Thomas Linden, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz*

The Privacy Policy Landscape After the GDPR

Abstract: The EU General Data Protection Regulation (GDPR) is one of the most demanding and comprehensive privacy regulations of all time. A year after it went into effect, we study its impact on the landscape of privacy policies online. We conduct the first longitudinal, in-depth, and at-scale assessment of privacy policies before and after the GDPR. We gauge the complete consumption cycle of these policies, from the first user impressions until the compliance assessment. We create a diverse corpus of two sets of 6,278 unique English-language privacy policies from inside and outside the EU, covering their pre-GDPR and the post-GDPR versions. The results of our tests and analyses suggest that the GDPR has been a catalyst for a major overhaul of the privacy policies inside and outside the EU. This overhaul of the policies, manifesting in extensive textual changes, especially for the EU-based websites, comes at mixed benefits to the users.

While the privacy policies have become considerably longer, our user study with 470 participants on Amazon MTurk indicates a significant improvement in the visual representation of privacy policies from the users' perspective for the EU websites. We further develop a new workflow for the automated assessment of requirements in privacy policies. Using this workflow, we show that privacy policies cover more data practices and are more consistent with seven compliance requirements post the GDPR. We also assess how transparent the organizations are with their privacy practices by performing specificity analysis. In this analysis, we find evidence for positive changes triggered by the GDPR, with the specificity level improving on average. Still, we find the landscape of privacy policies to be in a transitional phase; many policies still do not meet several key GDPR requirements or their improved coverage comes with reduced specificity.

Keywords: GDPR, Privacy, Security, Automated Analysis of Privacy Policy, Polisis

DOI 10.2478/popets-2020-0004
Received 2019-05-31; revised 2019-09-15; accepted 2019-09-16.

Thomas Linden: University of Wisconsin, E-mail: tlinden2@wisc.edu
Rishabh Khandelwal: University of Wisconsin, E-mail: rkhandelwal3@wisc.edu
Hamza Harkous: École Polytechnique Fédérale de Lausanne, E-mail: hamza.harkous@gmail.com
***Corresponding Author: Kassem Fawaz:** University of Wisconsin, E-mail: kfawaz@wisc.edu

1 Introduction

For more than two decades since the emergence of the World Wide Web, the "Notice and Choice" framework has been the governing practice for the disclosure of online privacy practices. This framework follows a market-based approach of voluntarily disclosing the privacy practices and meeting the fair information practices [25]. The EU's recent General Data Protection Regulation (GDPR) promises to change this privacy landscape drastically. As the most sweeping privacy regulation so far, the GDPR requires information processors, across all industries, to be transparent and informative about their privacy practices.

Research Question

Researchers have conducted comparative studies around the changes of privacy policies through time, particularly in light of previous privacy regulations (e.g., HIPAA[1] and GLBA[2]) [1, 3, 5, 26]. Interestingly, the outcomes of these studies have been consistent: (1) the percentage of websites with privacy policies has been growing, (2) the detail-level and descriptiveness of policies have increased, and (3) the readability and clarity of policies have suffered.

The GDPR aims to address shortcomings of previous regulations by going further than any prior privacy regulation. One of its distinguishing features is that non-complying entities can face hefty fines, the maximum of 20 million Euros or 4% of the total worldwide annual revenue. Companies and service providers raced to change their privacy notices by May 25^{th}, 2018 to comply with the new regulations [9]. With the avalanche of updated privacy notices that users had to accommodate, a natural question follows:

What is the impact of the GDPR on the landscape of online privacy policies?

Researchers have recently started looking into this question by evaluating companies' behavior in light of the GDPR. Their approaches, however, are limited to a small number of websites (at most 14) [8, 32]. Concurrent to our work, Degeling et al. [9], performed the first large-scale study focused on the evolution of the cookie consent notices, which have been hugely reshaped by the GDPR (with 6,579 EU websites). They also touched upon the

1 The Health Information and Portability Accountability Act of 1996.
2 The Gramm-Leach-Bliley Act for the financial industry of 1999.

growth of privacy policies, finding that the percentage of sites with privacy policies has grown by 4.9%.

Methodology and Findings

Previous studies have not provided a comprehensive answer to our research question. In this paper, we answer this question by presenting the first on-scale, longitudinal study of privacy policies' content in the context of the GDPR. We develop an automated pipeline for the collection and analysis of 6,278 unique English privacy policies by comparing pre-GDPR and post-GDPR versions. These policies cover the privacy practices of websites from different topics and regions. We approach the problem by studying the change induced on the entire experience of the users interacting with privacy policies. We break this experience into five stages:

A. Presentation (Sec. 4). To quantify the progress in privacy policies' *presentation*, we gauge the change in user perception of their interfaces via a user study involving 470 participants on Amazon Mechanical Turk. Our results find a positive change in the attractiveness and clarity of EU-based policies; however, we find that outside the EU, the visual experience of policies is not significantly different.

B. Text-Features (Sec. 5). We study the change in the policies' using high-level syntactic text features. Our analysis shows that both *EU* and *Global* privacy policies have become significantly longer, with (+35%, +25%) more words and (+33%, +22%) more sentences on average respectively. However, we do not observe a major change in the metrics around the sentence structure.

We devise an approach, inspired by goal-driven requirements engineering [35], to evaluate coverage, compliance, and specificity in the privacy policies. While previous longitudinal studies either relied on manual investigations or heuristics-based search queries [1, 3, 9], we build on the recent trend of automated semantic analysis of policies. We develop a total of 24 advanced, in-depth queries that allow us to assess the evolution of content among the set of studied policies. We conduct this analysis by building on top of the *Polisis* framework (Sec. 6), a recent system developed for the automated analysis of privacy policies [12]. Following this approach, we perform a deeper level of semantic analysis that overcomes the limitations of keyword-based approaches.

C. Coverage (Sec. 7). We evaluate the policies' coverage of high-level privacy practices. For both *EU* and *Global* policies, we find a significant improvement in the policies' coverage of topics highly relevant to the GDPR such as data retention (52% improvement for *EU*, 31% improvement for *Global*), handling special audiences (16% improvement for *EU*, 12% improvement for *Global*), and user access (21% improvement for *EU*, 13% improvement for *Global*).

D. Compliance (Sec. 8). We design seven queries that codify several of the GDPR *compliance* metrics, namely those provided by the UK Information Commissioner (ICO). The GDPR's effect is evident; for both of our analyzed datasets, we find a positive trend in complying with the GDPR's clauses: substantially more companies improved (15.1% for *EU* policies and 10.66% for *Global* policies, on average) on these metrics compared to those that worsened (10.3% for *EU* policies and 7.12% for *Global* policies, on average).

E. Specificity (Sec. 9). Finally, we design eight queries capturing how specific policies are in describing their data practices. Our results show that providers are paying special attention to make the users aware of the specific data collection/sharing practices; 25.22% of *EU* policies and 19.4% of *Global* policies are more specific in describing their data practice. Other providers, however, are attempting to cover more practices in their policies at the expense of specificity; 22.7% of *EU* policies and 17.8% of *Global* policies are less specific than before.

Building on the above, we draw a final set of takeaways across both the time and geographical dimensions (Sec. 12).

2 GDPR Background

The European Union began their expedition into privacy legislation since the 1990s. The Data Protection Directive (95/46/EC) [11] set forth a set of principles for data protection as a preemptive attempt to safeguard EU constituents in a data-heavy future. This directive was further amended with the ePrivacy Directive (2002/58/EC) [10] to enhance the privacy of high-sensitivity data-types such as e-communications data. While the two directives maintained some focus on protecting user privacy, article 1, paragraph 2 of the Data Protection Directive very clearly states that "Member States shall neither restrict nor prohibit the free flow of personal data between Member States for reasons connected with the protection afforded under paragraph 1" [11]. Thus, while the set of principles intended to increase user privacy, the primary focus of the legislation was to prevent the inhibiting of personal data flow. Furthermore, the directives designated the Member State where a particular controller is located as the supervising authority. Naturally, the decentralized approach led to a wide variety of approaches to enforcement and penalization at the national level.

As the most comprehensive privacy regulation to date, the General Data Protection Regulation (Regulation (EU) 2016/679), passed on April 14, 2016 and enforced on May 25, 2018, is the European Union's attempt to create a united approach to online privacy. The GDPR defines

four entities: data subjects, data controllers, data processors, and third parties. The data subjects are the users of the information systems from which data is collected. The data controller is typically the service provider (e.g., website or mobile app) with a vested interest in receiving and processing the user data. A data controller might employ a processor to process the data on its behalf. Finally, the data controller might authorize a third party (e.g., analytics agency) to process some of the user's data.

Chapter III of the GDPR describes the rights of the data subjects; the first (Article 12) is the right to be informed about the service provider's privacy practices "in a concise, transparent, intelligible and easily accessible form, using clear and plain language." The service provider has to communicate its practices regarding data collection and sharing (Articles 13 and 14) as well as the rights of users associated with data collection and processing (Articles 15-22).

Under the GDPR, the service provider has to inform the user about the contact information of the controller, the purposes for data collection, the recipients of shared data, the retention period and the types of data collected. Furthermore, the service provider has to notify the users about updates to its privacy practices promptly. Articles 13 and 14 make a distinction between data collected directly from the user or obtained indirectly. The service providers have to inform the users about the source and type of information when obtained indirectly. Articles 15-22 list the rights of users regarding data collection and processing. These rights include the right of access, rights of rectification and erasure, right to the restriction of processing, right to data portability and right to object. In this paper, we focus on the requirements set in Articles 12-22 of the GDPR and study how privacy policies evolved in meeting these requirements.

3 Creation of the Policies Dataset

We assembled two datasets of privacy policies, one to represent EU-based privacy policies, and a second representing policies from around the world. Hereafter, we will refer to the former as the *EU* set and the latter as the *Global* set. For each policy, we selected two snapshots between January 2016 and May 2019: pre-GDPR and post-GDPR. We define the pre-GDPR snapshot as the last stable version policy before the enforcement of the GDPR. The post-GDPR is the most recent version of the policy. We restrict our analysis to privacy policies in the English Language.

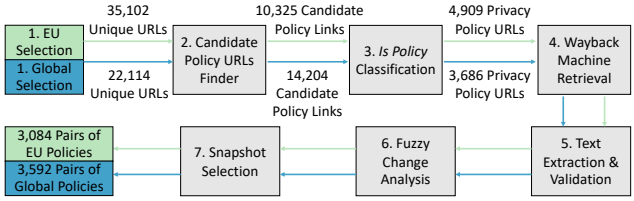


Fig. 1. Our methodology of retrieving the policy URLs.

3.1 Methodology

In the following, we describe our methodology to create a corpus of privacy policies as highlighted in Fig. 1.

Website Selection

To select the set of websites for the *EU* set, we used the Alexa TopSites API to query the top 3,000 domains for each of the 28 EU member-states. This step resulted in 82,389 URLs, as some of the smaller states, such as Malta, returned fewer results (2,394). Of the extracted domains, we found 35,102 unique instances. For the *Global* set, our methodology aimed at selecting websites that exhibit a topical and geographical mix. We used the Alexa Website Information Service to obtain the top links, ranked by popularity, in each of the 16 Alexa categories, spanning different topics (e.g., *Adult, Arts, Business, Regional*). We amended these categories by considering 9 subcategories of the *Regional* category (e.g., *North America, Middle East, Europe*). For each of the 25 categories, we considered the top 1,000 visited websites. This step resulted in a set of 25,000 URLs, of which we counted 22,114 unique URLs. We note that the starting *EU* set was significantly larger by design, as our restriction to English policies excludes a significant portion of the candidate *EU* domains.

Policy Finder

We automatically crawled the home page of each of the URLs identified in the previous stage. We crawled the HTML using the Selenium framework and a headless Chrome Browser. We identified a set of candidate privacy policy links on the home page based on regular expressions (e.g., the presence of words like *privacy, statement, notice,* or *policy* in the URL or the title). This stage resulted in candidate sets of 10,325 *EU* pages and 14,204 *Global* pages. In a lot of cases, initially distinct URLs share the same privacy policy link due to the same company owning multiple websites (e.g., YouTube owned by Google or Xbox owned by Microsoft).

Is-Policy Classification

We developed a custom Is-Policy classifier to decide whether each of the candidate pages belongs to a valid English-language privacy policy. The Is-Policy classifier consists of two modules. The first is a language detection module, using *langid.py* [24] that labels non-English websites as *invalid*. The second module is a one-layer Convolutional Neural Network (CNN) that we developed to output the probability that the input text belongs to a privacy policy (based on the classifier by Kim [14]). The accuracy of the classifier is 99.09% accuracy on the testing set. The details of the classifier's architecture and training are in Appendix A.

The Is-Policy classifier assigned a *valid* label (i.e., an English privacy policy) to 4,909 *EU* URLs, as well as 3,686 *Global* URLs. Besides language, additional reasons for a candidate URL's rejection included the domain's "robots.txt" policy, the lack of any privacy policy at the candidate link, and the privacy policy embedded within the longer terms of service. We considered these rejected cases to be non-suitable for our further automated analysis, prioritizing high precision over high recall.

Wayback Machine Retrieval

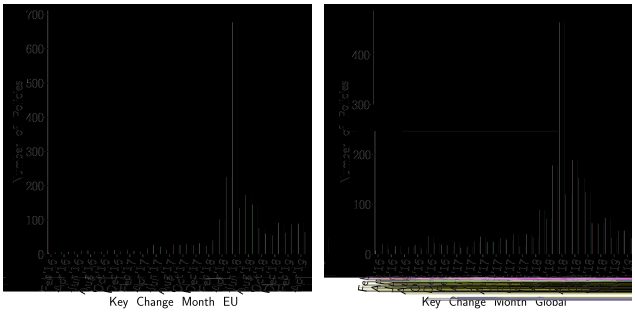
We used the Wayback Machine [13] from the Internet Archive to collect a series of raw HTML snapshots for each privacy policy from the last stage. We used the python library *Waybackpack*[31] to request every unique archived instance from Jan. 2016 to May 2019, with a monthly resolution (*i.e.* at most one instance was returned per month).

Text Extraction and Validation

For each downloaded HTML file of each policy, we used Boilerpipe [15] to get the cleaned HTML of the webpage without the unnecessary components (e.g., headers and footers). We extracted the body text using the library *Beautiful Soup*[30] . After these sanitization steps, a portion of the snapshots was reduced to very short text bodies and was removed from further processing. We performed further manual verification of a random set of the remaining snapshots to ensure that the false positives have been removed.

Analyzed Dataset

Next, we used the text of the valid snapshots for each policy to examine its change over time. We quantified the per-month change using a fuzzy string matching library [7] that employs Levenshtein Distance to calculate the distance between two strings. We defined changes in text between two snapshots to be *significant* if the similarity ratio between



(a) *EU* dataset. **(b)** *Global* dataset.

Fig. 2. The distribution of key-change dates for the privacy policies in our corpus.

the two text files was less than or equal to 95%. For both datasets, a portion of the policies exhibited no significant changes between 2016 and 2019. These policies have the same pre-GDPR and post-GDPR versions. For the rest, we define, for each policy, the *key-change date* as the closest month to the enforcement date of the GDPR that exhibited a significant change. The pre-GDPR snapshot of the policy is the first stable version of the policy before the key-change date with a timestamp preceding May 2018. This strategy accommodates the transient stage around the enforcement date of the GDPR by considering a stable version of the policy. We used the most recent snapshot (taken after May 2018) of each policy as the post-GDPR version since it captures any changes that may have occurred after the key-change date. The distribution of key-change dates for the *EU* and *Global* sets can be seen in Fig. 2a and Fig. 2b, respectively. The most frequent key-change date in both sets is June 2018, the month following the enforcement of the GDPR, as snapshots captured in this month contained the changes deployed in order to meet the deadline.

We removed policies that lacked a valid pre-GDPR or post-GDPR snapshot from further analysis (e.g., a policy was not indexed by the Wayback Machine before May 2018). This last pruning step resulted in the final *EU* dataset of 3,084 pairs of snapshots and a *Global* dataset of 3,592 pairs of snapshots. Between the two sets, 398 policy URLs are shared; these intersecting instances are globally popular sites (*e.g. Microsoft and Dictionary.com*), along with instances of websites added to the *Global* set as members of the "Europe" category (*e.g. Times of Malta and Munster Rugby*). Fig. 3 illustrates the cumulative distributions of fuzzy similarity between the pre-GDPR and post-GDPR instances for the *EU* and *Global* sets. As evident from the figure, the *EU* set demonstrates more signs of change during this period.

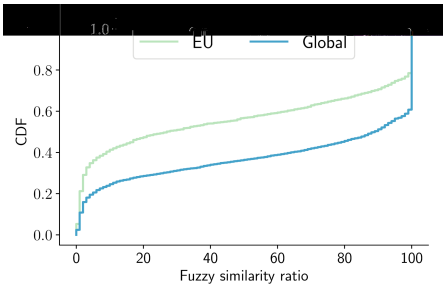


Fig. 3. Cumulative distribution of the fuzzy similarity metrics for the pre-GDPR and post-GDPR pairs of both datasets.

3.2 Takeaways

Our methodology highlights two takeaways regarding the initial impact of the GDPR on the privacy policy landscape. As expected, May 2018 was an inflection point; more than 45% of the policies that we studied have changed between March and July 2018. Second, the GDPR appears to have had a higher impact inside the EU than outside. The text content of the policies corresponding to EU websites has changed at a higher rate than their global counterparts.

4 Presentation Analysis

Our first step to understanding the evolution of the privacy policies is to test for changes in the presentation of the web pages through a user study. We followed a within-subjects study design, with two conditions: pre-GDPR and post-GDPR. Our goal was to have each participant evaluate how presentable a screenshot of a privacy policy is.

4.1 Hypothesis

Our null hypothesis for the user evaluation of privacy policies is that *there is no significant difference in users' perception of privacy policy appearance between the pairs of pre-GDPR and post-GDPR policies in our sample.* We reject the null hypothesis for $p < .05$.

4.2 Study Setup

We recruited 470 participants (so that each snapshot receives at least ten different evaluations) from Amazon Mechanical Turk. We chose participants who had >95% HIT approval rate and achieved masters status. We paid each respondent $1.45 to fill the survey that lasted 8 minutes on average. Out of the respondents, 48% were female, 46% had a Bachelors degree, and 26% did not have a degree. The average age of the respondents was 39 years. We did not ask for any personally identifiable information.

Study Material

We chose a random set of 200 unique privacy policies from the *EU* set and 200 unique privacy policies from the *Global* set (we excluded policies common between the two sets). Counting the pre-GDPR and post-GDPR versions, we obtained a total of 800 policies. We followed the approach used in previous studies around websites aesthetics [18, 29] to assess the presentability by using screenshots of the webpages instead of live versions. This approach avoids any bias from webpage loading times, internet speed or localized versions. We used the "webkit2png" tool to capture a full screenshot of each of the 800 privacy policies, which were all hosted by the Wayback Machine. As these 800 screenshots included the full policy scrolled to the bottom, we cropped 612×1028 pixels from the text body of each screenshot to display for the respondents. Two of the authors manually inspected each of the images and corrected any screenshot that was incorrectly captured.

Survey Design

We presented each respondent with a random set of 20 screenshots from the total set of 800 images. The image order was randomized per participant to compensate for the effects of learning and fatigue. The respondents were not primed about the comparative nature of the study. We explicitly asked the respondents not to read the content of each screenshot, but instead to give their assessment over the images' look and feel. For each screenshot, the respondents indicated how much they agree/disagree with a set of three statements over a 5-point Likert scale (Strongly Disagree(SD), Disagree (D), Neither (N), Agree (A), Strongly Agree (SA)). A snapshot of the survey is available in Fig. 13 of Appendix B. These statements, consistent with the usability measurement questions in [22], were as follows:

s1. This policy has an attractive appearance.
s2. This policy has a clean and simple presentation.
s3. This policy creates a positive experience for me.

Additionally, we placed two anchor questions that contain poorly formatted "lorem ipsum" text, using them to filter out respondents with low-quality answers. At the survey end, the respondents filled an optional demographics survey. After filtering out responses failing the quality-assurance questions, we analyzed only those images with at least five evaluations. The resulting image set was composed of 139 pairs of *EU* policies and 136 pairs of *Global* policies. On average, each screenshot received 7.8 evaluations.

Table 1. The resulting *EU* scores for the user study grouped by question and study condition.

Stmt	Condition	Disagree (%)	Neutral (%)	Agree (%)
s1	pre-GDPR	32.5	19.8	47.7
	post-GDPR	28.2	20.7	51.1
s2	pre-GDPR	24.8	16.1	59.2
	post-GDPR	20.9	16.6	62.5
s3	pre-GDPR	29.5	26.4	44.0
	post-GDPR	25.5	28.0	46.5

Table 2. The resulting *Global* scores for the user study grouped by question and study condition.

Stmt	Condition	Disagree (%)	Neutral (%)	Agree (%)
s1	pre-GDPR	35.0	20.4	44.6
	post-GDPR	33.6	20.0	46.4
s2	pre-GDPR	26.1	18.4	55.5
	post-GDPR	25.5	16.2	58.3
s3	pre-GDPR	32.5	27.2	40.2
	post-GDPR	32.1	29.2	38.7

4.3 Findings

The distribution of scores can be seen in Tables 1 and 2. Because we want to compare users' overall perception of policy interfaces, we group scores into three categories: disagree (Strongly Disagree, Disagree), neutral (Neither Agree nor Disagree), and agree (Agree, Strongly Agree). Given three discrete outcomes, we apply the Chi-Squared Test between the pre-GDPR and post-GDPR instances' distributions. For the *EU* set, we reject the null hypothesis for $p=0.05$ for *s1* (0.039) and *s2* (0.040), but fail to reject the null hypothesis for *s3* (0.057). On the other hand, for the *Global* set, we fail to reject the null hypothesis for $p=0.05$ for all questions, with all three scores having $p>0.4$. This result suggests that *EU* policies improved their visual interfaces to be more attractive and simplified to further their support of the GDPR's call for clarity in policies.

Another planned comparison was to investigate how the pre-GDPR *EU* policies compare to their *Global* counterparts as well as how the post-GDPR sets compare. We group the scores into the same three categories as before, and once again apply the Chi-Squared Test. We fail to reject the null hypothesis for $p=.05$ for all three questions when comparing the pre-GDPR sets; however, we reject the null-hypothesis for *s1* ($p=0.007$), *s2* ($p=0.014$), and *s3* ($p=9.06e-5$) when comparing the post-GDPR images of the two sets. Looking at the post-GDPR distributions in Table 1 and Table 2, the observable difference between the two sets of scores suggests that the *EU* policies created a notably more positive experience for users compared to the *Global* policies.

4.4 Takeaways

The major takeaway from this study is that the GDPR was a driver for enhancing the appearance and the presentation of the top privacy policies in the EU countries. We were not able to conclusively observe such an improvement for the *Global* set of websites. Compared to the *Global* set, the *EU* websites have improved the appearance, presentation, and experience of their policies.

5 Text-Feature Analysis

We complement our initial study of the visual features' evolution in the privacy policies with an assessment of syntactic textual changes. These features allow us to get a high-level understanding of how the policy's structure evolved before we go deeper into semantic features in the later sections.

5.1 Hypothesis

We applied five common text metrics to the documents that describe the length and sentence structure of the text. For each test we conduct: we evaluate the null-hypothesis that *there does not exist a significant change in the tested text-feature between the pre-GDPR and post-GDPR instances of a policy.* Since we run five tests on the same policy samples, we apply the Bonferroni correction [27]. Below, the null hypothesis is rejected when the p value of the test is less than $\frac{0.05}{5}$.

5.2 Text Metrics

We consider the following metrics:

1. **Syllables Count:** gives the total number of syllables available in the text.
2. **Word Count:** gives the number of words available in the text.
3. **Sentence Count:** gives the number of sentences present in a text.
4. **Word per Sentence:** gives the average number of words per sentence in the text.
5. **Passive Voice Index:** gives the percentage of sentences that contain passive verb forms. To compute this score, we tokenize the text into sentences and perform dependency parsing on each sentence using the

Table 3. EU Text Metrics results for pre-GDPR and post-GDPR instances.

Metric	pre-GDPR μ (σ)	post-GDPR μ (σ)	p
#Syllables	3534.18 (4580.4)	4670.45 (5291.8)	9.92e-143
#Words	1936.15 (2010.3)	2621.38 (2564.9)	3.05e-148
#Sentences	53.49 (53.35)	71.36 (69.62)	1.02e-100
#Words/Sent.	53.23 (121.10)	50.47 (144.46)	2.84e-06
#Passive	10.75 (7.35)	10.90 (6.89)	1.85e-01

Table 4. Global Text Metrics results for pre-GDPR and post-GDPR instances.

Metric	pre-GDPR μ (σ)	post-GDPR μ (σ)	p
#Syllables	2977.43 (2941.5)	4108.18 (6606.6)	6.36e-136
#Words	1709.45 (1609.2)	2140.44 (2167.4)	6.16e-141
#Sentences	48.04 (42.02)	58.61 (54.42)	1.52e-86
#Words/Sent.	42.47 (47.61)	43.76 (54.51)	4.14e-04
#Passive	11.45 (7.21)	11.52 (6.99)	4.29e-01

Spacy library [2] . We consider a sentence to contain a passive voice if it follows the pattern of: *nsubjpass* (that is Nominal subject (passive)), followed by *aux* (Auxiliary), and then followed by *auxpass* (Auxiliary (passive)). This pattern would match sentences similar to *"Data is collected."*

5.3 Findings

We compute the value of each text metric for the pre-GDPR and post-GDPR versions of each policy. Given the high variability in the text content of policies, we avoid using statistics requiring normal-distributions. Instead, we use the Wilcoxon signed rank test to compare the pre-GDPR and post-GDPR pairs of text for each metric, correcting for multiple comparisons. Table 3 and Table 4 describe the key statistics of the study for the *EU* and *Global* sets, respectively.

For the *EU* set, we reject the null hypothesis for the changes in the number of syllables (+50%), the number of words (+35%), and the number of sentences (+33%). The magnitude of these changes indicates that policies in this set are becoming much longer. These results are consistent with the expected effects of adopting the GDPR. *EU* policies must cover more material than before to bring transparency to data subjects.

For the *Global* set, there has been a statistically significant increase in the number of syllables (+38%), the number of words (+25%), and the number of sentences (+21%). This increase in the size of policies outside the EU suggests that the GDPR has significantly impacted the global landscape of privacy policies. Still, the magnitude of this increase is noticeably smaller compared to the *EU* set. Finally, we fail to reject the null hypothesis for the passive voice index as with the *EU* set.

5.4 Takeaways

Consistent with recent findings [20, 23], we find privacy policies to have increased in length considerably, with *EU* policies being longer than *Global* policies. We, however, find that this increase did not accompany sentence structure improvements. The passive voice index did not exhibit any change in the post-GDPR versions for both the datasets. Further, while there are statistically significant changes in the number of words per sentence, these changes are too small (-5% for *EU* and +3% for *Global*) to conclude a change in the sentence structure of the policies. On average, a privacy policy sentence spans more than 43 words for both datasets.

6 Automated Requirement Analysis

While the text-feature analysis captures whether the privacy policies have syntactically changed, the metrics are domain-agnostic. Hence, we follow a different methodology to delve deeper and assess the coverage, the compliance, and the specificity angles in the following sections.

6.1 Methodology Overview

Our methodology starts by defining a set of goals or requirements. These goals are high-level and are independent of the implementation methodology. Then, we code these goals by extending a technique called *structured querying* over privacy policies, introduced by Harkous et al. [12]. This technique builds two levels of abstraction on top of the raw privacy policy as demonstrated in Fig. 4.

On the first level, a privacy policy is converted from a set of text segments written in natural language to a set of automatic labels that describe the embedded privacy practices. For instance, one label can indicate that the segment discusses "third-party" sharing and another can indicate that the sharing purpose is "advertising". These labels are assigned by a set of machine-learning text classifiers trained on human annotations.

On the second level, a first-order logic query is constructed to reason about the set of labels across segments. For example, if one were to query the level of specificity of a policy describing the purpose of data collection, they would list the segments that have a purpose label. Then,

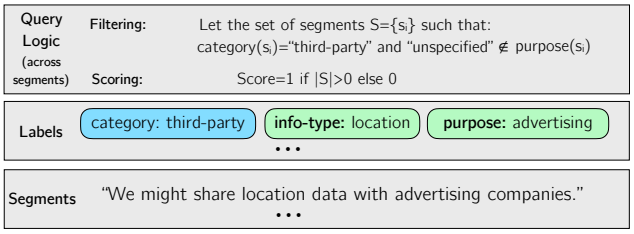


Fig. 4. Structured querying with two-levels of abstraction on top of text

they would count how many of these segments do not have the purpose label equal to "unspecified." We further formalize the querying logic in this work by splitting the querying into two steps. The first step is **filtering**, where the first-order logic query is used to filter the total set of segments into a subset of relevant segments. The second step is **scoring**, where a scoring function is computed based on the relevant segments. We provide an example in Fig. 4, where the filtering step is used to decide on the subset of segments discussing third-party sharing with specified purposes and the scoring step assigns a score of 1 if this subset is not empty. We follow the spirit of this **filtering-scoring** approach for our in-depth analysis in the following sections.

This structured querying technique offers an advantage over approaches based on heuristics and keyword analysis (e.g., [4, 9]); it allows us to better cover text with varying wordings but similar semantics. Further, this technique avoids the shortcomings of the other approaches that directly use machine learning to quantify the goals (e.g., [8, 16, 19]); it is more flexible for adapting the goals (i.e., queries) as needed, without having to create new labeling data for each new goal.

In this work, we are the first to conduct a comprehensive analysis of privacy-related goals using structured querying. Our main contributions lie in the goals' definition, the translation of these goals into queries, the volume of goals we measure, and the comparative nature of this study.

6.2 Polisis

We utilize the Polisis system described by Harkous et al [12] to generate the automated labels described within structured querying. Polisis pre-processes a privacy policy and breaks it into a set of smaller segments (one example of such a segment is in Fig. 4). A segment is a set of consecutive sentences of a privacy policy that are semantically coherent. Polisis passes each segment through a set of classifiers to assign automatic labels describing the embedded privacy practices. These classifiers have been trained on the OPP-115 dataset created by Wilson et al. [36]. The dataset consists of 115 privacy policies (267K words) with

Table 5. Description of the relevant high-level privacy categories from Wilson et al. [36].

Privacy Category	Description
First Party Collection /Use	Service provider's collection and use of user data.
Third Party Sharing /Collection	Sharing and collection of user data with third parties (e.g., advertisers)
User Choice/Control	Options for choices and control users have for their collected data
International & Specific Audiences	Practices related to a specific group of users (e.g., children, Europeans).
Data Security	The protection mechanisms for user's data.
User Access, Edit, & Deletion	Options for users to access, edit or delete their stored data.
Data Retention	The period and purpose of storing user's data.
Policy Change	Communicating changes to the privacy policy to the users.
Privacy Contact Info	The contact information for privacy related matters.

manual annotations for 23K fine-grained data practices. The privacy-policy taxonomy used for the annotation task is depicted in (Fig. 5).

Polisis labels each segment with high-level privacy categories (blue labels in Fig. 4 prefixed with "category:") as well as values for lower-level privacy attributes (green labels in Fig. 4 prefixed with "info-type:" and "purpose:"). In particular, Polisis assigns a segment s, of a privacy policy, a set: $category(s)$. This set is a subset of the nine high-level privacy categories which are dark shaded in Fig. 5. Also, Polisis labels each segment with a set of values, corresponding to 20 lower-level privacy attributes (light-shaded in Fig. 5). The values corresponding to each attribute are shown as tables in Fig. 5. For example, the attribute "purpose" indicates the purposes of data processing and is represented by the set $purpose(s)$.

If $category(s) = \{first\text{-}party\text{-}collection\text{-}use\}$ and $purpose(s) = \{basic\text{-}feature, personalization, marketing\}$, we conclude that the segment s describes multiple purposes for first party data collection, which are to provide basic features, personalize the service, and use data for marketing. In addition to the labels, Polisis returns a probability measure associated with each label. Elements of the sets mentioned above are the ones classified with a probability larger than 0.5.

Following guidelines from the developers of Polisis, we retrained its classifiers so that they can execute locally. We refer the reader to the work of Harkous et al., [12] for a breakdown of the accuracy of the high-level and each of the low-level classifiers. Also, a detailed description of all the taxonomy attributes and their values is present within the OPP-115 dataset (https://usableprivacy.org/data).

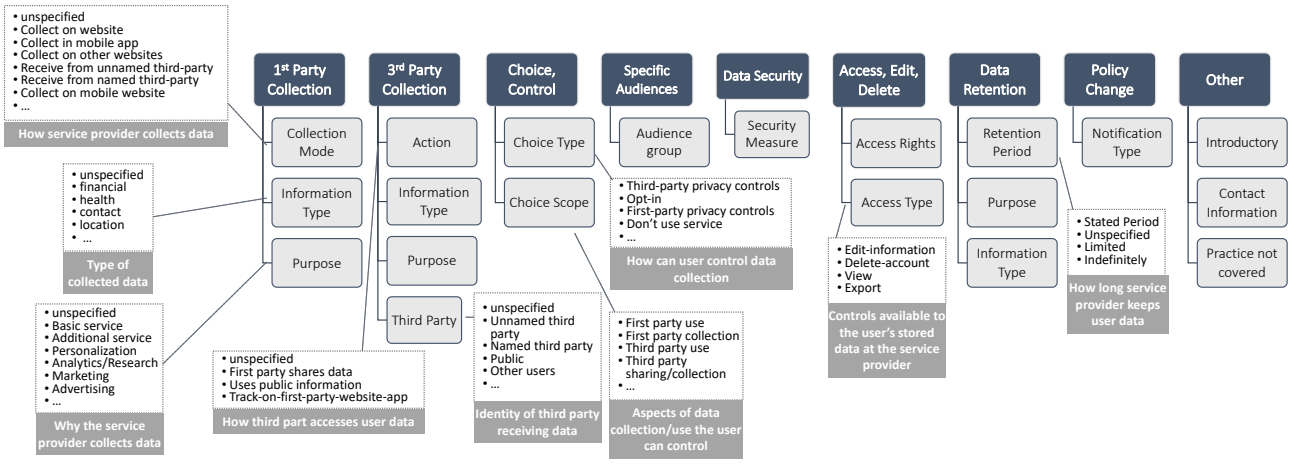


Fig. 5. The privacy taxonomy of Wilson *et al.* [36]. The top level of the hierarchy (darkly shaded blocks) defines high-level privacy categories. The lower level defines a set of privacy attributes (light shaded blocks), each assuming a set of values. We show examples of values for some of the attributes. The taxonomy has more attributes that we do not show for space considerations.

7 Coverage Analysis

As the first step in the context-based analysis of privacy policies, we study the policies' coverage of the high-level privacy categories described in Table 5. This analysis highlights the difference in category coverage between pre-GDPR and post-GDPR policies. A policy covers a specific category if it contains at least one segment with this category as a label, according to Polisis.

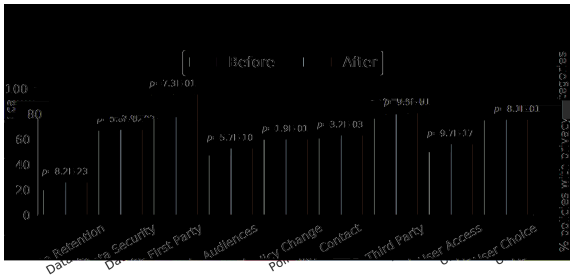


Fig. 6. Category Coverage for *EU* policies' pre-GDPR and post-GDPR instances; p values are derived from applying the Chi-Square test.

Figures 6 and 7 display the fraction of *EU* and *Global* policies with coverage scores of 1 for each of the high-level privacy categories pre-GDPR and post-GDPR. For this analysis, we consider the hypothesis that *there is no significant change in the category coverage of privacy policies between pre-GDPR and post-GDPR instances*. We use the Chi-Squared test to evaluate this hypothesis. However, since we are doing multiple tests on the same labeled data set, we apply Bonferroni correction [27] and reject the null hypothesis only when the p value is less than $\frac{.05}{9}$. The p values for each category are shown in Fig. 6 and Fig. 7 for the *EU* and *Global* sets, respectively.

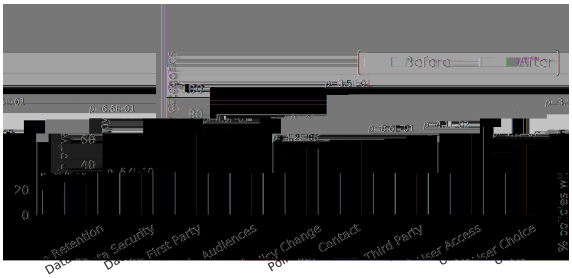


Fig. 7. Category Coverage for *Global* policies' pre-GDPR and post-GDPR instances; p values are derived from applying the Chi-Square test.

We observe that categories already exhibiting high coverage did not have a major change between the pre-GDPR and the post-GDPR sets for both the *EU* and the *Global* cases. On the other hand, categories that were underrepresented in the pre-GDPR set showed significant change. For the *EU* set, we observe statistically significant improvement in four of the nine categories ('Data Retention', 'International & Specific Audiences', 'Privacy Contact Information', and 'User Access, Edit & Deletion'). For the *Global* set, we observe statistically significant improvement in three of the nine categories (Data Retention, International & Specific Audiences, and User Access, Edit & Deletion).

7.1 Comparison to Manual Annotations

As mentioned in Sec. 6, this context-based analysis uses Polisis [12] as the core querying engine. A question that comes to mind is: *How well does Polisis work with these*

queries? To better understand how Polisis behaves at the query level, we decided to compare its results with Subject Matter Experts (SMEs). We leverage the raw annotations of the OPP-115 dataset [36] in which three SMEs have annotated each policy.

After consulting with the authors of Polisis, we selected a set of 15 policies outside the training and validation sets of Polisis. We then generated the automatic labels for these policies. Therefore, we ended up with four annotations of the same policy; three by SMEs and one by Polisis. We then passed these annotations through the querying engine to get the query results. Next, we computed the level of disagreement in these results among SMEs themselves and between Polisis and SMEs. The disagreement rate per policy is quantified by the ratio of (the number of queries with different scores between the two groups) to (the total number of queries). We then average this rate across policies[3].

We find that the disagreement rate for Polisis-SMEs was 0.10, which is only slightly worse than the SME-SME disagreement rate of 0.07. This observation indicates that we can rely on Polisis' annotations as a reasonable proxy to what human annotators produce. Note that the disagreement rate is not equivalent to the error rate as the latter assumes the existence of ground truth.

7.2 Takeaways

It is evident from Fig. 6 and Fig. 7 that the GDPR has had a positive effect on the coverage of categories. Traditionally, privacy policies covered clauses mostly related to first party collection and third party sharing. With the introduction of the GDPR, it is clear that there is a trend of covering additional categories of particular importance to the GDPR requirements, including data retention periods (Article 13(2.a)), notices to special audiences, safeguarding the user data, and providing the users with the options to access and rectify their information (c.f. Sec. 2). Interestingly, the improvement in coverage of *EU* policies for these categories is higher than that of *Global* policies. Finally, our manual analysis with subject matter experts shows that automatically extracting these results did not hinder their quality.

8 Compliance Analysis

Next, we study the content of the policies in light of the compliance requirements introduced by the GDPR. We rely on the UK's Information Commissioner's officer's (ICO) guide to the GDPR [**?**] , which contains a set of guidelines for organizations to meet the provisions set in the GDPR. In the form of a checklist for organizations to inform users of their rights, the ICO guide provides an official and structured interpretation of the GDPR. It obviates the need for our customized interpretation of the law. We translate these requirements via the filtering-scoring approach of Sec. 6 in a way that allows us to compare the privacy practices of the service providers before and after the introduction of the GDPR. Table 6 shows the ICO checklist items, their descriptions, and their corresponding filtering and scoring logic.

Since the taxonomy employed in Polisis precedes the GDPR, some of the items in the ICO's checklists are incompatible with the framework, i.e., they cover newer concepts and are not quantifiable in the current framework. We considered only those items from the ICO framework that are compatible with Polisis taxonomy. We manually compared these covered items to the GDPR text to find them representative of articles 12-20. We consider the compliance evidence for all the checklist items as a binary metric denoted by the existence of a segment satisfying the associated clause.

To assess the change of compliance for each policy, according to these seven requirements, we compare the scores of pre-GDPR and post-GDPR versions by breaking down the change for each requirement into four cases:

- **Requirement Worsened:** A policy used to cover the requirement in the pre-GDPR version, but does not cover it in the post-GDPR version; i.e., score drops from 1 to 0.
- **Requirement Still Missing:** A policy did not cover the requirement in either the pre-GDPR or post-GDPR snapshots; i.e., the score is 0 for both versions.
- **Requirement Still Covered:** A policy continues to cover the requirement in both the versions; i.e., the score is 1 for both versions.
- **Requirement Improved:** A policy did not cover the requirement in the pre-GDPR version, but does cover it in the post-GDPR version; i.e., the score rises from 0 to 1.

8.1 Findings

Fig. 8 and 9 show the percentage of *EU* and *Global* policies falling into each of the four cases for each ICO requirement. Similar to the methodology described in Sec. 7.1, we computed the disagreement rate between queries built on automated labels of Polisis and those built on annotations of Subject Matter Experts (SMEs) from the OPP-115 dataset [36]. The disagreement rate averaged across 15 policies for Polisis-SMEs was found to be 0.21, which is

3 We use this metric vs. Cohen's Kappa or Krippendorff's alpha as the latter does not apply for when the three annotators are not the same across policies, which was the case with the OPP-115 dataset.

Table 6. The list of the queries derived from ICO's GDPR checklists. ICO-Q1 – ICO-Q7 are from the "Right to be Informed" checklist. ICO-Q8 is from the "Right of Access" checklist. $S_{actions}$= {collect-from-user-on-other-websites, receive-from-other-parts-of-company-affiliates, receive-from-other-service-third-party-named, receive-from-other-service-third-party-unnamed, track-user-on-other-websites }

ICO Checklist Item	GDPR Ref.	Filtering Logic	Scoring Func.
ICO-Q1: "The purposes of processing user data."	13(1.c)	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{first\text{-}party\}$ $purpose(s_i) \neq \phi$ and $unspecified \notin purpose(s_i)$	Score= 1 if $\lvert S\rvert > 0$ else 0
ICO-Q2: "The categories of obtained personal data (if personal data is not obtained from the individual it relates to)."	14(1.d)	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{first\text{-}party\}$ $action\text{-}first\text{-}party(s_i) \subset S_{actions}$ $unspecified \notin info\text{-}type(s_i)$	Score= 1 if $\lvert S\rvert > 0$ else 0
ICO-Q3: "The recipients of the user's personal data."	14(1.e)	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{third\text{-}party\}$ $unspecified \notin third\text{-}party\text{-}entity(s_i)$	Score= 1 if $\lvert S\rvert > 0$ else 0
ICO-Q4: "The right for the user to withdraw consent from data processing."	17(1.b)	Consider the set $S=\{s_i\}$ such that $category(s_i) \in \{first\text{-}party, user\text{-}choice\text{-}control\}$ $choice\text{-}type(s_i) = \{$op-out-link, op-out-via-contacting-company$\}$ $choice\text{-}scope(s_i) = \{$first-party-use$\}$	Score= 1 if $\lvert S\rvert > 0$ else 0
ICO-Q5: "The source of the personal data (if the personal data is not obtained from the individual it relates to)."	15(1.g)	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{first\text{-}party\}$ $action\text{-}first\text{-}party(s_i) \subset S_{actions}$	Score= 1 if $\lvert S\rvert > 0$ else 0
ICO-Q6: "If we plan to use personal data for a new purpose, we update our privacy information and communicate the changes to individuals before starting any new processing."	13(3)	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{policy\text{-}change\}$ $type\text{-}of\text{-}policy\text{-}change(s_i) = \{$privacy-relevant-change$\}$ $unspecified \notin how\text{-}notified(s_i)$	Score= 1 if $\lvert S\rvert > 0$ else 0
ICO-Q7: "Individuals have the right to access their personal data."	15(1)	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{user\text{-}access\text{-}edit\text{-}deletion\}$ $access\text{-}type(s_i) \in \{$view, export, edit-information$\}$	Score= 1 if $\lvert S\rvert > 0$ else 0

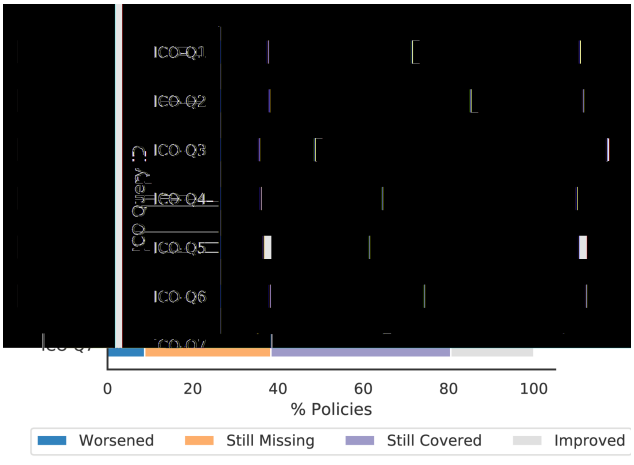


Fig. 8. The comparison of ICO scores of *EU* policies' pre-GDPR and post-GDPR instances. The queries for the ICO checklist can be found in Table 6.

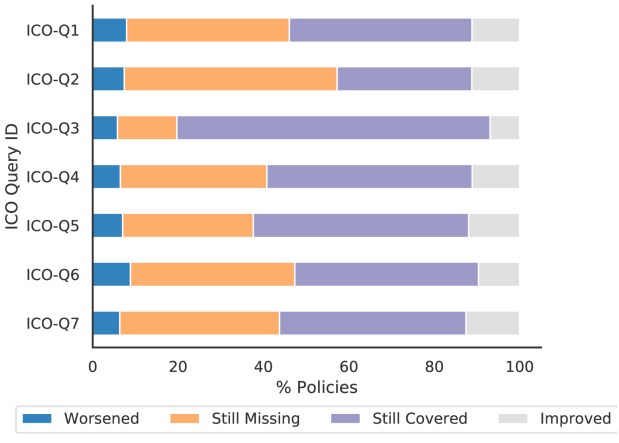


Fig. 9. The comparison of ICO scores of *Global* policies' pre-GDPR and post-GDPR instances. The queries for the ICO checklist can be found in Table 6.

comparable to the SME-SME disagreement rate of 0.19. We note that the disagreement rate is higher than that of coverage because of the added level of classification involved in the ICO queries.

Case 1: Requirements Worsened

For the vast majority of the policy sets, there has not been a noticeable decline in the covered ICO requirements. This decline has been limited to less than 12% for both sets

across all queries. We observe that ICO-Q6, referring to updating privacy policies if the site uses personal data for new purposes, exhibits the highest decline (11.7%) among the seven requirements for both the *EU* and *Global* sets. An example of a policy with a worse score is Only Fans, whose pre-GDPR policy contained the segment: "We post any changes we make to the Policy on this page. If we make material changes to how we treat our Users' personal information, we will notify you.", had no similar statement in its post-GDPR instance.

Case 2: Requirements Still Missing

Despite the GDPR's emphasis on the right of the users to be informed, we find that many privacy policies are still lacking. For instance, when it comes to specifying the sources of personal data obtained by first parties, there is still a large gap to fill. This is captured in ICO-Q2 for which we could not find evidence in 47.1% of *EU* policies and 49.8% of *Global*. Although companies are not always obtaining such data from third parties, these numbers are large enough to raise concerns about compliance with the ICO requirements.

One exception is ICO-Q3, which is focused on specifying the third party entities **receiving** the data (only 13.0% of *EU* policies and 13.9% of *Global* policies maintained a score of 0 in their pre-GDPR and post-GDPR versions).

Case 3: Requirements Still Covered

From Fig. 8 & 9, we observe that except for ICO-Q2, at least 38% of policies from both sets were, and are still compliant with the ICO requirements. This suggests that the policies are still being forthcoming with their practices, especially for ICO-Q3, 4, 5, and 7, which require the policy to talk about the source and recipients of collected data and users' rights to withdraw and access their data.

Case 4: Requirements Improved

Finally, a considerable portion of both policy sets, on average about 15% for *EU* and 10.7% for *Global*, have improved their coverage of the ICO requirements. ICO-Q7, informing users of their right to access their personal data, has seen the highest improvement (19.5% for *EU* and 12.6% for *Global*) among the policies in our datasets. We observe that only a minority of the policies (mostly from the more popular websites) have started addressing the previously missing requirements (such as ICO-Q4, ICO-Q6). For example, NYTimes has added a new clause to address the requirement about notifying users about changes to the privacy policies:

"We evaluate this privacy policy periodically in light of changing business practices, technology, and legal requirements. As a result, it is updated from time to time. Any such changes will be posted on this page. If we make a significant or material change in the way we use or share your personal information, you will be notified via email and/or prominent notice within the NYT Services at least 30 days prior to the changes taking effect."

8.2 Takeaways

Similar to the coverage analysis, we find that the majority of the privacy policies, both inside and outside the EU, have incorporated the GDPR's privacy requirements we analyzed. On average, 59.3% of the *EU* post-GDPR policies and 58.2% of the *Global* post-GDPR policies meet our seven queries (combining Case 3 and Case 4). We also notice that policies state the recipients of user data (77.8% *EU*, 80.3% *Global*) and the collection sources of users' personal data (65% *EU*, 62.4% *Global*) fairly well across both sets of policies. On the other hand, both sets struggle in describing the categories of the collected data (41.3% *EU* and 42.8% *Global*); for our compliance analysis, only this single query had less than 50% of the post-GDPR policies comply.

Comparing the two sets, we observe that the average score of compliance is higher for the *EU* policies. Also, the average margin of improvement (from pre-GDPR to post-GDPR) across the chosen ICO requirements for the *EU* set (4.8%) is larger than that of the *Global* set (3.5%).

9 Specificity Analysis

Compliance and coverage describe whether a policy mentions a privacy practice. Merely mentioning a privacy practice, however, does not fully satisfy transparency; it is not clear whether these practices are covered in general or specific terms. We quantify this aspect through specificity queries. For example, the statement "We collect your personal information . . . " covers collection by the first party but is not specific as to which type of personal information is being collected; a specific statement would be "We collect your health data . . . " In this section, we aim to assess the change in the level of specificity present in the privacy policies.

We use the filtering-scoring approach of Sec. 6 to quantify the policy's specificity of describing a privacy practice. Table 7 describes the eight specificity queries (Q1 $\rightarrow$ Q8) that quantify how explicit the privacy policy is in describing: how first party is collecting data, how third-party is obtaining data, the information type collected, information type shared, purposes for data collection, purposes for data sharing, and purposes for data retention. For all the queries, a higher score indicates higher specificity.

The reader might notice a discrepancy in Table 7 in the scoring step for queries focusing on the *purpose* attribute (first party (Q6) and third party (Q7)) vs. the rest of the

Table 7. Table of the specificity queries applied to each pre-GDPR and post-GDPR policy. Note the separate scoring Function for Q6 and Q7; for these queries, $purpose(s_i)$ is the set of all purposes seen for the current policy being analyzed.

Description	Filtering Logic	Scoring Function
Q1: Quantify how specifically the policy indicates how the first party is obtaining user data.	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{first\text{-}party\}$ $action\text{-}first\text{-}party(s_i) \neq \phi$	Take $S_a \subset S$ such that $action\text{-}first\text{-}party(s_i) = unspecified$ The specificity score is: $1-\lvert S_a\rvert/\lvert S\rvert$.
Q2: Quantify how specifically the policy indicates how the third party is collecting user data.	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{third\text{-}party\}$ $action\text{-}third\text{-}party(s_i) \neq \phi$	Take $S_a \subset S$ such that $action\text{-}third\text{-}party(s_i) = unspecified$ The specificity score is: $1-\lvert S_a\rvert/\lvert S\rvert$.
Q3: Quantify how specifically the policy indicates the type of information accessed by the first party.	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{first\text{-}party\}$ $info\text{-}type(s_i) \neq \phi$	Take $S_a \subset S$ such that $info\text{-}type(s_i) = unspecified$ The specificity score is: $1-\lvert S_a\rvert/\lvert S\rvert$.
Q4: Quantify how specifically the policy indicates the type of information shared with the third party.	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{third\text{-}party\}$ $info\text{-}type(s_i) \neq \phi$	Take $S_a \subset S$ such that $info\text{-}type(s_i) = unspecified$ The specificity score is: $1-\lvert S_a\rvert/\lvert S\rvert$.
Q5: Quantify how specifically the policy indicates how the third party is receiving user information.	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{third\text{-}party\}$	Take $S_a \subset S$ such that $third\text{-}party\text{-}entity(s_i) = unspecified$ The specificity score is: $1-\lvert S_a\rvert/\lvert S\rvert$.
Q6: Quantify how specifically the policy covers first party collection purposes relative to all possible purposes in our taxonomy.	Let P be the set of all purposes. Let P_s be the set of all purposes p such that $\exists$ a segment s_i with: $category(s_i) = \{first\text{-}party\}$ $p \in purpose(s_i)$	The specificity score is $\lvert P_s\rvert/\lvert P\rvert$.
Q7: Quantify how specifically the policy covers third party sharing purposes relative to all possible purposes in our taxonomy.	Let P be the set of all purposes. Let P_s be the set of all purposes p such that $\exists$ a segment s_i with: $category(s_i) = \{third\text{-}party\}$ $p \in purpose(s_i)$	The specificity score is $\lvert P_s\rvert/\lvert P\rvert$.
Q8: Quantify how specifically the policy indicates the purpose for data retention.	Consider the set $S=\{s_i\}$ such that $category(s_i) = \{data\text{-}retention\}$ $purpose(s_i) \neq \phi$	Take $S_a \subset S$ such that $purpose(s_i) = unspecified$ The specificity score is: $1-\lvert S_a\rvert/\lvert S\rvert$.

queries. We treated these cases differently due to the way Polisis interprets privacy policies. Within the Polisis system, $purpose(s) = unspecified$ does not always imply a missing purpose of data collection/sharing. Instead, it might indicate that the purpose is not the subject of that segment. Hence, most of the segments that focus on the data types being collected or shared will carry an *unspecified* purpose label. Accordingly, we quantify purpose specificity, in the first party and third party contexts, as the ratio of the number of stated purposes in the policy ($|P_s|$) to the total number of possible purposes ($|P|$). On the other hand data retention is typically addressed by one or two segments in the policy; the segments almost always describe the purpose. If those segments have $purpose(s) = unspecified$, then we expect that the policy is not being specific in explaining the purpose for data retention.

We analyze the evolution of the eight specificity scores between each policy's pre-GDPR and post-GDPR instances. We also note that manual verification for specificity queries is not suitable here because segments in Polisis are different than the segments in the OPP-115 dataset, and the specificity scores depend heavily on the number of segments with a particular label. We consider five cases in Fig. 10 and Fig. 11:

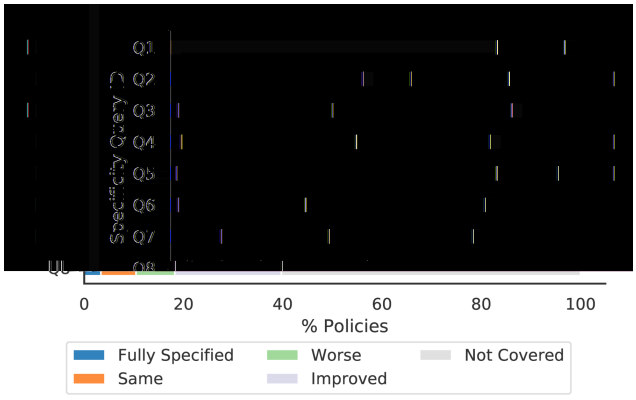


Fig. 10. The comparison specificity scores of *EU* policies' pre-GDPR and post-GDPR instances.

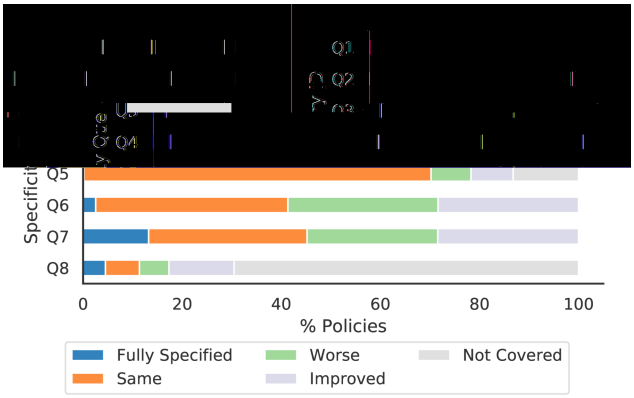


Fig. 11. The comparison specificity scores of *Global* policies' pre-GDPR and post-GDPR instances.

- **Query Not Covered:** A policy did not cover the requirement in both its pre-GDPR and post-GDPR versions; i.e. $|S|=0$ or $|P|=0$ for both the versions.
- **Same Specificity:** A policy maintains the same specificity level about a practice between the two versions, but the score is not equal to 1.
- **Fully Specified:** A policy continues to be fully specific about a practice; i.e. the score $=1$ for both the versions.
- **Worse Specificity:** The post-GDPR version of the policy has a lower specificity score than its pre-GDPR instance counterpart.
- **Improved Specificity:** The post-GDPR version of the policy has a higher specificity score than its pre-GDPR instance counterpart.

9.1 Findings

Case 1: Query Not Covered

Consistent with the results of Sec. 7, we find that the purpose of data retention practice (Q8) is not frequently covered among both sets of studied policies. Also, we find that approximately 10.6% of *EU* policies and 13.2% of *Global* policies do not cover the queries Q2, Q4, and Q5 because they do not cover the third party sharing category. This observation is relatively consistent with the results of Sec. 7 where around 20% of the pre-GDPR and post-GDPR instances of policies do not cover this category.

Case 2: Same specificity

A large portion of the policies in both sets exhibited the same specificity scores for the analyzed privacy practices, particularly for queries Q1 and Q5. This result is not surprising given that about 20% of *EU* policies and 35% of *Global* policies did not change between pre-GDPR and post-GDPR instance, as seen in Fig. 3. For the other policies, they maintain the same specificity levels even when their content changes.

Case 3: Fully Specified

For the privacy practices covered in Q2, the specificity values stay at one for 40% (and to a lower degree Q7) of the studied policies for both datasets. These subsets of policies mention the specific methods of collecting user and sharing data. We also observe that the portion of fully specified policies is fairly consistent between the *EU* and *Global* sets for all queries.

Case 4: Worse Specificity

Interestingly, we observe a considerable portion of policies exhibiting lower specificity in describing their privacy practices. We attribute this reason to the policies trying to be more comprehensive and general in describing the data practices at the expense of the specificity of the data practices clauses. This is a representative example from the post-GDPR snapshot of hrc.org [6] :

"We also may be required to release information if required to do so by law or if, in our sole discretion, such disclosure is reasonably necessary to comply with legal process or to protect the rights, property, or personal safety of our web site, HRC, our officers, directors, employees, members, other users, and/or the public. We may also share personal information in the event of an organizational restructuring". While the pre-GDPR snapshot contained segments related to sharing data with third-party entities, this newly added segment does not specify the type of personal information released.

Case 5: Improved Specificity

Finally, we observe that a large portion of the privacy policies have improved their specificity by using more precise phrases to describe the data collected and shared along with the purposes. For both sets, five of the eight queries had a higher number of policies with improving specificity than the number of policies with reduced specificity. This event occurred for queries Q1, 2, 7, and 8 for each set. Additionally, the *EU* set had more positive changes than negative changes for query Q6 (around first party purposes), and the *Global* set had more for query Q5 (around third party entities). Here is a representative example of improved specificity for Q1 (around data collection method) from the privacy policy of legacy.com [17]

pre-GDPR: "Examples of Personally Identifiable Information we may collect include name, postal address, email address, credit card number and related information, and phone number...."

post-GDPR: "...the personal data you provide *when you register to use our Services*, including your name, email address, and password; the

personal data that may be contained in any video, photo, image..." We note here that pre-GDPR version did not mention how the personal data was being collected, while the post-GDPR version specifically mentions it as during registration.

9.2 Takeaways

In conclusion, privacy policies appear to be more specific post the GDPR. While many of them have maintained the same specificity levels (due to unchanged policies or low coverage of data practices), a considerable number of policies has been changed. Of those policies, for both sets, a minority encountered reduced specificity according to our metrics; to comply with the GDPR, they have tried to be more comprehensive in describing their practices at the expense of being less specific. The majority of the policies that changed, however, have been more specific in informing the users about their privacy practices.

When comparing the two sets, the *EU* set exhibits a higher portion of improving policies than the *Global* set for all the specificity queries. The *EU* set also had a higher portion of policies with lower scores for each of the eight specificity queries. Recall that the *EU* dataset has higher coverage of the related privacy practices (Sec. 7). This result suggests that simply modifying a policy to increase coverage does not guarantee transparency to the data subjects. This can also be seen by analyzing policies whose coverage improved for a particular category, *e.g.* data retention. We find that more than 58% of policies with improved data retention coverage are not fully specific (*i.e.* they have a specificity score of less than 1) suggesting that, despite better coverage, the transparent disclosure of the privacy information is yet to be fully achieved.

10 Limitations

Despite our efforts to cover a diverse set of websites and to understand privacy policies' evolution from multiple angles, we acknowledge that this work has several limitations.

First, our approach in assessing the content of privacy policies does not fully capture all the efforts introduced due to the GDPR. For instance, the concept of layered privacy notices has been adopted by several companies to give users two levels of understanding: an overview of high-level practices and an in-depth description of these practices. Unfortunately, this is difficult to automatically analyze as it can come in a variety of formats, such as multi-page policies.

Second, our study is limited to the English language policies as we wanted to leverage the existing techniques of advanced semantic understanding of natural language. We made this trade-off for depth vs. language-coverage and decided not to use a keyword-based analysis.

Third, the use of automated approaches for selecting privacy policies and analyzing them is inherently error-prone. Hence, our specificity and compliance analysis might have been affected by the errors made with the machine learning models. This is accentuated by the fact that privacy policies are complex and even human annotators disagree on the interpretations sometimes, as we found in our manual verification. Nevertheless, with the recent success and increasing accuracy levels across different tasks [12, 36], we believe that such techniques, coupled with manual post-analysis, are a highly effective venue for in-depth analysis at scale.

Fourth, we restricted our user study to focus on the appearance of privacy policies, rather than their content. A more comprehensive user study, as a dedicated research project, is needed to gain a deeper understanding of how much the readability of the policies has evolved.

11 Related Work

In the following, we survey the evolution of the privacy policies' landscape, particularly in relation to regulatory intervention. We also describe the recent research that studies the GDPR's impact on privacy policies.

Evolution of the Privacy Policies Landscape

In 2002, the Progress and Freedom Foundation (PFF) studied a random sample of the most visited websites and found that, compared to two years earlier, websites were collecting less personally identifiable information and offering more opt-in choices and less opt-out choices [1]. Another longitudinal analysis has been performed by Milne and Culnan in the 1998-2001 period, confirming the positive change in the number of websites including notices about information collection, third-party disclosures, and user choices [26]. In the same period, Liu and Arnett found that slightly more than 30% of Global 500 Web sites provide privacy policies on their home page [21].

Despite the increased proliferation of privacy policies, their lack of clarity was one of the primary motivations of the regulatory measures before the GDPR. In 2004, Antón et al. showed that 40 online privacy statements from 9 financial institutions have questionable compliance with the requirements of the Gramm-Leach-Bliley Act (GLBA). They assessed the requirements of having "clear and conspicuous" policies via keyword-based investigation and readability metrics [3]. In 2007, Antón et al. studied the effect of the Health Information and Portability Accountability Act (HIPAA) via a longitudinal study of 24

healthcare privacy policy documents from 9 healthcare Web sites. A similar conclusion held: although HIPAA has resulted in more descriptive policies, the overall trend was reduced readability and less clarity. Resorting to a user study, in 2008, Vail et al. showed that users perceive traditional privacy policies (in paragraph-form) to be more secure than shorter, simplified alternatives. However, they also demonstrated that these policies are significantly more difficult to comprehend than other formats [34].

In a recent study, Turow et al. studied the surveys around the "privacy policy" as a label between 2003 and 2015 in the US. They found that the users' misplaced confidence in this label, not only carries implication on their personal lives but affects their actions as citizens in response to government regulations or corporate activities [33].

Privacy Policies After the GDPR

Since the GDPR had been enforced on May 25, 2018, a few studies have investigated its impact on privacy practices of companies. Despite the initial trials with automated approaches in these studies, they have been limited in terms of scale, which is the main goal behind automation. Contissa et al. conducted a preliminary survey of 14 privacy policies of the top companies as an attempt to measure the compliance of these companies with the GDPR automatically. They found a frequent presence of unclear language, problematic processing, and insufficient information. They used Claudette, a recent system designed for the detection of such types of issues in privacy policies [19]. Tesfay et al. introduced a tool, inspired by the GDPR to classify privacy policy content into eleven privacy aspects [32]. They validated their approach with ten privacy policies.

The first large-scale study concurrent to ours is that by Degeling et al. who performed a longitudinal analysis of the privacy policies and cookie consent notices of 6,759 websites representing the 500 most popular websites in each of the 28 member states of the EU [9]. They found that the number of websites with privacy policies has increased by 4.9% and that 50% of websites updated their privacy policies just before the GDPR came into action in May 2018. Unlike our work, however, their work has been focused on cookie consent notices and terminology-based analysis of privacy policies, without an in-depth tackling of the semantic change of privacy policies.

12 Takeaways and Conclusions

In this paper, we seek to answer a question about the impact of the recently introduced General Data Protection Regulation on website privacy policies. To answer this question, we analyze a sample of 6,278 unique English-language privacy policies from inside and outside the EU. Using the pre-GDPR and post-GDPR of each policy we study the changes along five dimensions: presentation, textual features, coverage, compliance, and specificity.

The results of our tests and analyses suggest that the GDPR has been a catalyst for a major overhaul of the privacy policies inside and outside the EU. This overhaul of the policies, manifesting in extensive textual changes, especially for the EU-based websites, does not necessarily come at a benefit to the users. Policies have become considerably longer (a fifth longer in the *Global* set and a third longer in the *EU* set). Our presentation analysis, however, identified a positive trend in user experience, specifically for the *EU* policies; such a trend was not found for *Global* policies.

We discovered another positive development in the high-level privacy category coverage of policies. For both *EU* and *Global* sets, the portion of policies mentioning these topics increased, especially in categories that were previously underrepresented. Interestingly, once more, we observed a more significant upward trend from the *EU* policies, and this disparity between the two sets continued in our compliance analysis. The average portion of compliant *EU* policies was larger than that of the *Global* policies (according to our queries). While the majority of policies in both sets scored full marks for each compliance query, the analysis revealed inconsistency between coverage and compliance. Simply covering a high-level privacy topic does not ensure the level of detail required by the GDPR is met. This discrepancy emerges in our specificity analysis. Again the margin of improvement was larger for *EU* policies compared to *Global*. However, we observed that in several cases the improvement in coverage resulted in reduced specificity.

In summary, we observe two major trends from our analysis. First, while the GDPR has made a positive impact on the overall incorporation of privacy rights and information, its influence is more pronounced inside the EU, its primary target region. Second, although policies are becoming consistently longer and covering more privacy topics, more change is to come before we reach a stable status of full-disclosure and transparency.

13 Acknowledgements

This study was supported in part by the Wisconsin Alumni Research Foundation and the NSF award 1838733. All the data gathered and analyzed in this study, as well as the methods of analysis are publicly available at https://github.com/wi-pi/GDPR/

References

[1] W. F. Adkinson, J. A. Eisenach, and T. M. Lenard, "Privacy online: A report on the information practices and policies of commercial web sites," *Progress and Freedom Foundation*, 2002.

[2] E. AI. [Online]. Available: https://spacy.io/

[3] A. I. Anton, J. B. Earp, Q. He, W. Stufflebeam, D. Bolchini, and C. Jensen, "Financial privacy policies and the need for standardization," *IEEE Security & privacy*, vol. 2, no. 2, pp. 36–45, 2004.

[4] A. I. Antón, J. B. Earp, and A. Reese, "Analyzing website privacy requirements using a privacy goal taxonomy," in *Requirements Engineering, 2002. Proceedings. IEEE Joint International Conference on*. IEEE, 2002, pp. 23–31.

[5] A. I. Anton, J. B. Earp, M. W. Vail, N. Jain, C. M. Gheen, and J. M. Frink, "Hipaa's effect on web site privacy policies," *IEEE Security & Privacy*, vol. 5, no. 1, pp. 45–52, 2007.

[6] T. H. R. Campaign. [Online]. Available: https://www.hrc.org/hrc-story/privacy-policy

[7] A. Cohen. [Online]. Available: https://github.com/seatgeek/fuzzywuzzy

[8] G. Contissa, K. Docter, F. Lagioia, M. Lippi, H.-W. Micklitz, P. Pałka, G. Sartor, and P. Torroni, "Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence," 2018.

[9] 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 *26th Annual Network and Distributed System Security Symposium, NDSS 2019, San Diego, California, USA, February 24-27, 2019*. The Internet Society, 2019. [Online]. Available: https://www.ndss-symposium.org/ndss2019/

[10] E.-P. Directive, "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)," *Official Journal L*, vol. 201, no. 31, p. 07, 2002.

[11] E. 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," *Official Journal of the EC*, vol. 23, no. 6, 1995.

[12] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K. Shin, and K. Aberer, "Polisis: Automated analysis and presentation of privacy policies using deep learning," in *27th USENIX Security Symposium (USENIX Security 18)*. USENIX Association, 2018.

[13] B. Kahle. [Online]. Available: https://archive.org/help/wayback_api.php

[14] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 2014, pp. 1746–1751. [Online]. Available: http://aclweb.org/anthology/D/D14/D14-1181.pdf

[15] C. Kohlschütter, P. Fankhauser, and W. Nejdl, "Boilerplate detection using shallow text features," in *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 2010, pp. 441–450.

[16] L. Lebanoff and F. Liu, "Automatic detection of vague words and sentences in privacy policies," *arXiv preprint arXiv:1808.06219*, 2018.

[17] Legacy.com. [Online]. Available: https://www.legacy.com/about/privacy-policy

[18] G. Lindgaard, G. Fernandes, C. Dudek, and J. Brown, "Attention web designers: You have 50 milliseconds to make a good first impression!" *Behaviour & information technology*, vol. 25, no. 2, pp. 115–126, 2006.

[19] M. Lippi, P. Palka, G. Contissa, F. Lagioia, H.-W. Micklitz, G. Sartor, and P. Torroni, "Claudette: an automated detector of potentially unfair clauses in online terms of service," *arXiv preprint arXiv:1805.01217*, 2018.

[20] K. Litman-Navarro, "We read 150 privacy policies. they were an incomprehensible disaster." https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html, 2019, accessed: 2019-06-13.

[21] C. Liu and K. P. Arnett, "Raising a red flag on global www privacy policies," *Journal of Computer Information Systems*, vol. 43, no. 1, pp. 117–127, 2002.

[22] E. T. Loiacono, R. T. Watson, D. L. Goodhue *et al.*, "Webqual: A measure of website quality," *Marketing theory and applications*, vol. 13, no. 3, pp. 432–438, 2002.

[23] N. Lomas, "Privacy policies are still too horrible to read in full." https://techcrunch.com/2019/06/13/privacy-policies-are-still-too-horrible-to-read-in-full/, 2019, accessed: 2019-06-13.

[24] M. Lui and T. Baldwin, "langid. py: An off-the-shelf language identification tool," in *Proceedings of the ACL 2012 system demonstrations*. Association for Computational Linguistics, 2012, pp. 25–30.

[25] F. Marotta-Wurgler, "Self-regulation and competition in privacy policies," *The Journal of Legal Studies*, vol. 45, no. S2, pp. S13–S39, 2016.

[26] G. R. Milne and M. J. Culnan, "Using the content of online privacy notices to inform public policy: A longitudinal analysis of the 1998-2001 us web surveys," *The Information Society*, vol. 18, no. 5, pp. 345–359, 2002.

[27] M. A. Napierala, "What is the bonferroni correction," *AAOS Now*, vol. 6, no. 4, p. 40, 2012.

[28] R. Ramanath, F. Liu, N. M. Sadeh, and N. A. Smith, "Unsupervised alignment of privacy policies using hidden markov models," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, 2014, pp. 605–610. [Online]. Available: http://aclweb.org/anthology/P/P14/P14-2099.pdf

[29] K. Reinecke, T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, and K. Z. Gajos, "Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2013, pp. 2049–2058.

[30] L. Richardson. [Online]. Available: https://www.crummy.com/software/BeautifulSoup/

[31] J. Singer-Vine, "WayBackPack: Open source scientific tools for Python." [Online]. Available: https://pypi.org/project/waybackpack/

[32] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna, "Privacyguide: Towards an implementation of the eu gdpr on internet privacy policy evaluation," in *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics*. ACM, 2018, pp. 15–21.

[33] J. Turow, M. Hennessy, and N. Draper, "Persistent misperceptions: Americans' misplaced confidence in privacy policies, 2003–2015," *Journal of Broadcasting & Electronic Media*, vol. 62, no. 3, pp. 461–478, 2018.

[34] M. W. Vail, J. B. Earp, and A. I. Antón, "An empirical study of consumer perceptions and comprehension of web site privacy policies," *IEEE Transactions on Engineering Management*, vol. 55, no. 3, pp. 442–454, 2008.

[35] A. Van Lamsweerde, "Goal-oriented requirements engineering: A guided tour," in *Requirements Engineering, 2001. Proceedings. Fifth IEEE International Symposium on*. IEEE, 2001, pp. 249–262.

[36] 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, T. B. Norton, E. H. Hovy, J. R. Reidenberg, and N. M. Sadeh,

"The creation and analysis of a website privacy policy corpus," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016. [Online]. Available: http://aclweb.org/anthology/P/P16/P16-1126.pdf

A Policy Classifier Architecture

Fig. 12 shows the detailed architecture of the single-label classifier used in Sec. 3 for checking whether the crawled pages are valid privacy policies. The input text, obtained from the web page, is tokenized into words, using the Penn Treebank Tokenizer in *nltk*[4]. Then the words are mapped into vectors at the word embeddings layer. The word vectors are input to a convolutional layer, followed by a Rectified Linear Unit (ReLU) and a Max-pooling layer. The next layer is a fully connected (dense) layer followed by another ReLU. Finally, we apply a softmax on the output dense layer to obtain a probability distribution over the two labels "valid" and "invalid". For more details about this kind of classifiers we refer the reader to the work of Kim [14] on sentence classification.

The data used to train the classifier was composed of (1) a set of 1,000 privacy policies labeled as *valid* from the ACL/COLING 2014 privacy policies' dataset released by Ramanath et al. [28] and (2) an *invalid* set consisting of the text from 1,000 web pages, fetched from random links within the homepages of the top 500 Alexa websites. We ensured that the latter pages do not have any of the keywords associated with privacy policies in their URL or title. The data was split into an 80% training set and a 20% testing set, and the classifier yielded a 99.09% accuracy on the testing set.

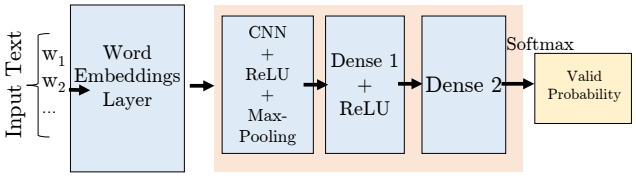
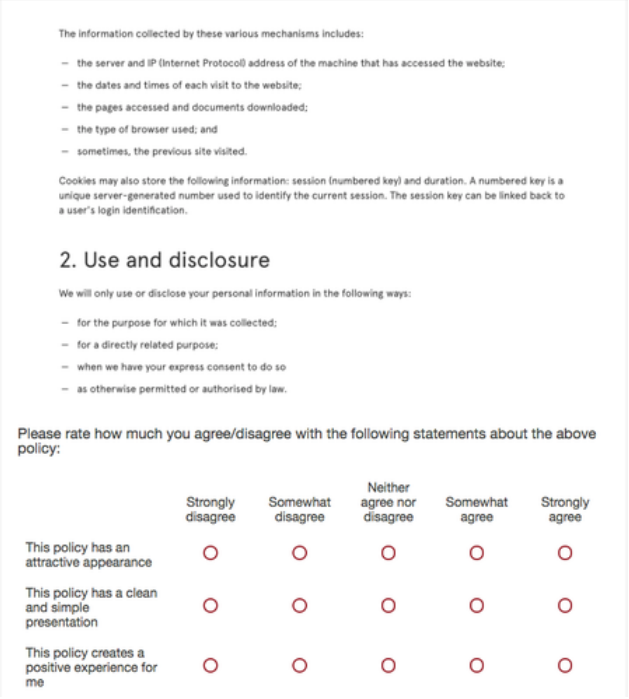


Fig. 12. Architecture of the *"Is Policy?"* classifier used to determine whether a privacy policy is valid. Hyperparameters used: Embeddings size: 100, Number of filters: 250, Filter Size: 3, Dense Layer Size: 250, Batch Size: 32

4 http://www.nltk.org/

B User Survey

In Fig. 13, we show a screenshot of an example step from the user survey that we presented to the users in Sec. 4.


The information collected by these various mechanisms includes:

- the server and IP (Internet Protocol) address of the machine that has accessed the website;
- the dates and times of each visit to the website;
- the pages accessed and documents downloaded;
- the type of browser used; and
- sometimes, the previous site visited.

Cookies may also store the following information: session (numbered key) and duration. A numbered key is a unique server-generated number used to identify the current session. The session key can be linked back to a user's login identification.

2. Use and disclosure

We will only use or disclose your personal information in the following ways:

- for the purpose for which it was collected;
- for a directly related purpose;
- when we have your express consent to do so
- as otherwise permitted or authorised by law.

Please rate how much you agree/disagree with the following statements about the above policy:

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
This policy has an attractive appearance	○	○	○	○	○
This policy has a clean and simple presentation	○	○	○	○	○
This policy creates a positive experience for me	○	○	○	○	○

Fig. 13. Example step from our user survey where the users had to respond to the three questions