies by Bui et al. [7] can further improve the automation by extracting the names of the third parties defined in the privacy policy. However, this work was published after we finished our privacy policies analysis using only Polisis.

8 Conclusions and future directions

We manually registered on 666 out of 1000 websites and annotated the registration procedures and emails that these websites sent. We proposed a decision procedure that, based on the annotated legal properties, detects potential violations of opt-in and consent for sending marketing emails. We then evaluated the emails that we received, finding services that send marketing emails without valid consent in 17.3% of the cases. Furthermore, 17.7% of the services sent us an email that potentially violated the legal requirements on email content. In total, 21.9% of the websites committed at least one potential violation.

The results of our study indicate that a substantial number of websites may be violating European privacy and unfair competition rules as far as marketing emails are concerned. The non-compliance with such rules is not too surprising, given that it is cumbersome to detect violations and enforce these rules.

Our study can inform the policy and regulatory debate about privacy and unfair competition law on the Internet in several ways. First, it provides policymakers and regulators with an estimate of the prevalence of non-compliance. Second, it shows a path of how to increase compliance: A next step is to automate the procedure outlined in this study, helping overloaded and underfunded regulatory agencies to police the Internet more efficiently and increase compliance with legal requirements.

As future work, we plan to train machine learning models from the annotated dataset to automatically detect potential violations. Combining this tool with an automated registration procedure, we could detect potential violations in the wild, without any of the time-consuming manual work done by annotators for the present study. This could open up novel and cost-effective ways for ensuring compliance of websites with legal rules that are aimed at protecting millions of consumers on the Internet.

Acknowledgment

The authors would like to thank Martin Monsch, Nils Heinemann, Ozan Yildrim, and Patrick Krebs for excellent legal research and Joachim Posegga for an access to a web proxy at the University of Passau. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] F. Al Maqbali and C. J. Mitchell. "Web Password Recovery: A Necessary Evil?" In: *Proceedings of the Future Technologies Conference*. Springer. 2018, pp. 324–341.

[2] R. Amos, G. Acar, E. Lucherini, M. Kshirsagar, A. Narayanan, and J. Mayer. "Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset." In: *Proceedings of The Web Conference 2021*. WWW '21. Association for Computing Machinery, Apr. 19, 2021, p. 22. DOI: 10.1145/3442381.3450048.

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

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

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

[6] Y. Bakos, F. Marotta-Wurgler, and D. R. Trossen. "Does anyone read the fine print? Consumer attention to standard-form contracts." In: *The Journal of Legal Studies* 43.1 (2014), pp. 1–35.

[7] D. Bui, K. G. Shin, J.-M. Choi, and J. Shin. "Automated Extraction and Presentation of Data Practices in Privacy

Policies." In: *Proceedings on Privacy Enhancing Technologies* 2021.2 (2021), pp. 88–110.

[8] M. Chatzimpyrros, K. Solomos, and S. Ioannidis. "You Shall Not Register! Detecting Privacy Leaks Across Registration Forms." In: *Computer Security*. Springer, 2019, pp. 91–104.

[9] J. Cohen. "A coefficient of agreement for nominal scales." In: *Educational and psychological measurement* 20.1 (1960), pp. 37–46.

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

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

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

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv preprint arXiv:1810.04805* (2018).

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

[15] K. Drakonakis, S. Ioannidis, and J. Polakis. "The Cookie Hunter: Automated Black-box Auditing for Web Authentication and Authorization Flaws." In: *Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security*. 2020, pp. 1953–1970.

[16] L. Edwards. *The New Legal Framework for E-Commerce in Europe*. ISBN 978-1-847-31261-7, Hart Publishing, 2005.

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

[18] S. Englehardt, J. Han, and A. Narayanan. "I never signed up for this! Privacy implications of email tracking." In: *Proceedings on Privacy Enhancing Technologies* 2018.1 (2018), pp. 109–126.

[19] L. Epstein and A. D. Martin. *An introduction to empirical legal research*. Oxford University Press, 2014.

[20] European Commission. *Guidance on the implementation/application of Directive 2005/29/EC on Unfair Commercial Practices*. May 25, 2016.

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

[22] European Data Protection Board. *Guidelines 05/2020 on consent under Regulation 2016/679 (GDPR)*. May 2020.

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

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

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

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

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

[28] N. Gelernter, S. Kalma, B. Magnezi, and H. Porcilan. "The password reset MitM attack." In: *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2017, pp. 251–267.

[29] J. Gluck, F. Schaub, A. Friedman, H. Habib, N. Sadeh, L. F. Cranor, and Y. Agarwal. "How short is too short? Implications of length and framing on the effectiveness of privacy notices." In: *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. 2016, pp. 321–340.

[30] I. J. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and harnessing adversarial examples." In: *arXiv preprint arXiv:1412.6572* (2014).

[31] M. Hamin. *"don't ignore this:" Automating the Collection and Analysis of Campaign Emails*. Tech. rep. Princeton University, 2018.

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

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

[34] A. Javanmard and M. Soltanolkotabi. "Precise statistical analysis of classification accuracies for adversarial training." In: *arXiv preprint arXiv:2010.11213* (2020).

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

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

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

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

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

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

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

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

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

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

[45] Legal team of the Certified Senders Alliance. *DOI: if not now, then when?!* https://certified-senders.org/blog/doi-if-not-now-then-when/. 2017. (Visited on 08/25/2021).

[46] R. Liepin, G. Contissa, K. Drazewski, F. Lagioia, M. Lippi, H.-W. Micklitz, P. Palka, G. Sartor, and P. Torroni. "GDPR privacy policies in CLAUDETTE: Challenges of omission, context and multilingualism." In: *3rd Workshop on Automated Semantic Analysis of Information in Legal Texts, ASAIL 2019*. Vol. 2385. CEUR-WS. 2019.

[47] T. Linden, R. Khandelwal, H. Harkous, and K. Fawaz. "The privacy policy landscape after the GDPR." In: *Proceedings on Privacy Enhancing Technologies* 2020.1 (2020), pp. 47–64.

[48] D. Machuletz and R. Böhme. "Multiple purposes, multiple problems: A user study of consent dialogs after GDPR." In: *Proceedings on Privacy Enhancing Technologies* 2020.2 (2020), pp. 481–498.

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

[50] *Is email marketing dead?* https://optinmonster.com/is-email-marketing-dead-heres-what-the-statistics-show/.

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

[52] A. Mathur, G. Acar, M. J. Friedman, E. Lucherini, J. Mayer, M. Chetty, and A. Narayanan. "Dark patterns at scale: Findings from a crawl of 11K shopping websites." In: *Proceedings of the ACM on Human-Computer Interaction* 3.CSCW (2019), pp. 1–32.

[53] A. Mathur, M. Kshirsagar, and J. Mayer. "What makes a dark pattern... dark? Design attributes, normative considerations, and measurement methods." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 2021, pp. 1–18.

[54] A. Mathur, A. Wang, C. Schwemmer, M. Hamin, B. M. Stewart, and A. Narayanan. *Manipulative tactics are the norm in political emails: Evidence from 100K emails from the 2020 U.S. election cycle*. https://electionemails2020.org. 2020.

[55] C. Matte, N. Bielova, and C. Santos. "Do Cookie Banners Respect my Choice? Measuring Legal Compliance of Banners from IAB Europe's Transparency and Consent Framework." In: *2020 IEEE Symposium on Security and Privacy (SP)*. IEEE. 2020, pp. 791–809.

[56] A. M. McDonald and L. F. Cranor. "The cost of reading privacy policies." In: *ISJLP* 4 (2008), p. 543.

[57] D. Mederle. *The regulation of spam and unsolicited commercial emails (Die Regulierung von Spam und unerbetenen kommerziellen E-Mails)*. Heymanns, 2010. ISBN: 3452272680.

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

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

[60] M. Nouwens, I. Liccardi, M. Veale, D. Karger, and L. Kagal. "Dark Patterns after the GDPR: Scraping Consent Pop-ups and Demonstrating their Influence." In: *Proceedings of the 2020 CHI conference on human factors in computing systems*. 2020, pp. 1–13.

[61] J. Oh, J. Hong, C. Lee, J. J. Lee, S. S. Woo, and K. Lee. "Will EU's GDPR Act as an Effective Enforcer to Gain Consent?" In: *IEEE Access* (2021).

[62] C. Routh, B. DeCrescenzo, and S. Roy. "Attacks and vulnerability analysis of e-mail as a password reset point." In: *2018 Fourth International Conference on Mobile and Secure Services (MobiSecServ)*. IEEE. 2018, pp. 1–5.

[63] C. Santos, N. Bielova, and C. Matte. "Are cookie banners indeed compliant with the law? Deciphering EU legal requirements on consent and technical means to verify compliance of cookie banners." In: *Technology and Regulation (2020)*. 2019, pp. 91–135.

[64] J. Sim and C. C. Wright. "The kappa statistic in reliability studies: use, interpretation, and sample size requirements." In: *Physical therapy* 85.3 (2005), pp. 257–268.

[65] M. Trevisan, S. Traverso, E. Bassi, and M. Mellia. "4 years of EU cookie law: Results and lessons learned." In: *Proceedings on Privacy Enhancing Technologies* 2019.2 (2019), pp. 126–145.

[66] J. Weiser. "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)." In: *GRUR-Prax, (Gewerblicher Rechtsschutz und Urheberrecht, Praxis im Immaterialgüter- und Wettbewerbsrecht)* 2018.10 (2018), p. 291.

[67] S. Wilson, F. Schaub, A. A. Dara, F. Liu, S. Cherivirala, P. G. Leon, M. S. Andersen, S. Zimmeck, K. M. Sathyendra, N. C. Russell, et al. "The creation and analysis of a website privacy policy corpus." In: *Proceedings of the 54th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2016, pp. 1330–1340.

[68] S. Zimmeck, P. Story, D. Smullen, A. Ravichander, Z. Wang, J. Reidenberg, N. C. Russell, and N. Sadeh. "MAPS: Scaling privacy compliance analysis to a million apps." In: *Proceedings on Privacy Enhancing Technologies* 2019.3 (2019), pp. 66–86.

[69] K. A. Zscherpe. "Direct marketing by e-mail – How can companies proceed legally? (Direktmarketing per E-Mail – Wie können Unternehmen rechtlich einwandfrei vorgehen?)" In: *Journal of Business and Consumer Law, (Zeitschrift für Wirtschafts- und Verbraucherrecht)* 2008.9 (2008), pp. 327–322.

A Appendix

A.1 Annotation process

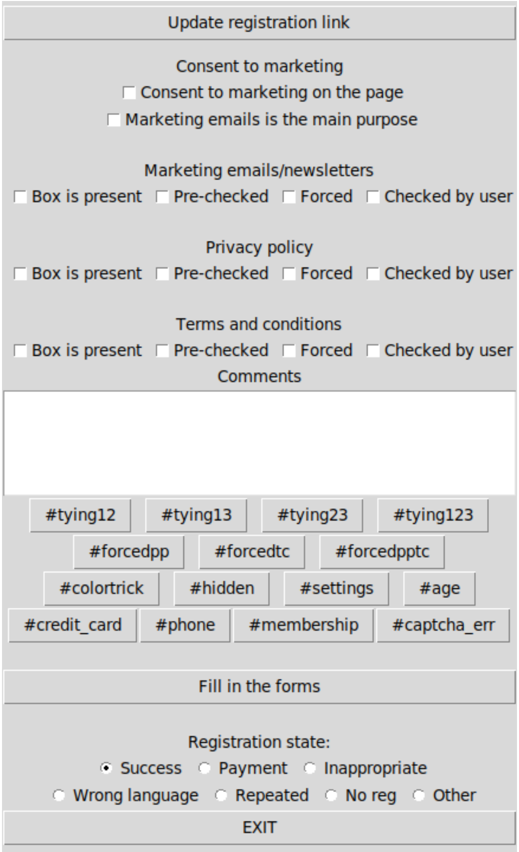
The dataset, legal instructions, and supplementary materials are available on request at `https://forms.gle/dTGpfs5vKqdLz8sQ7`. In this section, we provide additional information to annotation process.

A.1.1 Pilot study

The exploratory pilot study 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.1.2 Annotating tool

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



Update registration link

Consent to marketing
□ Consent to marketing on the page
□ Marketing emails is the main purpose

Marketing emails/newsletters
□ Box is present □ Pre-checked □ Forced □ Checked by user

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

Terms and conditions
□ Box is present □ Pre-checked □ Forced □ Checked by user

Comments

#tying12 #tying13 #tying23 #tying123
#forcedpp #forcedtc #forcedpptc
#colortrick #hidden #settings #age
#credit_card #phone #membership #captcha_err

Fill in the forms

Registration state:
● Success ○ Payment ○ Inappropriate
○ Wrong language ○ Repeated ○ No reg ○ Other

EXIT

Fig. 10. *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 Selenium library, which loads the registration page in the first place and auto-fills the forms.*

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 10. 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.

Table 4. The individual Cohen's κs of legal properties. Note that $\kappa = 1$ implies full agreement, while $\kappa = -1$ is full disagreement.

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

Table 5. Contingency tables of checkbox values. Rows represent the first 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 6. Contingency tables of hashtag values. Rows represent the first 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

A.1.3 Inter-annotator agreement

Sim et al. [64] 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 4. Therefore, we also present the contingency tables for every legal properties in Tables 5 and 6.

A.1.4 Linkage to dark patterns

In this section, we compare our defined potential violation types to the taxonomy of dark patterns by Marthur et al. [53]. We refer to terms from [53] in *italics*.

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

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

A.1.5 Registered accounts

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

Note that if we would have to split the registration and annotation processes, we would loose significant information. The annotators need to see the whole registration to determine all the legal properties. In addition, the annotators would be provided a potentially wrong form, which by our approach would not be resolved by resolving annotation. Moreover, we would not have subscribed to many of the 701 websites.

A.1.6 Email address generation

We considered two options for generating emails: setting up a custom email server or using Gmail "+ suffixes." An appended + sign and any combination of alphanumerical characters are ignored for resolving the recipient for Gmail addresses. This way, `john@gmail.com` also receives emails from `john+friends@gmail.com`. We chose the custom email server as it cannot be detected and exploited by marketing services. This differs from [31] that used Gmail suffixes.

A.2 Datasets content

We now elaborate on our analyses in Section 6, showing insights that help to understand the content of the datasets 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.

A.2.1 Successful form annotations

In Figure 11, 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 12 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.

A.2.2 Email classification

As we stated in Section 4.2, the pattern-matching classification of emails misclassified 32 emails from manually inspected 1000. Those were 11 marketing emails classified as servicing and 21 servicing emails classified as

Table 7. Results of logistic regression for legal properties based on DE dataset, with the percentage of positive samples (ps) in the last column. The confidence intervals are based on five-fold cross-validation.

Property	Precision	Recall	F1	ps
ma_consent	82.3% ± 7.2%	73.0% ± 9.3%	77.1% ± 7.1%	44%
ma_purpose	12.7% ± 7.4%	36.7% ± 19.4%	18.7% ± 10.4%	5%
ma_checkbox	80.6% ± 7.5%	72.7% ± 4.6%	76.2% ± 4.5%	38%
ma_pre_checked	6.7% ± 8.2%	20.0% ± 24.5%	10.0% ± 12.2%	4%
ma_forced	16.7% ± 11.5%	60.0% ± 37.4%	23.7% ± 13.7%	4%
pp_checkbox	81.4% ± 5.9%	72.0% ± 7.9%	76.0% ± 3.8%	45%
pp_forced	79.2% ± 10.4%	75.5% ± 11.1%	76.1% ± 4.8%	44%
tc_checkbox	56.9% ± 5.5%	71.7% ± 14.8%	63.2% ± 9.1%	28%
tc_forced	54.6% ± 6.6%	66.5% ± 6.4%	59.6% ± 4.8%	27%
#hidden	2.9% ± 5.7%	20.0% ± 40.0%	5.0% ± 10.0%	3%
#forced	41.0% ± 7.1%	61.9% ± 17.0%	47.9% ± 6.4%	27%
#tying1	4.0% ± 8.0%	20.0% ± 40.0%	6.7% ± 13.3%	3%
Marketing email	88.3% ± 1.4%	97.3% ± 0.7%	92.6% ± 0.7%	69.7%

Table 8. The figure shows the five most important features based on the model that decides ma_purpose property, i.e., if the form serves as an email subscription. We identify these features by the highest absolute values of the coefficients of a logistic regression model. A coefficient value interprets similarly as a correlation. A positive coefficient means the feature needs to be true for an email subscription form, while a negative coefficient signalizes a negative correlation between the feature and decision. The model correctly identified that forms without passwords serve more likely only as an email subscription. Such forms also more often contained multiple checkboxes.

Feature	Coefficient
Is there a password input field in the form?	-0.442
The number of input fields of the form	-0.212
Is password input field required to submit the form?	-0.21
Contains the form multiple unidentified checkboxes?	0.203
The number of password input fields of the form	-0.202

Table 9. Selection from the most important features for classifying whether an email is marketing. The interpretation is the same as in Table 8. The model successfully identifies keywords that denote confirmation emails, as well as unsubscribe keyword and high number of links is typical for marketing emails.

Feature	Coefficient
"Account" in the email body	-0.197
"Confirm" in the email body	-0.117
"Address" in the email body	-0.097
"Account" in the email subject	-0.087
"Thanks" in the email body	-0.082
"Unsubscribe" in the email body	0.06
Number of <a> the email body	0.056

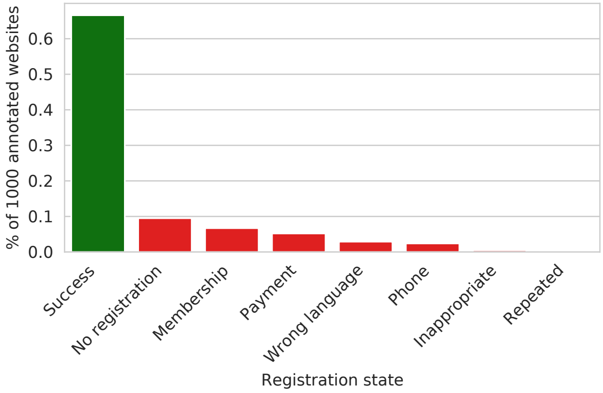
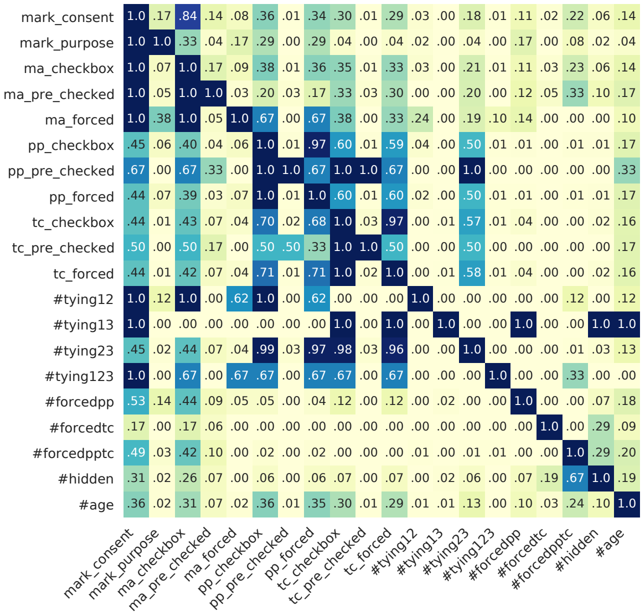


Fig. 11. *Registration state of the resolved annotations. For agreement, Cohen's κ for distinction between successful and failed registrations is 0.64.*



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

Fig. 12. *Interdependence of legal properties as a ratio of annotations with the property of the row that has also the property of the column. A cell in the first row, second column, marks how many websites with marketing consent (row label) have the marketing purpose (column label).*

marketing. In addition to these misclassifications, the information was incomplete in an additional 30 cases:

- missed a confirmation code in 3 emails,
- collected a wrong code in 1 email,
- found a wrong confirmation link in 16 emails, and
- another method of activation was specified in 10 emails. Once the sender required us to send them an email and 9 times the email contained a generated password, which serves as a confirmation.

A.2.3 Third party email sharing

As we stated in Section 4.5, we classified four websites as "other." In the first case, apart from the same physical address, we did not have enough indications that the two Chinese companies were part of the same group. In the second case, the third party was maintaining a reward system on behalf of the website. The third website was offline and could no longer be analyzed. The last service's data likely breached (reported by other users), which lead to us receiving fraudulent emails. The service did not notify its users about any breach.

A.2.4 Marketing trends in newsletters

For our study, we annotated emails during the period starting in September 2020 and ending in February 2021, so we were able to observe several marketing trends influencing the email content. We observed that 5.8%, 11.7%, and 4.2% of marketing emails were related to Black Friday, Christmas, and New Year, respectively. These topics become relevant during autumn and winter, but we did not observe an overall increase in the number of marketing emails. Also, 17.2% of all processed emails were related to the Covid pandemic. As the frequency of marketing emails did not change during these periods (see Figure 13), the observations suggest that trending topics are used to improve marketing campaigns, but they do not generate new newsletter traffic. This hypothesis is based on the fact that during the limited period of the study, we did not observe any spikes in the number of newsletters during these periods. However, to confirm this hypothesis, we would need a more longitudinal study.

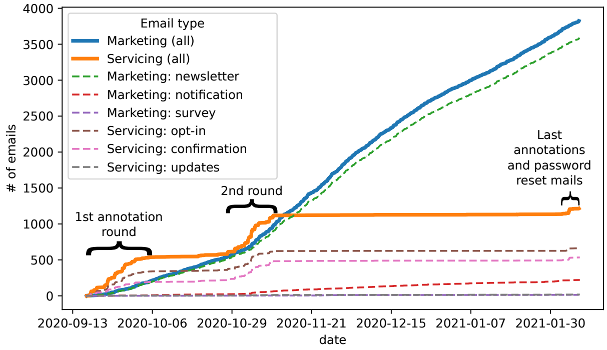


Fig. 13. *Classification of the manually annotated emails, where reported marketing and servicing numbers are the sum of number of email of each subtype. The x-axis is continuous over the period of our study. We can see that the number of servicing emails is constant function in number of registrations ($\approx 1.2 \cdot$ number of accounts), while number of emails linearly increases over time (≈ 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.*