Legal Knowledge and Information Systems
M. Palmirani (Ed.)

doi:10.3233/978-1-61499-935-5-51

Automated Processing of Privacy Policies Under the EU General Data Protection Regulation

Giuseppe CONTISSA [a], Koen DOCTER [b], Francesca LAGIOIA [b], Marco LIPPI [c], Hans-Wolfgang MICKLITZ [b], Przemyslaw PALKA [d], Giovanni SARTOR [a] and Paolo TORRONI [e]

[a] *CIRSFID, Alma Mater – Università di Bologna, Italy*
[b] *Law Department, European University Institute, Florence, Italy*
[c] *DISMI – Università di Modena e Reggio Emilia, Italy*
[d] *Center for Private Law, Yale Law School, New Haven, United States*
[e] *DISI, Alma Mater – Università di Bologna, Italy*

Abstract. Two years after its entry into force, the EU General Data Protection Regulation became applicable on the 25th May 2018. Despite the long time for preparation, privacy policies of online platforms and services still often fail to comply with information duties and the standard of lawfulness of data processing. In this paper we present a new methodology for processing privacy policies under GDPR's provisions, and a novel annotated corpus, to be used by machine learning systems to automatically check the compliance and adequacy of privacy policies. Preliminary results confirm the potential of the methodology.

1. Introduction: the legal and technological context

In Europe the processing of online personal data falls under the the General Data Protection Regulation (GDPR), which aims at making all data processing (from collection, to usage to transfers) lawful, fair and transparent. The enforcement of GDPR is based on two complementary approaches: (1) the administrative control by independent supervisory authorities and (2) the exercise of private rights by data subjects and/or civil society. The supervisory authority can either act on its own motion, or as a result of a complaint by a data subject or an NGO. To ensure transparency and enable the effective exercise of data subjects' rights, the GDPR requires controllers to provide the data subject with the information enlisted in Art. 13 and 14. Art. 12 stipulates that all this information must be given "in a concise, transparent, intelligible and easily accessible form, using clear and plain language". The document containing this information, namely the privacy policy, fails to be GDPR compliant if it foresees unlawful processings, if it does not contain required information, or if it uses unclear language. Our research indicates that many privacy policies fail to meet the requirements of the GDPR (see 4).

This undesirable state of affairs is due to the fact that even though data subjects, civil society and public authorities are legally empowered to conduct the control, they lack factual capabilities to do so, given the large amount of privacy policies to be checked and their complexity. Recent research has shown that tools for legal text analytics can be used to assess the completeness of privacy policies [2] and to automatically extract, categorize, and summarize information from privacy documents [3, 7, 6].

This work builds on previous research [4] where we have used machine learning methods to address the automated detection of potentially unfair clauses in online contracts. Its purpose is threefold: (1) to define the standard for a correctly designed privacy policy, in form and in content, under the currently existing standards put forward by the GDPR; (2) to analyse the privacy policies of 14 relevant online platforms and services in accordance with that standard; and (3) to verify to what extent such analysis can be automated, in order to empower consumers as a response to the technological supremacy of companies and business. If we succeed at (even partially) automating the analysis of privacy policies, this can pave the way for the development of tools that increase the efficiency and quality of the work of supervisory authorities and NGOs, and/or empower data subjects themselves. Section 2 provides an overview of the document corpus and describes the methodology adopted for evaluating privacy policies. In particular, it describes all the legal requirements that a properly designed privacy policy should meet. Further, we provide an overview of the document annotation procedures. Section 3 explains the machine learning methodology employed in the system, and some preliminary results. Section 4 concludes with a look at future research.

2. Classification of clauses and annotation guidelines

In this section, we shall provide the methodology adopted for evaluating privacy policies, and an overview of the annotated corpus.

According to the GDPR, privacy policies should be comprehensive, regarding the information they provide; comprehensible, regarding the form of expression; and substantively compliant, regarding GDPRs rules and principles (see 1). Thus, we defined a **Golden Standard** including the following three top-level dimensions: (1) Comprehensiveness of information: the policy should include all the information that is required by articles 13 and 14 of the GDPR; (2) Substantive compliance: the policy should only allow for the types of processing of personal data that are compliant with the GDPR; and (3) Clarity of expression: the policy should be framed in an understandable and precise language. With regard to these three dimensions we distinguished optimal and suboptimal achievement. In the first case the privacy policy clearly meets the GDPR requirements along the dimension at issue; while in case of suboptimal achievement the privacy policy apparently fails to reach the threshold required. In some cases we have distinguished two levels of suboptimal achievement: (a) questionable achievement: it may be reasonably doubted that the suboptimal policy reaches the threshold required (the clause could have been better framed, but still there is the possibility that the competent authorities view it as being good enough, i.e., that its improvement is only

supererogatory); and (b) Insufficient (or no) achievement: the suboptimal policy clearly fails to reach the threshold. For each dimension, we have distinguished different aspects relatively to which the clause could be assessed. Each of these aspects of profile was denoted by a tag to be used, together with a number indicating the level of achievement, in the annotation of our corpus, as we shall see in the next section.

The corpus for our exploratory inquiry consisted of 14 relevant online privacy policies, i.e. Google, Facebook, Amazon, Apple, Microsoft, WhatsApp, Twitter, Uber, AirBnB, Booking, Skyscanner, Netflix, Steam and Epic Games. These privacy policies were selected among those provided by the main online platforms, taking into account their significance in terms of number of users and global relevance. Due to the limitation of our resources and time constraints we had to focus on such a limited number of documents (thanks to additional resources, we will be able to expand it substantially in the future), but the corpus still has a significant size. In fact, it contains overall 3,658 sentences (80,398 words), 401 sentences (11.0%) of which were manually marked as containing unclear language; 1,240 (33.9%) were marked as potentially problematic or as providing insufficient information. We used XML as a mark-up language. In cases where a single clause fell into multiple categories according to our classification, we applied to it multiple tags. If a clause span included multiple sentences, we tagged all such sentences. Readers can review full privacy policies annotated here: http://www.claudette.eu/gdpr/.

In the following subsection, we shall introduce, for each dimension of our golden standard, the different aspects of it that were distinguished in our annotation, being denoted by different tags.

2.1. Comprehensiveness of Information

The dimension of comprehensiveness of information concerns whether a privacy policy meets all the information requirements of Articles 13 and 14 of the GDPR, or fails to do so, either by not providing at all the required item of information, or by providing it insufficiently or imprecisely. We identified 12 types of required information clauses, for which we defined corresponding XML tags, as specified below. For each type of required information, we classified the corresponding clause either as optimal, i.e., fully informative (all the required information is present and well specified); or as suboptimal, i.e., insufficiently informative (information is hinted at, but non-comprehensive), appending to each XML tag respectively number 1 or 2. As noted above, a single clause in some cases may fall in different categories and consequently may have multiple taggings. In the following we present each category.

Identity of the controller and, where applicable, of the controller's representative (label:`<id>`). According to the GDPR this information must be provided both when personal data are collected from the data subject (Article 13(1)(a)) and when they have not been obtained from the data subject (Article 14(1)(a)). As an example of a suboptimal clause in this regard, thus labelled as `<id2>`, consider the following example taken from the Airbnb privacy policy (last updated on 16 April 2018):

```
<id2>If you change your Country of Residence, the Data Controller and/or Payments
Data Controller will be determined by your new Country of Residence as specified
above, from the date on which your Country of Residence changes.</id2>
```

Contact details of the controller and, where applicable, of the controller's representative (label:<contact>). Contact details should allow for different forms of communication with the data controller (See Article 29 Working Party Guideline on Transparency under Regulation 2016/679 (WP260), hereinafter "Transparency Guidelines", p. 26). In order to facilitate the exercise of data subjects rights, the data controller should "also provide means for requests to be made electronically, especially where personal data are processed by electronic means" (See GDPR, Recital 59). We labelled a clause on contact details as <contact2> when it did not allow for different forms of communication with the data controller (e.g. phone number, email, postal address etc) or it failed to provide adequate specifications.

Contact details of the data protection officer (label:<dpo>). They must be published and communicated by the controller or processor to the relevant supervisory authorities (Article 37(7)). This information should allow data subjects to contact the DPO easily and directly, without having to contact another part of the organisation. Article 37(7) does not require that the published contact details should include the name of the DPO. Whilst it may be a good practice to do so. We labelled a clause as <dpo2> when it only reached a low standard for the clarity and accessibility of the information, e.g. when it only provided a dedicated email address, omitting both the name of the DPO and a postal address.

Purposes of the processing (label:<purp>). Purpose specification must be provided by the controller (Articles 13(1)(c) and 14(1)(c)), to ensure a degree of user control and transparency for the data subject (See GDPR Art. 6(1)(a); Article 29 WP Guidelines on consent under Regulation 2016/679 (WP259 rev.01) p. 13; and Recital 42 GDPR). We labelled as <purp2> those cases where, for instance, it was unclear (i) what type of data would be processed; and (ii) what the correlation was between the collected information and the specific purposes, since a number of different purposes were listed one after the other.

Legal basis for the processing (label:<basis>). The legal basis of the processing (Articles. 13(1)(c) and 14(1)(c)) must be specified with regard to both personal data (Article 6(1)) and special categories of personal data (Article 9). For instance, we labelled as <basis2> those clauses that specified the purpose with reference to broad marketing practices, or that mixed multiple unrelated purposes.

Categories of personal data concerned (label:<cat>). This specification must be provided where personal data have not been obtained from the data subject (Article 14(1)(d)), as well as whenever the data subject consent constitutes the legal basis for the processing (Articles 6 and 9). A kind of clauses that we labelled as <cat2> were those that only provided a non-exhaustive list of examples of the collected data.

Recipients or categories of recipients of the personal data (label:<recep>) (Article 13(1)(e)). In accordance with the principle of fairness, controllers must provide information on the recipients that is meaningful for data subjects, i.e. the named recipients, so that data subjects know exactly who has their personal

data, or the categories of recipients, by indicating the type of recipient (i.e. the activities it carries out), the industry, sector and sub-sector and its location. We labelled clauses as `<recep2>`, when they did not clearly indicate the type, the industry, the sector and the location of the mentioned recipients.

The period for which the personal data will be stored, or if that is not possible, the criteria used to determine that period (label:`<ret>`). 'We labelled as `<ret2>` those clauses that fail to specify the time for which the personal data will be stored, or at least the criteria used to determine that period, such as those clauses generically stating that personal data will be kept as long as necessary.

The rights to access; rectification; erasure; restriction on processing; objection to processing and data portability (label:`<correct>`) (Articles 13.2(b) and 14.2(c)). This information should be specific to the processing scenario and include a summary of what the right involves and how the data subject can take steps to exercise it, as well as any limitations on the right. We classified as `<correct2>` those clauses that failed to specify under what condition data subjects could exercise their rights and what steps were needed to exercise them.

The right to lodge a complaint with a supervisory authority (label:`<complain>`) (Articles 13.2(b) and 14.2(c)). We labelled as `<complain2>` the clauses that failed to specify in which State a complain should be presented, or that provided a wrong or misleading specification.

Information about source from which the personal data originate, and if applicable, whether it came from publicly accessible sources (label:`<source>`). Information on the source must be provided if personal data are not coming directly from the data subject (Article 14(2)(f)). We labelled as `<source2>` those clauses failing to specify the nature of the sources (i.e. publicly/ privately held sources; the types of organisation/ industry/ sector; and where the information was held (EU or non-EU) etc.).

The existence of automated decision-making, including profiling (label:`<auto>`). The policy must also include meaningful information about the logic involved, the significance and the envisaged consequences of such processing for the data subject (Articles 13.2(f) and 14.2(g))[1]. We labelled as `<auto2>` those clauses that failed to provide meaningful information about the logic involved, as well as the significance and the envisaged consequences of such processing for the data subject; or that did not inform the data subject about the right not to be subject to a decision based solely on automated decision making, including profiling; and that the decisions referred would not be based on sensitive data.

2.2. Substantive compliance

This dimension concerns whether the types of processing stipulated are themselves GDPR compliant. We identified 10 categories of clauses and for each category, we defined a corresponding XML tag, as specified below. We assumed that each category could be classified either as a *fair processing*, clause; a *problematic processing* clause; and as an *unfair processing clause*. To this end, we appended a numeric value to each XML tag, with 1 meaning fair; 2 problematic; and 3 unfair.

[1]See Article 29 Working Party Guideline on Automated Individual Decision-making and Profiling for the Purposes of Regulation 2016/679 (WP251rev.01).

Processing of special categories of personal data (label:<sens>). Processing of sensitive data (e.g. data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, concerning health or a natural person's sex life or sexual orientation, etc.) are prohibited unless an exception applies (Articles 9 and 9(2)). We labelled as <sens2>, those clauses that allowed for the processing of sensitive data without providing full information e.g. clauses stating that the explicit consent is required for processing certain sensitive data, but failing to indicate the purpose of such processing. We labelled as <sens3> the clauses allowing for the processing of sensitive data outside the conditions specified in Article 9. As an example of a <sens2>-labelled clause, consider the following fragment taken from the Facebook privacy policy (last updated on 19 April 2018):

<sens2>To create personalized products that are unique and relevant to you, we use your connections, preferences, interests and activities based on the data we collect and learn from you and others (including any data with special protections you choose to provide where you have given your explicit consent); how you use and interact with our Products; and the people, places, or things you're connected to and interested in on and off our Products.</sens2>

Consent by using (label:<cuse>). Consent should be given "by a statement or by a clear affirmative action" (art 4(11)). Thus we labelled as <cuse3>, among others, those clauses stating that by simply using the service, the user consents to the terms of the privacy policy. For instance, consider the following example taken from the Epic games privacy policy (last updated on 24 May 2018):

<cuse3> when you use our websites, games, game engines, and applications, you agree to our collection, use, disclosure, and transfer of information as described in this policy, so please review it carefully.</cuse3>

Take or leave it approach (label:<tol). "When assessing whether consent is freely given, utmost account shall be taken of whether, inter alia, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract."(Article 7(4)). The situation of "bundling" consent with acceptance of terms or conditions, or "tying" the provision of a contract or a service to a request for consent to process personal data that are not necessary for the performance of that contract or service, is considered highly undesirable". We always labelled all clauses implementing a "take it or leave it approach" as <tol2>.

Transfers to third parties (label:<tp>). The data subjects should be informed on whether their information will be transferred to third parties, on the identity of the transferee, and on the legal basis for the transfer (such as consent by the party, necessity for the execution of a contract, or legitimate interest). We labelled a clause as <tp2> when the purpose of the transfer or identity of the transferee were not specified, but (a) the transfer presupposed consent of the data subject, which was not necessary to access the service, or (b) the transfer was needed to perform the contract. We labelled a clause as <tp3> when the purpose of the transfer or the identity of the transferee were not specified, but consent was necessary to access the service and the transfer was not needed to perform the contract.

Policy change (label:<pch>). The controller should adhere to the transparency principle when communicating both the initial privacy statement/ notice and

any subsequent substantive or material changes to this statement/ notice, and consider several factors in assessing what is a substantive or material change. We labelled a clause as <pch2> when stating that a notice of the change would be provided, but would not require new consent or a confirmation of reading; and as <pch3> when even the commitment to a fresh notice was rejected (e.g. clauses stating that it is a responsibility of the data subject to check for the last updated version of the privacy policy).

Transfer to third country (label:<cross>). Articles 13(1)(f) and 14(1)(f) require the controller to inform the data subject about (i) whether he/she intends to transfer personal data to third countries and (ii) the existence or absence of adequacy decision by the Commission or appropriate safeguards. Besides, Chapter V (art. 44–49) governs the transfer of personal data to third countries. We labelled a clause as <cross2> when it only mentioned one of the transfer mechanisms listed by Articles 44–49 and did not provide any specific information allowing the data subject to be effectively informed. We labelled as <cross3> clauses failing to provide any information on the transfer mechanism requirements.

Processing of children's data (label:<child>). According to Article 8(1) The GDPR requires parental consent for the processing of data concerning children below 16 years (Article 8), and recommends a cautious and proportionate approach for all children (individuals under 18). We labelled as <child2> those clauses failing to mention and/or specify the types of efforts made to verify that the consent is given/authorised by the holder of parental responsibility. We labelled as <child3>, for instance, those clauses failing to specify what efforts would be taken to verify parental authorisation.

Advertising (label:<ad>). Whenever profiling for marketing purposes involves the use of personal data that were originally collected for something else, these marketing purposes must be compatible with the original purposes for which the data were collected, and moreover, the data subject has a right to opt-out (Articles 21(2), 21(3), and Recital 70). Consequently, we labelled a clause as <ad2> when the consent was not required, but the opt-out was possible, and as <ad3> when the consent was not requested, and the opt-out was not possible.

Any other type of consent (label:<c>). Since consent cannot be given through the general acceptance of a privacy policy or terms of use (Articles 4(11) and 7(2)), we labelled as <ad2> all clauses where "hidden" consents were present.

Any other type of clause we find "outstandingly problematic" (label:<out>). We labelled as <out2> clauses stating anything else that did not fall under the scope of the above-mentioned clauses, and yet could be considered as problematic for different reasons. For example, this included clauses that were obscure in their meaning, or that claimed the responsibility of the data subject regarding the processing of the data of third parties by the data controller.

2.3. Clarity of Expression

Article 5(1)(a) requires that personal data must be processed lawfully, fairly and in a transparent manner. Further, Article 12(1) requires that information must be provided in a concise, transparent intelligible and easily accessible form, using clear and plain language. Therefore, complex sentence and language structures

should be avoided. Besides, the information should not be phrased in abstract or ambivalent terms or leave room for different interpretations (see Guideline on Transparency, pp. 9-10). The use of the language qualifiers such as "may", "might", "some", "often", and "possible" should be avoided (see Guideline on Transparency, p. 9), as well as "including" and "such as" when they are present within a list, for example of the category of data collected. We identified two types of clauses: clauses expressed in a clear language (not tagged) and clauses expressed in an unclear language, for which we defined the following XML tag: `<vag>`. As an example that fails to meet optimality concerning the *clarity* of expression, consider the following example taken from the Apple privacy policy (last updated on 22 May 2018):

```
<vag>Apple and its affiliates may share this personal information with each
other and use it consistent with this Privacy Policy. They may also combine it
with other information to provide and improve our products, services, content,
and advertising.</vag>
```

3. Machine Learning Methods and Experiments

Although this work focuses on offering a novel methodology for the labeling of privacy policies, with the purpose of automating the evaluation of such documents, we also present some preliminary results of the machine learning experiments conducted on the corpus.

Given the complexity of the problem, there are several ways in which the automatic system could be developed. For example, regarding problematic clauses, one could first detect all problematic clauses (in general) and then distinguish each category (data regarding children, advertising, etc.). Or the other way around, one could first detect all sentences regarding a certain category (e.g. advertising) and then decide whether they are problematic or not. The same holds for required clauses. In contrast, vague clauses do not have sub-categories.

As for the adopted algorithms, in our current approach we started experimenting with the technologies that have been successfully employed in the detection of potentially unfair clauses in online Terms of Service [5], including support vector machines (SVM) and deep networks. In some cases, a solution based on manually defined rules and patterns could also be used to detect some specific categories of problematic or required clauses, as often done in data mining.

We employed a standard leave-one-document-out (LOO) procedure, where training is repeated N times (N being the number of documents), with a different document of the corpus used as the test set, and the remaining $N - 1$ forming the training set. Performance is measured with standard metrics: precision, as the percentage of predicted positive sentences that are indeed positive in the corpus; recall, as the percentage of positive sentences that are indeed classified as positive; F_1, as the harmonic mean between precision and recall (the set of positive clauses depending on the task).

In a first experiment, we considered unclear language clauses only. Here, a simple grammar that detects whether some keyword (or combination of keywords) is present in each sentence is capable of recognizing 89% of vague clauses, yet

with a low precision of 25%. A machine learning classifier based on Support Vector Machines and bag-of-words, instead, detects 72% of vague clauses, yet with a low precision of 30%. A combination of grammar and machine learning achieves 81%/32% recall/precision. Yet, a detailed analysis of the false positives (sentences detected as unclear, which actually were not tagged as such) shows that most of them are indeed problematic clauses. This observation made us argue that probably a machine learning classifier could take advantage of observing the combination of both problematic and unclear clauses. Therefore, we repeated the same experiments, this time considering the positive class (to be detected) as the union of problematic and vague clauses. Following the same approach described above, a hand-crafted grammar correctly detects 92% of positive clauses, yet with 31% precision. A pure machine learning classifier achieves instead 70% recall and 50% precision. A combination of the two approaches reaches a 75% recall with 47% precision, with an overall 57% F_1. Although these numbers could seem unimpressive to a lay observer, we shall remark that as preliminary results they are not bad at all. Indeed, they are comparable to the results obtained with the analysis of Terms of Service with a corpus of 20 documents [4], where we had initially obtained a 72%/62% recall/precision, which further increased to 80%/83% when the corpus was extended to include 50 documents. We could imagine a similar trend also for privacy policies.

Once problematic clauses have been detected, automatic categorization of such sentences into unlawfulness classes is a much simpler task: SVMs are capable of identifying the correct category with precision/recall usually around 80%/75%.

Some required information clauses can also be easily detected with grammars and regular expressions. For example, the category of automatic decision making can be identified with 95% precision and 83% recall, and the required information about complaints can be identified with 94% precision and 91% recall. Similarly, the data protection officer clause can be detected with 78% precision and 85% recall. Other tags are much more heterogeneous, and thus difficult to detect with hand-crafted rules (e.g., the purposes and the legal basis of data processing): in these cases, machine learning achieves performance comparable to the detection of problematic clauses.

4. Conclusions

This paper presented a first experimental study that used machine learning to evaluate privacy policy under the GDPR. Our inquiry was based on the identification of a golden standard for privacy policies, and on the definition of a methodology for assessing the extent to which policies get closer to such standard. From the legal perspective, our analysis of 14 privacy policies of online platforms and services suggests that there is still a significant margin for improvement. None of the analysed privacy policies gets close to meeting the standards put forward by the GDPR. Unsatisfactory treatment of the information requirements (e.g. with regard to contact details of the DPO; we could not retrieve an example of a fully informative clause from the policies we analysed); large amounts of sentences employing vague language; and an alarming number of "problematic" clauses cannot

be deemed satisfactory. The results we obtained in our machine learning experiments, though still limited and provisional – mostly due to the limited extension of our corpus – show that there are promising prospects for developing automated tools to support the detection of unlawful clauses in privacy policies. Providing such tools will be an important contribution to the protection of individuals, in particular consumers or internet users. In the era of big data and automated decision-making, there is indeed a strong connection between data protection and consumer protection, as we are tracked, profiled, and directed/manipulated especially in our role as potential consumers. Thus there is a strong synergy between the purpose of our first Claudette project (empowering consumers and their associations to assess the legality and the fairness of online contracts) and the purpose of the project here presented (empowering data-subjects to assess privacy policies ([1])). Addressing privacy policies has been more difficult than examining consumer contracts: policies are less modular than contracts, they use a more variable and open-textured language, may have multiple layers, etc. Moreover, privacy policies may be defective not only for including wrong clauses, but also for omitting required information: thus to assess policies, negative tests are also needed. In the future, we plan to develop our research in different directions: addressing multilingualism in privacy policies, providing argument-based explanations for the assessment of clauses, detecting inconsistencies in policies, assessing the overall quality of privacy documents. Besides, as data protection regulation is rapidly evolving we will need to consider ways to minimize the effort involved in adapting our tool to new regulations and decisions by competent authorities.

References

[1] G. Contissa, F. Lagioia, M. Lippi, P. Palka, H.-W. Micklitz, G. Sartor, and P. Torroni. Towards consumer-empowering artificial intelligence. In *Proceedings of IJCAI Conference*, pages 5150–7. 2018.

[2] E. Costante, Y. Sun, M. Petković, and J. den Hartog. A machine learning solution to assess privacy policy completeness:(short paper). In *ACM workshop on Privacy in the electronic society*, pages 91–96. ACM, 2012.

[3] K. P. Joshi, A. Gupta, S. Mittal, C. Pearce, A. Joshi, and T. Finin. Alda: Cognitive assistant for legal document analytics. *AAAI Fall symposium*, 2016.

[4] M. Lippi, P. Palka, G. Contissa, F. Lagioia, H.-W. Micklitz, Y. Panagis, G. Sartor, and P. Torroni. Automated detection of unfair clauses in online consumer contracts. *Legal Knowledge and Information Systems*, page 145, 2017.

[5] M. Lippi, P. Palka, G. Contissa, F. Lagioia, H. Micklitz, G. Sartor, and P. Torroni. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *CoRR*, abs/1805.01217, 2018. URL `http://arxiv.org/abs/1805.01217`.

[6] N. Tomuro, S. Lytinen, and K. Hornsburg. Automatic summarization of privacy policies using ensemble learning. In *Proceedings of the Sixth ACM Conference on Data and Application Security and Privacy*, pages 133–135, 2016.

[7] R. N. Zaeem, R. L. German, and K. S. Barber. Privacycheck: Automatic summarization of privacy policies using data mining. *ACM Transactions on Internet Technology (TOIT)*, 18(4):53, 2018.