Computing (2022) 104:2053–2076
https://doi.org/10.1007/s00607-022-01076-3

REGULAR PAPER

A systematic mapping study on automated analysis of privacy policies

Jose M. Del Alamo[1] · Danny S. Guaman[1,2] · Boni García[3] · Ana Diez[1]

Received: 12 November 2021 / Accepted: 10 March 2022 / Published online: 10 May 2022

Abstract

A privacy policy describes the operations an organization carries out on its users' personal data and how it applies data protection principles. The automated analysis of privacy policies is a multidisciplinary research topic producing a growing but scattered body of knowledge. We address this gap by conducting a systematic mapping study which provides an overview of the field, identifies research opportunities, and suggests future research lines. Our study analyzed 39 papers from the 1097 publications found on the topic, to find what information can be automatically extracted from policies presented as textual documents, what this information is applied to, and what analysis techniques are being used. We observe that the techniques found can identify individual pieces of information from the policies with good results. However, further advances are needed to put them in context and provide valuable insight to end-users, organizations dealing with data protection laws and data protection authorities.

Keywords Privacy policy · Natural language processing · Data protection · Privacy

Jose M. Del Alamo, Danny S. Guaman, Boni García and Ana Diez have contributed equally to this work.

✉ Jose M. Del Alamo
jm.delalamo@upm.es

Danny S. Guaman
danny.guaman@epn.edu.ec

Boni García
boni.garcia@uc3m.es

Ana Diez
ana.diezm@upm.es

1 ETSI Telecomunicación, Universidad Politécnica de Madrid, Avda. Complutense 30, 28040 Madrid, Spain

2 Escuela Politécnica Nacional, 179417 Quito, Ecuador

3 Universidad Carlos III de Madrid, 28911 Leganés, Spain

Mathematics Subject Classification 68-02 (Research exposition), 68P7 (Privacy of data)

1 Introduction

A privacy policy, also known as a privacy notice, is a statement through which an organization informs its users about the operations on their personal data (e.g. collection, transfer) and how it applies data protection principles.

The mandatory contents of a given privacy policy depend on the applicable privacy law. For example, in the European Economic Area (EEA), the General Data Protection Regulation (GDPR) Articles 12–14 set the requirements on the information to be provided to EEA citizens whenever an organization wishes to process their personal data. In China, the new Personal Information Protection Law (PIPL) sets similar requirements. In the US, the requirements vary according to the specific circumstances. For example, the Children's Online Privacy Protection Act (COPPA) sets requirements when child data are processed, the Health Insurance Portability and Accountability Act (HIPAA) sets requirements when health data are processed, or the California Consumer Privacy Act (CCPA) sets requirements when California's residents personal data are processed. Most countries have similar legislation in place mandating organizations to inform their users about their personal data practices in clear and plain language so that they can understand the privacy concerns.

Privacy policies are typically presented as textual documents [1]. Their automated analysis is becoming a pressing need for different stakeholders. Global organizations need to know whether their policies comply with the varied local privacy laws where they offer their products and services. Supervising authorities overseeing privacy laws require automated means to cope with the myriad of privacy policies disclosing the practices of online systems processing personal data (e.g. websites, smart devices). Users demand new ways of understanding the verbose and complex legal texts they are confronted with e.g., when browsing the web or installing a new fancy app.

The automated analysis of written privacy policies is a multidisciplinary problem involving legal (i.e. privacy and data protection legislation) and technical (e.g. natural language processing) domains. Research efforts are scattered across several research communities, resulting in a growing body of knowledge presented at different symposia, conference tracks, and publications. To the best of our knowledge, the state of the art still lacks an overview of techniques that can support the different stakeholders in automatically analyzing privacy policies presented as textual documents.

To fill this gap, this paper presents the first overview of the different techniques used to analyze privacy policy texts automatically, obtained through a systematic mapping study. It also identifies the concrete information obtained from the policies, and the goals pursued with this analysis. Finally, it discusses the most promising future research lines found.

2 Background

This section contains a summary of the main aspects covered in the research.

2.1 The content and readability of privacy policies

Although different privacy and data protection laws set out different requirements on policies, they also mandate some common contents to be found in any privacy policy. One of the salient contributions to the automated analysis of privacy policies is that proposed by Wilson et al. [2], which identified a set of privacy practices usually disclosed in privacy policies. We have leveraged the Wilson et al. scheme to understand what contents can be expected in a privacy policy (Table 1).

The main purpose of privacy policies is to inform users so that they can understand the privacy risks faced. However, while privacy policies may disclose detailed information on the privacy practices carried out, studies have demonstrated that users generally do not understand them [5]. Thus, policy readability is also of the utmost importance. For example, the European GDPR Recital 58 [6] requires that "any information addressed to the public or to the data subject be concise, easily accessible and easy to understand, and that clear and plain language [...] be used".

Text readability depends on its contents e.g., vocabulary and syntax, as well as its presentation e.g., font type or font size [5]. Different metrics have been devised to predict text readability, such as word/sentence length, percentage of difficult words or legibility, among others. They are introduced in readability formulas that provide a score or level that predict the overall readability of a given text. Some examples of well-known readability scores are the Flesch-Kincaid Grade Level score, the Gunning-Fog score, the Coleman-Liau Index, the SMOG Index, or the Automated Readability Index. Other aspects can influence readability, such as inconsistent or vague texts.

The content and readability metrics extracted from privacy policies can be applied to different goals such as assessing compliance with some laws (Law compliance), checking that the statements of the privacy policy are coherent with the behavior of the system under analysis (System check), informing users of the system of some practices declared in the privacy text (User information), or gathering new knowledge to inform further research (Researcher insight).

2.2 Natural language processing

Privacy policies are typically written texts and, as such, can be automatically analyzed using natural language processing (NLP) techniques. There are two main approaches in implementing NLP systems [7]:

1. Symbolic NLP (also known as "classic") is based on human-developed grammar rules and lexicons to process text and model natural language.
2. Statistical NLP (also known as "empirical") applies mathematical techniques using actual datasets (text corpora) to develop generalized models of linguistic phenomena.

Table 1 Information stated in a privacy policy

Policy content	Definition
Controller	Identification of those organizations collecting the personal data and their contact details
Data	Personal data types that will be subject to some processing operation. Organizations collecting personal data are expected to disclose what data they are collecting e.g., device identifier, location data, etc
Operations	Processing operations carried out on the personal data such as collection, organization, storage, disclosure, transmission, etc. The operations can be carried out by the controller itself (i.e., first party) or by other organizations (i.e., third parties)
Purpose	The business goal behind the processing operations carried out on personal data. For example, an organization can collect and analyze personal data to display personalized ads to the data subject
Consent	Data subjects' options to opt-in/out of the data operations described in the privacy policy
Access	Information on the rights of data subjects to access, edit and delete their personal data once they have been collected and how to enforce them
Retention	Amount of time the parties obtaining the personal data will keep them
Security	Measures taken to keep the data protected
Change	Details on how changes to the privacy policy will be communicated to the data subjects
Children	Informational aspects related to the processing of children's personal data. Children are considered as vulnerable individuals and as such, processing their personal data usually requires further information e.g., how parents can exercise control and limit the information collected
Cookies	A cookie is a small text file stored on the user's device by a website owner (first-party cookie) or other external services (third-party cookie) when users visit a website. Cookies have become a serious privacy threat [3], and under different legislations, websites are required to inform their users about who stores the data, the types of data they store, together with the purpose
Do Not Track	Do Not Track (DNT) is a World Wide Web Consortium (W3C) Recommendation [4] for an HTTP Header to be sent by users' devices to signal websites they do not want to be tracked e.g. by placing cookies on their devices. Websites were expected to inform their users whether they respond to the DNT request
Other	Privacy-related information not covered by the previous categories. For example, the GDPR mandates organizations sending personal data out of the EEA (international data transfers) to inform their data subjects of the privacy policy

A text preprocessing stage is usually required in any NLP pipeline to transform text from human language to some more convenient format for further processing. Text preprocessing is done before applying symbolic and statistical approaches. The typical steps in text preprocessing are:

1. Tokenization, which is the process of chopping input text into small pieces (called tokens).
2. Stop words removal, which consists of eliminating terms that do not add relevant meaning (e.g. "the", "a" or "an" in English).
3. Normalization, which is the process of generating the root form of the words. There are several types of normalization, such as stemming (i.e., transforming related

words without knowledge of the context) and lemmatization (i.e., normalization considering the morphological analysis of the sentences).

Traditionally, symbolic NLP is broken down into several levels, namely, morphological, lexical, syntactic, semantic, discourse and pragmatic. The morphological analysis deals with morphemes, which are the smallest unit of meaning within words. The lexical analysis studies individual words as regards their meaning and part-of-speech. The syntactic analysis studies words grouped as sentences. The semantic level is used to capture the meaning of a sentence. Ontologies are closely connected to the semantic analysis to model a domain knowledge and reason on a natural language. The discourse level is concerned with how sentences are related to others. Finally, the pragmatic level deals with the context (external to the input text).

The Statistical NLP approaches typically use Machine Learning (ML) algorithms to develop generalized models of some linguistic phenomena. These algorithms are usually classified into two groups depending on the type of learning on which they are based: supervised learning and unsupervised learning. The term supervised learning is applied to the algorithms that need a labeled dataset as input so they can learn a specific characteristic of the text that they will have to predict. Some examples of supervised algorithms are Random Forest, Naive Bayes, Support Vector Machines (SVMs), Decision Trees, or K-nearest neighbors (kNN). Instead, unsupervised algorithms do not need the input data to be labeled since their objective is to find hidden patterns in the data to understand and organize it. An example of unsupervised algorithms is K-Means used for clustering tasks. In recent times, Artificial Neural Networks (ANNs) have been applied as another ML approach to generate prediction models for NLP [8]. Knowledge is spread across the ANN, and the connectivity between units (called neurons or perceptrons) reflects their structural relationship.

The application of statistical approaches requires the converting of some natural language (i.e. text) into a mathematical data structure (i.e. numbers), used as input of the ML algorithm. This process is commonly known as text data vectorization. Second, a prediction model is created using some training data. After a model is built (or "trained"), it should be evaluated, i.e., to measure its ability to be generalized (in other words, to make accurate predictions on new, unseen data with the same characteristics as the training set). Several metrics are used to measure the performance of the model, such as precision, recall, F1-score, or accuracy.

3 Related work

To the best of our knowledge, there are currently no systematic mapping studies, surveys, or reviews that fall into the intersection of the two domains specified in the scope of this study, those being privacy policies and text analysis techniques. Basically, the secondary studies found, which can be considered as related work, can be classified into two groups: those related to text analysis techniques applied to a specific area of knowledge and those related to the analysis of privacy and related aspects. Nevertheless, we have found two reviews touching on privacy policies and

text analysis techniques. The following paragraphs describe these related works in more detail.

On the borders of our related work, we can find many reviews covering text analysis techniques, mainly NLP techniques, applied to specific areas of knowledge. We have gathered a few of the most relevant. Opinion mining systems is a very active research area in recent years in the field of NLP techniques. In their review, Sun et al. present [9] an overview of all the approaches in this field and the challenges and open problems related to opinion mining. NLP is also widely used in the healthcare sector, and one very interesting application is the generation of structured information from unstructured clinical free text. A systematic review of the advances in this sector has been carried out by Kreimeyer et al. [10]. A completely different field is covered by Nazir et al. [11]. Text analysis techniques are likewise applied in software requirement engineering in order to achieve goals such as requirement prioritization and classification, and this systematic literature review gathers the main contributions. Finally, Kang et al. carried out a literature review [12] into NLP techniques applied to management research.

As regards the second domain of our research, many reviews are published concerning privacy aspects, and some of them make references to privacy policies. That is the case of the systematic mapping study published by Guaman et al. [13]. There is only one paper in common between their research and ours (ID848) since one of their paper assessment criteria specifically excludes papers exclusively assessing privacy policies (as their focus was on the privacy assessment of information systems). Nevertheless, they report a great number of articles that use the privacy policy text to check compliance albeit manually, which was an exclusion criterion in our case, as we are looking for automated means. As in our research, they highlight that the most studied privacy law is the GDPR.

Another aspect closely related to privacy is transparency, and Murmann et al. conducted a survey into the available tools for achieving transparency [14]. They also take the GDPR as a point of reference for the definition of transparency and divide it in two types: Ex ante transparency, the one that informs about the intended actions in privacy policies, and ex post transparency, the one that provides insights into what practices have being carried out. Since their work is more focused on ex-post transparency, there are no common articles between their research and ours.

A different approach is followed by Becher et al. [15], in which they present a broad literature review about Privacy Enhancing Technologies (PETs). They include tools that allow users to perform personal privacy policy negotiation, involving the representation of the privacy policy and its personalization. Just one of our papers gathers these two characteristics (ID28), and so it is included in their study.

A study closer to ours is the review presented by Kirrane et al. [16]. They analyzed the Semantic Web research domain applied to privacy, security and/or policies. Around 40% of their analyzed papers were related to privacy policies and they found that the semantic web was being used with two purposes with regard to privacy policies: policy communication in order to help producers write policies and policy interpretation to help users understand privacy policies. This latter purpose is the one most closely related to our work, and one of the papers of our research (ID30) is included in this group.

Finally, Morel and Pardo [1] studied the different means of expression of privacy policies, namely textual, graphical and machine-readable. They analyzed the information each policy type usually discloses, the tools supporting authoring and analysis, and the benefits and limitations. However, they only report seven analysis techniques for textual policies while we found 39, including three papers they also found (ID28, ID62 and ID72).

Considering all this, our review differs from all the available surveys and reviews since no one before has focused their attention on the existing techniques to analyze privacy policies automatically. We believe that our research is necessary since privacy compliance is becoming more and more important nowadays, and automatizing this task is the only way to start making a high-quality assessment of privacy compliance at scale.

4 Methodology

A mapping study is a systematic approach to provide an overview of a research area of interest by showing quantitative evidence to identify trends [17]. We have organized our research in three stages:

1. Planning. In this stage, we defined the scope of the research, the main goal, and the Research Questions (RQs). We also formulated the search strategy, the inclusion and exclusion criteria and procedure, and finally the classification scheme and procedure.
2. Conducting. The objective of this phase was to answer the RQs. With this purpose in mind, we carried out the paper search, filtered the results based on our defined criteria, and classified the remaining papers using the classification scheme
3. Reporting. We analyzed the results to answer the RQs and discussed interesting trends and gaps discovered during the research process.

4.1 Scope and research questions

The scope of this research is the intersection between two topics: (1) privacy policies and (2) text analysis techniques. Within this scope, our overall goal is (i) to identify the techniques used to analyze privacy policy texts and (ii) to identify what information is retrieved from the privacy policies. These objectives have been divided into three specific RQs.

RQ1: What information is obtained from the privacy policies?
RQ2: What is the purpose of the policy analysis?
RQ3: What techniques have been used to analyze privacy policy texts?

4.2 Paper search strategy

We used Scopus and Web of Science (WoS) databases to find high-quality peer-reviewed literature. Scopus indexes the most important digital libraries such as IEEE

Table 2 Inclusion and exclusion criteria

Publication year	2021	2020	2019	2018	2017	2016	2015	2014	<2013
Minimum citations	0	0	2	3	4	5	5	5	6

Xplore, Springer Link, Elsevier, Science Direct, or ACM. WoS complements the Scopus database by indexing other journals and conference papers [18].

We created a search string using terms related to our two topics, privacy policies and text analyses. We used the IEEE Thesaurus to find these terms. To obtain a wider search string, we simply used the stem of these terms in the search string and used the nearby operator ('W/3' in Scopus, 'NEAR/3' in WoS). The resulting strings for each database were these:

Scopus: *((privacy OR "data protection") W/3 (text* OR polic* OR statement* OR term* OR condition* OR notice*)) W/3 (analy* OR process* OR min* OR recogni* OR learn* OR classif*))*

Web of Science: *(((privacy OR "data protection") NEAR/3 (text* OR polic* OR statement* OR term* OR condition* OR notice*)) NEAR/3 (analy* OR process* OR min* OR recogni* OR learn* OR classif*))*

To validate the completeness of the set of papers obtained, a senior privacy researcher selected 10 papers to create a test set that should be taken into consideration in the research. In every iteration of the search string definition process, we manually checked how many of them were included. We carried out the final search in these databases, searching on title, abstract and keywords.

To mitigate the threat to validity by missing relevant papers, after filtering the results of the database search, we carried out a snowballing [19] with the selected papers. This technique consists of analyzing the papers cited by the selected ones (backward snowballing), and those citing the selected ones (forward snowballing).

4.3 Inclusion and exclusion procedure

We conducted an inclusion and exclusion procedure to filter out papers. This procedure consists of an automated filter followed by a manual one.

4.3.1 Automated inclusion-exclusion

All the following inclusion criteria must be met for a paper to pass to the manual filter:

1. Language: English.
2. Document type: Conference paper and journal article.
3. Number of citations: For papers published up to 2019, the minimum number of citations of a paper must be more than the 50 percentile of citations to papers in computer science, as per Thomson Reuters. For papers published in 2020 and 2021, zero citations as a minimum. The citations per year are stated in Table 2.
4. Number of pages: We are looking for papers proposing contributions with some form of validation, and this requires extensive works with detailed publications.

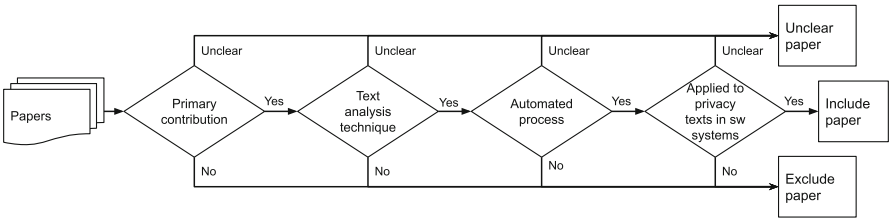

Fig. 1 Decision tree for the manual selection of papers

Table 3 Inclusion and exclusion criteria

Inclusion criteria	The paper is a primary contribution
	The paper describes a text analysis technique
	The technique described is not completely manual
	The technique described is applied to texts describing privacy aspects in software systems
Exclusion criteria	The paper reports a secondary study
	The paper does not report a text analysis technique
	The paper only reports manual techniques for analyzing texts
	The text analysis techniques are not applied to texts related to privacy policies in software systems (e.g., applied to privacy laws)
	The paper only reports the generation of a dataset of annotated texts
	The paper only reports a technique to analyze text but does not apply it to any privacy aspect
	The paper only reports a privacy policy model
	The paper only sets requirements for a privacy policy but does not analyze existing texts
	The paper only reports a technique for text processing, but it is not applied to privacy texts
	The paper only reports the use of existing metrics or scores to assess some text aspects such as readability or legibility
	The paper only reports a tool or a technique to analyze machine-oriented privacy documents

Thus, we exclude short papers, i.e., heuristically, papers with less than 5 double-column pages or 8 single-column pages.

4.3.2 Manual inclusion-exclusion

In the manual stage, two screening phases were carried out: a title and abstract screening followed by a full-text screening, both performed through CADIMA (https://www.cadima.info/). We followed the decision tree shown in Fig. 1.

The list of inclusion-exclusion criteria used to evaluate the papers is included in Table 3. All criteria must hold for inclusion, but if any exclusion criterion holds then the paper is excluded.

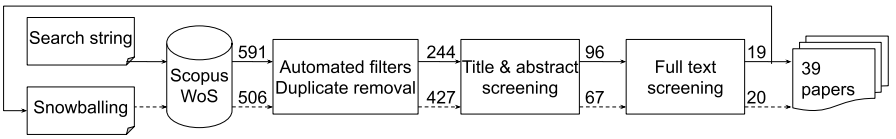

Fig. 2 Paper selection process

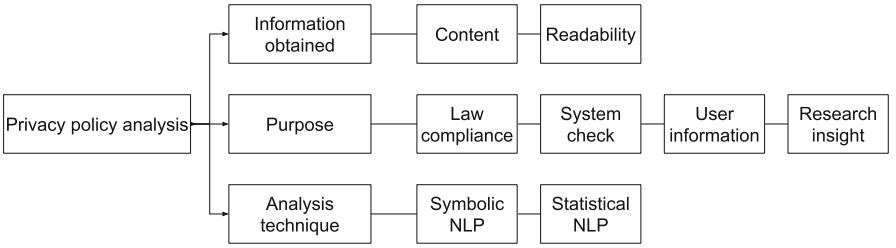

Fig. 3 Classification scheme

Each paper was reviewed by two researchers and inconsistencies were resolved in daily meetings with all the team members. At the beginning of each stage, a pilot, divided into iterations, was conducted to align the criteria of all the researchers. In each iteration, five papers were analyzed by all of the team members and the Krippendorff's alpha inter-coder reliability coefficient [20] was used to calculate the inter agreement. To finish a pilot, the agreement coefficients must be above the 'good' agreement threshold (0.8). Figure 2 shows the numbers of papers being considered in each step, distinguishing the ones retrieved from a database search (solid line) from the ones retrieved through snowballing (dotted line). The list of the 39 selected papers can be found in the "Appendix A".

4.4 Classification scheme and procedure

We created a classification scheme (Fig. 3) based on our two research areas to obtain all the relevant information to answer our RQs.

Before starting the classification stage, a pilot was performed by all the team members to align the coding criteria and to clear any possible doubt about the categories in the scheme. Once again, Krippendorff's coefficient was used to measure the level of agreement between researchers. Once the coefficient was above 0.8 ('good' agreement) in every category, we moved on to the classification. Each paper was classified by two researchers. A daily meeting was performed to check inconsistencies between coders and to reach agreements between the whole team. When the classification was over, Krippendorffs's coefficient was calculated for all the papers, and in all categories, the value was above 0.8, which is the recommended value.

5 Results

5.1 RQ1—What information is obtained from privacy policies?

RQ1 seeks to provide insight into what information has been automatically extracted from privacy policies by previous research. To this end, we focus on the policy contents and text readability.

5.1.1 Policy contents

Most (nearly 90%) of the papers we have found identified specific content in the privacy policies (Table 4). In this table, the 'Other' group papers focus on contents related to specific privacy laws, i.e., CCPA (ID81) and GDPR (ID136). Remarkably, ID136 provides a GDPR conceptual model and a set of classifiers for identifying all these concepts in policy texts, including, for instance, information about automated decision making.

5.1.2 Policy readability

Only 15% of the papers analyzed focused on how the policy was written. We intentionally excluded papers that only report the use of existing readability scores, as their value is on the conclusions extracted from the application of the score to a given set of policies rather than on the novelty of the technique. However, we found novel approaches that might become useful to improve a privacy policy readability by detecting vagueness and inconsistencies.

Vagueness introduces ambiguous language that undermines the purpose and value of privacy policies as transparency elements. The system described in paper ID996 (2016) is aimed to detect if the words in a privacy policy are vague. Paper ID815 (2018) advances these results by classifying sentences in a privacy policy into different levels of vagueness, resulting in a more complete analysis.

An inconsistent privacy policy introduces contradictions between its contents, thus making it more difficult to understand. The authors of ID175 present a system capable of identifying contradictions in a given privacy policy. To this end, they perform an analysis of the privacy contents included in the text, model these contents, and then check if there is any contradiction between them. Other authors perform similar approaches but comparing the privacy policy of an application with the privacy policies of the software libraries it integrates (ID983) or comparing the privacy policy with the textual description provided by the developer (ID763).

5.2 RQ2—What is the purpose of the policy analysis?

Table 5 shows the different purposes described by the papers analyzed, namely law compliance, system check, user information, or research insight. There are papers fitting more than one purpose (e.g., comparing the policy with the system code and further assessing policy compliance with applicable laws).

Table 4 Papers identifying specific contents of a privacy policy

Policy content	Papers
Controller	ID81, ID136
Data	ID10, ID17, ID19, ID28, ID30, ID33, ID48, ID55, ID60, ID62, ID64, ID65, ID81, ID110, ID136, ID200, ID763, ID770, ID783, ID804, ID805, ID848, ID885, ID983, ID989, ID993, ID1044
Operations	ID10[1], ID17, ID19[1], ID28[1], ID30[1], ID48[1], ID55, ID59, ID62[1], ID64[1], ID72, ID81, ID110[1], ID136[1], ID175, ID763, ID770, ID783[1], ID804[1], ID805[1], ID885[1], ID978[1], ID983, ID993[1], ID1044
Purpose	ID10[1], ID17, ID28[1], ID30[1], ID59, ID62[1], ID64, ID72, ID81, ID110[1], ID136, ID885[1], ID978[1], ID993[1]
Consent	ID10, ID17, ID19, ID28, ID30, ID48, ID59, ID62, ID81, ID136, ID770, ID796, ID885, ID886, ID978, ID993
Access	ID10, ID17, ID19, ID28, ID30, ID48, ID62, ID64, ID81, ID136, ID770, ID885, ID993
Retention	ID10, ID19, ID28, ID30, ID59, ID62, ID72, ID81, ID136, ID885, ID983, ID993
Security	ID10, ID19, ID28, ID30, ID48, ID59, ID62, ID72, ID81, ID136, ID770, ID885, ID993
Change	ID10, ID17, ID19, ID28, ID30, ID48, ID62, ID81, ID885, ID99
Children	ID10, ID17, ID19, ID28, ID30, ID62, ID81, ID136
Cookies	ID59, ID81, ID773
DNT	ID10, ID28, ID30, ID62, ID885, ID993
Other	ID81, ID136

[1]This article further distinguishes the organization role, i.e., first party or third party

Most papers assessing compliance with privacy regulations and laws focus on the GDPR (83.33%). Regardless of legislation, their main focus is on transparency requirements i.e. whether the policy includes mandatory information e.g. controller details or users' rights (see RQ1 for details). Most of them focus on (a few) specific pieces of information, but a salient exception is ID136 that provides a full conceptual model of GDPR transparency requirements, including dependencies between individual elements, which makes this analysis the most exhaustive among all. ID773 is also remarkable, not only checking the presence of some specific information in the

Table 5 Purposes described by the papers analyzed

Law compliance	EU ePrivacy	ID773
	EU GDPR	ID19, ID136, ID770, ID773, ID783
	US CCPA	ID885
System check	Data collection	ID200, ID770, ID773, ID783, ID804, ID848, ID885, ID983, ID989
	Data sharing	ID804, ID885, ID989
User information	Presenting	ID796, ID978, ID1045
	Summarizing	ID17, ID19, ID28, ID59, ID72, ID81, ID805
	Answering	ID28, ID30
Research insight	Data extraction	ID10, ID30, ID33, ID48, ID55, ID60, ID62, ID64, ID65, ID110, ID796, ID886, ID993, ID1044
	Policy characterization	ID175, ID763, ID810, ID815, ID996, ID1018

policy (i.e. purpose in this case) but also assessing its legal compliance as for GDPR and ePrivacy legal rules. Interestingly, only one paper (ID885) refers to collaborations with enforcing authorities to assess their results.

Another group of papers (23.08%) checks whether a software system meets its privacy policy. Particularly, they check if the policy properly declares: (1) the personal data actually collected by the application and (2) the data shared with third parties . These papers mainly rely on static analysis techniques to identify calls to methods retrieving or sending personal information. Static analysis techniques may yield false positives e.g. when a piece of code is never invoked at runtime. Only two papers (ID783, ID804) apply dynamic analysis techniques (e.g. Frida) to observe the actual behaviour of the system.

The language in privacy policies is complex and verbose, and most users do not understand it [5]. A set of papers seek to improve users' understanding of privacy policies, following three different paths: extracting and presenting specific information, summarizing different aspects, , and answering user-posed questions with the contents available in the policy . Some works focus on a given privacy aspect, and extract and present the related information to the user e.g. ID796 presents opt-out choices given in privacy policies to the users in their web browsers. Summaries address more than one aspect, and usually take the form of a set of fixed answers (e.g. yes, no, unknown) to predetermined questions (e.g. whether the system collects personal data). ID805 is remarkable as it provides human-like summaries at different compression ratios by applying risk- or coverage-focused content selection mechanisms. ID28 and ID30 further advance these works by supporting free-from queries that are resolved to specific policy text snippets. However, while the latter requires annotated policies to reason over the former works over previously unseen policies.

Finally, most of the articles do not address any specific stakeholder, but provide new valuable techniques for other researchers to leverage upon. Here we find data extraction techniques focusing on e.g. data types (ID30), data practices (ID62, ID993,

Table 6 NLP techniques applied by the papers analyzed

Symbolic	Morphological and lexical analysis	ID72, ID136, ID773
	Syntactic and semantic analysis	ID48, ID55, ID65, ID110, ID200, ID763, ID848, ID886, ID983, ID1044
	Ontology reasoning	ID30, ID33, ID60, ID175, ID804, ID989
Statistical	Supervised	ID10, ID17, ID19, ID59, ID62, ID64, ID72*, ID81, ID136*, ID770, ID783, ID796, ID810, ID885, ID886*, ID978, ID993
	Unsupervised	ID1018, ID1045
	Artificial Neural Networks	ID10, ID28, ID805, ID810, ID815, ID996

*This article combines symbolic and statistical techniques

ID1044), opt-out statements (ID796, ID886), goals (ID48, ID55, ID110), or several of them (ID10, ID33, ID60, ID62, ID64, ID65). Also, a set of contributions focused on characterizing policies as inconsistent (ID175, ID763), vague (ID815, ID996), or able to answer a specific question (ID810). Still, none of them explicitly apply their results to assess the policy or system under research, or nudge users into privacy aspects, and thus were not included in the other categories.

5.3 RQ3—What techniques have been used to analyse privacy policy texts?

Table 6 shows the two broad categories of NLP techniques used to analyze privacy policy texts reported by the papers analyzed, namely symbolic and statistical. There are contributions that combine techniques from both categories as a pipeline, where the outcome of one technique is the input of the other (ID136), or in parallel (ID72, ID886), combining their outputs to obtain the final result.

5.3.1 Symbolic NLP approaches

As detailed in Table 6, contributions apply symbolic NLP at three different levels: morphological and lexical analysis of the words, syntactic and semantic analysis, and using ontologies to extract the meaning of the sentences.

The first levels of NLP (morphological and lexical analysis) have similar results to more complex techniques when targeting certain privacy practices. There are privacy practices (e.g., those related to encryption) that often use very specific distinctive terms (e.g., SSL). In such cases, a basic keyword-based analysis performs best (e.g., see ID72).

Most symbolic NLP techniques use some form of semantic analysis. The typical procedure followed in these cases consists of five phases: (1) splitting the policy into sentences, (2) parsing the words, e.g. through Part-of-Speech (PoS) tagging, (3) eliciting syntactic patterns related to a privacy practice, such as collection or disclosure,

(4) detecting these patterns, and (5) deriving semantic meaning from them. The main differences between the authors are the proprietary tools or techniques implemented to carry out these tasks and the lexicons or taxonomies used.

We would like to highlight some of the most useful tools found in the study. The tool most used for carrying out syntactic analysis is the Stanford dependency parser, which is available in five different languages. One of the most critical stages is the creation of semantic patterns, which many authors manually create based on collections of privacy policies or taxonomies such as that created by Anton et al. [21]. By contrast, the authors of paper ID983 use a bootstrapping mechanism from Slankas et al. [22] to automatically find patterns from privacy policies according to a simple seed pattern. This process allows them to generate more inclusive patterns. Finally, there are some papers that report the use of specific programming languages to make annotations in the text and finding the patterns. This is the case of paper ID55 that uses Eddy [23], and paper ID1044 that uses Jape [24].

In our research, one in three papers using symbolic NLP techniques report the use of ontologies. Once a privacy policy text is represented as an ontology, information can be automatically extracted with query languages such as SPARQL. The most challenging part of the use of ontologies is the definition and the creation of the ontology. In most of the cases (50%), the creation of the ontology is a manual process conducted by a group of experts that annotate the privacy policies texts. Although this is the first step, what is really interesting is the automatic creation of the ontologies, which would allow researchers to analyze large amounts of policies without human intervention. Some examples of this are papers ID33, ID60, and ID175, each of which use a different approach. The authors of paper ID33 use Tregex patterns to detect information types automatically. In the case of ID60, they use semantic rules to extract relationships from information types. Finally, the authors of paper ID175 have created a method to "capture both positive and negative statements of data collection and sharing practices" based on Hearst patterns. This paper is relevant due to its ability to detect negative statements in comparison with other more limited approaches like those used by Zimmeck et al. in ID885 and by Yu et al. in ID983.

5.3.2 Statistical NLP approaches

As Table 6 shows, contributions based on Statistical NLP use supervised (60%), unsupervised (7%), ANN-based techniques (26%), or a combination of them (7%). While supervised techniques are mostly used, ANN have begun to gain strength since their appearance around 2016.

Supervised algorithms are primarily used for classification tasks, such as detecting which personal data is collected, while unsupervised algorithms are used for clustering tasks such as topic modeling. As for the supervised algorithms, geometric algorithms such as Support Vector Machine (SVM) (ID59, ID62, ID64) and Logistic Regression (LR) (ID886, ID978, ID993) are mainly applied. Decision tree-based models are also used, like Decision Tree (DT) (ID81), the ensembles Random Forest (RF) (ID770, ID783), and AdaBoost (ID783), which tends to outperform the results of DT.

Unsupervised learning techniques apply Hidden Markov Models (HMM) to group segments of policies based on the privacy topic they address (ID1018) and a Latent

Dirichlet Allocation (LDA) algorithm to determine the underlying topics in privacy policies (ID1045). Although one of the most interesting attributes of these approaches is the absence of a labeled corpus, it is important to highlight that in both cases, the authors had to create a labeled dataset to evaluate the accuracy of their models.

ANN-based techniques have been applied to tasks such as text classification (ID10), answerability prediction (ID810), or vagueness identification (ID815). We have found different approaches and papers using different kinds of neural networks. Convolutional Neural Networks (CNNs) (ID28, ID805), Recurrent Neural Networks (RNN) (ID996) and Google's algorithm BERT (Bidirectional Encoder Presentations from Transformers) (ID796, ID810) are mostly used. Certain works comparing ANN and supervised learning techniques (ID10, ID810) report better performance in the case of ANN-based predictions.

Another important aspect in ANN is the technique used to represent every word so that it can be used by the neural network. The analyzed papers have used different techniques to create this word representation: fastTest (ID28), Word2Vec (ID810), Glove (ID815, ID996) and ELMo (ID805). These tools can create a word representation from a pre-trained model or from a specific model trained for the occasion. Authors seem to agree that training the model with a related corpus improves the results.

5.3.3 Annotated privacy policies datasets

Learning algorithms require annotated datasets for training or validation. The creation of this dataset is a time-consuming task. On the other hand, these datasets are of paramount importance since the results and performance of the final model depend on the quality and the completeness of this data. Table 7 collates the information on all the public datasets of annotated privacy policies found in our research.

6 Discussion

The analysis of privacy policies seems to be a promising area of research. Having found the first work published in 2005, we have identified a growing interest especially in the last five years in which we found 72% of the papers published (Fig. 4).

This increasing interest might have been boosted by the adoption of the European GDPR in 2016, aligned with a stepped increase in publications. Indeed, we found that all papers explicitly focusing on law compliance have been published since 2016, and two-thirds of them target the GDPR. This evidence is aligned with the findings of previous work that highlights that the GDPR has inspired different privacy legislations worldwide [25] and its greatest impact on privacy assessment research [13].

Overall, the identification of the policy contents shows a good performance. Our findings reveal a preference for classical ML techniques (i.e. non-ANN-based) for analyzing policy texts (Table 6), yet the use of ANN-based techniques is quite recent (Fig. 5). Researchers usually train different classifiers and compare their results selecting the one that demonstrates better performance for the problem at hand. However, some of the papers applied both approaches and compared them (ID10, ID810). The authors of ID10 highlight that ANN-based models favor precision while other models favor recall, this may be taken into consideration according to the type of task at hand.

Table 7 Annotated privacy policies datasets

Dataset name	Description	Num. of items	Papers
OPP115[1]	Website Privacy Policies annotated with 10 privacy practices	115	ID10, ID28, ID30, ID885, ID886, ID978, ID993
APP350[1]	Android App Privacy Policies annotated with the information accessed by and shared with different parties	350	ID783
Opt-out Choice WWW[1]	Website Privacy Policies with annotated links of opt-out choices	236	ID796
ACL/ COLING[1]	Website Privacy Policies annotated with personal information accessed and shared	1010	ID996
Vagueness Data[2]	Sentences from Privacy Policies with vague words annotated and the global level of vagueness	>100,000	ID815
PP Summaries[3]	Risk Level and Summaries of Privacy Policies segments	151	ID805
PrivacyQA[4]	Questions about the contents of Privacy Policies	1750	ID810
ToS;DR[5]	Terms of service ranked from A to E based on user rights	>1000	ID805

[1] https://www.usableprivacy.org/data
[2] https://loganlebanoff.github.io/data/vagueness_data.tar.gz/
[3] https://github.com/senjed/Summarization-of-Privacy-Policies
[4] https://github.com/AbhilashaRavichander/PrivacyQA_EMNLP
[5] https://tosdr.org/en/frontpage

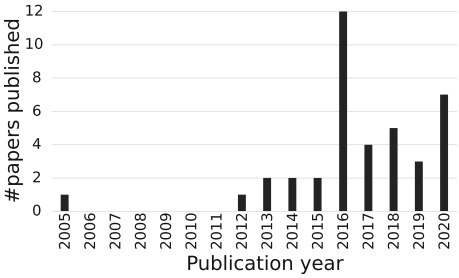

Fig. 4 Distribution of publications per year

The authors of ID810 use three different approaches for answerability prediction and answer sentence selection, namely SVM, CNN and BERT. Their results show that BERT achieves the best F1-Score in both tasks. This fact suggests there is room for improvement in the research of privacy policy texts using ANN-based approaches (e.g., deep learning).

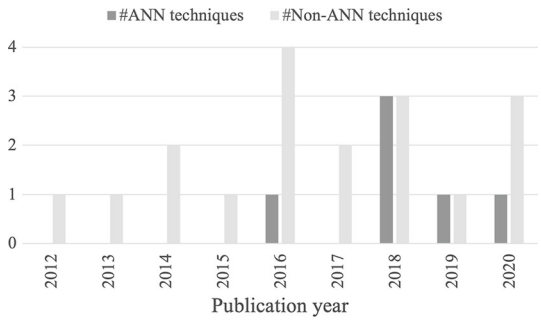

Fig. 5 Distribution of ANN and non-ANN techniques per year

6.1 Research challenges

Named-entity recognition is needed to allow for fully automated analysis at scale We found ample coverage on identifying the presence/absence of the different contents expected in a privacy policy (Table 4). However, oftentimes they do not obtain its specific value. For example, many researchers detect the presence of information on data retention time (which is useful in assessing policy completeness) but not the specific retention time (which would be useful for automatically assessing privacy risks due to excessive retention time). Named-entity recognition would support automated workflows to first identify if a given content is present and then extract and assess its specific value.

The analysis of specific policy contents still requires further research, particularly, those mandated only by specific privacy laws Online products and services are offered worldwide, and their policies must meet the requirements set by all the privacy legislations where they are consumed. A clear example is the transfers of personal data to other countries or international organizations. While the CCPA does not restrict them, the GDPR and PIPL set detailed requirements. As a result, future work is needed to identify more specific policy contents, so that policy compliance can be automatically assessed against different applicable laws. To this end, the techniques must be generalized to other languages, as all contributions found focus on policies written in English.

Context must be considered to improve the privacy analysis Many articles focus on identifying contents in isolation but do not investigate the relationship between them. For example, gathering a list of personal information types being collected yields less utility than contextualizing an information flow including the personal information type, the associated data processing operation, the organization carrying it out, and the purpose for it. Future work is expected to contextualize data processing practices, particularly to improve users' awareness and understanding of privacy risks. Also, new contributions are needed to analyze not just one but different inter-related policies to detect inconsistencies among them e.g. between a 1st party policy and all its 3rd parties collecting and processing data.

Integrating the results into tools for the benefit of different stakeholders More than 50% of the papers found do not apply their results to support any specific stakeholder, describing that as a future work. The contributions identified can support different stakeholders e.g. end-users in gaining awareness and understanding through privacy scores and summaries, legal counsels to clarify the legal texts provided by organiza-

tions processing personal data, developers to spot potential non-compliance earlier in the development processes, and app stores to improve their app vetting processes. However, future work is needed to increase the maturity of the techniques and integrate them into user-oriented tools.

7 Threats to validity

The main threat to the construct validity of a mapping study is that the research questions may not completely cover all of the aspects addressed by the studied publications. To deal with this threat, an annotation scheme was created by experts in the field based on known taxonomies and classifications being iteratively updated until it was able to identify all the essential aspects of the selected papers.

Another threat for the construction validity is a bias at the encoding stage. Different actions were taken to avoid this threat. First, the classes and values included in the encoding scheme were discussed by all the team members until a common understanding of all the covered aspects was reached. Second, a pilot of the codification scheme was carried out with 10% of the publications, and Krippendorf's coefficient was measured to assess the agreement between coders. All the measured questions values were above the threshold of a good agreement (0.8). Finally, two researchers analyzed and coded each paper, and their codifications were compared to avoid failure. All team members discussed inconsistencies until an agreement was reached.

We addressed the threats to the internal validity by identifying all the publications matching our criteria and creating the more unbiased process possible. First, two different databases were selected, namely, Scopus and Web of Science, as they complement each other by indexing different journals and conference papers [18]. Second, the search strings are based on known taxonomies and classifications such as the IEEE Thesaurus. Furthermore, a group of ten articles, identified by the experts as matching the criteria, were used to assess the completeness of the search string, which evolved until it matched with all the selected articles. Third, a snowballing technique was performed to include all cited papers and all papers that cited them; this technique is particularly useful for expanding the coverage of a systematic mapping study [19]. This step ultimately ensures that related papers are reviewed despite using other terms to refer to the main topics.

Once all the papers were obtained from the databases, inclusion, and exclusion criteria were applied. The number of citation criteria was created considering the percentile of papers in computer science, as per Thomson Reuters, which is a reliable source of publication relevance. The number of paper criteria was established taking into consideration the characteristics of short papers that normally do not include a validation section. The manual criteria were defined based on the definition of the scope. Their formulation allows the researchers to specify which cases are included and which are not.

Two researchers reviewed each publication to ensure that the bias of one of the researchers did not affect the selection process. A pilot, divided into phases, was performed to assess the coders' agreement. In each stage, five papers were analyzed by all team members, and Krippendorf's coefficient was measured. The inclusion/exclusion

phase started once the coefficient value was above the threshold of a good agreement (0.8).

After the extraction of all the necessary information of the selected publications in the codification stage, the data was analyzed to obtain aggregate results and conclusions. One of the researchers cleaned and aggregated the data to present it to the rest of the group. The meaning of the results and the more relevant aspects were discussed by all the team members until an agreement was reached. Therefore, all the results and conclusions presented came from common agreements and not from individual thoughts.

The external validity of this study is determined by its scope, the intersection between privacy policies,and text analysis techniques. Any other article that does not concern these two topics may affect the generalization of the results, and so the conclusions reached are not applicable to them. Accordingly, conclusions reached do not apply to publications on the generation of privacy policies, publications on the analysis of other kinds of texts, or publications solely containing the manual process of creation of a dataset of labeled privacy policies.

8 Conclusion

This paper has identified, classified, and analyzed the existing approaches and techniques to analyze privacy policies automatically. As a result, it provides an overview of the contents that can be automatically extracted from privacy policies and the most promising analysis techniques for each task. These techniques have been applied to check the policy's compliance with applicable privacy laws and the system compliance with its privacy policy, as well as to improve the user awareness and understanding of the privacy risks.

Our future work is focused on the exhaustive compliance analysis of transparency requirements mandated by privacy laws. For that, we will leverage and combine privacy policy analysis techniques with system behavior analysis techniques, to compare both results and detect legal breaches.

Acknowledgements This work was partially supported by the Comunidad de Madrid and Universidad Politécnica de Madrid through the V-PRICIT Research Programme Apoyo a la realización de Proyectos de I+D para jóvenes investigadores UPM-CAM, under Grant APOYO-JOVENES-QINIM8-72-PKGQ0J. The work of Danny S. Guaman was supported by Escuela Politécnica Nacional under the PII-DETRI-2021-06 research project. This work was supported in part by the Ministerio de Ciencia e Innovación-Agencia Estatal de Investigación, through the H2O Learn project under Grant PID2020-112584RB-C31, and in part by the Madrid Regional Government through the e-Madrid-CM Project under Grant S2018/TCS-4307.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Appendix A: List of examined publications

Paper ID	Reference
ID10	S. Wilson et al.: Analysing privacy policies at scale: From crowdsourcing to automated annotations. ACM Trans. Web 13:1, 2018. doi: 10.1145/3230665
ID17	R. Nokhbehv et al.: PrivacyCheck: Automatic summarization of privacy policies using data mining. ACM Trans. Internet Tech. 18:4, 2018. doi: 10.1145/3127519
ID19	W.B. Tesfay et al.: Privacyguide: Towards an implementation of the EU GDPR on internet privacy policy evaluation. 4th ACM Int. Workshop on Sec. and Privacy Analytics, 2018. doi: 10.1145/3180445.3180447
ID28	H. Harkous et al.: Polisis: Automated analysis and presentation of privacy policies using deep learning. 27th USENIX Sec. Symp., 2018
ID30	A. Oltramari et al.: PrivOnto: A semantic framework for the analysis of privacy policies. Semant. Web 9:2, 2018. doi: 10.3233/SW-170283
ID33	M.C. Evans et al.: An Evaluation of Constituency-Based Hyponymy Extraction from Privacy Policies. 25th Int. Req. Eng. Conf., 2017. doi: 10.1109/RE.2017.87
ID48	R.L. Rutledge et al.: Privacy impacts of IoT devices: A SmartTV case study. 24th Int. Req. Eng. Conf. Workshops, 2017. doi: 10.1109/REW.2016.40
ID55	J. Bhatia et al.: Mining privacy goals from privacy policies using hybridized task recomposition. ACM Trans. Softw. Eng. Methodol. 25:3, 2016. doi: 10.1145/2907942
ID59	N. Guntamukkala et al.: A machine-learning based approach for measuring the completeness of online privacy policies. 14th Int. Conf. Machine Learning and Apps., 2016. doi: 10.1109/ICMLA.2015.143
ID60	M.B. Hosseini et al.: Lexical similarity of information type hypernyms, meronyms and synonyms in privacy policies. AAAI Fall Symposia, 2016
ID62	S. Wilson et al.: The creation and analysis of a Website privacy policy corpus. 54th Annual Meeting of the ACL. doi: 10.18653/v1/p16-1126
ID64	A.R. Da Silva et al.: Improving the specification and analysis of privacy policies: The RSLingo4Privacy approach, 18th Int. Conf. on Enterprise Inf. Systems, 2016. doi: 10.5220/0005870503360347
ID65	J. Bhatia, T.D. Breaux: Towards an information type lexicon for privacy policies, 8th Int. Workshop on Req. Eng. and Law, 2015. doi: 10.1109/RELAW.2015.7330207
ID72	S. Zimmeck and S.M. Bellovin: Privee: An architecture for automatically analysing web privacy policies, 23rd USENIX Sec. Symp., 2014
ID81	E. Costante et al.: A machine learning solution to assess privacy policy completeness, ACM Conf. on Comp. and Comm. Sec., 2012. doi: 10.1145/2381966.2381979
ID110	T.D. Breaux, A.I. Antón: Analysing goal semantics for rights, permissions, and obligations, Int. Conf. on Req. Eng., 2005. doi: 10.1109/re.2005.12
ID136	D. Torre et al.: An AI-Assisted Approach for Checking the Completeness of Privacy Policies against GDPR, IEEE Int. Conf. on Req. Eng., 2020. doi: 10.1109/RE48521.2020.00025
ID175	B. Andow et al.: Policylint: Investigating internal privacy policy contradictions on google play, 28th USENIX Sec. Symp., 2019
ID200	L. Yu et al.: Revisiting the Description-to-Behavior Fidelity in Android Applications, Int. Conf. on Softw. Analysis, Evolution and Reengineering, 2016. doi: 10.1109/saner.2016.67

Paper ID	Reference
ID763	S. Liao et al.: Measuring the Effectiveness of Privacy Policies for Voice Assistant Applications, Annual Comp. Sec. Apps. Conf., 2020. doi: 10.1145/3427228.3427250
ID770	M. Fan et al.: An empirical evaluation of GDPR compliance violations in android mhealth apps, Int. Symp. on Softw. Reliability Eng., 2020. doi: 10.1109/ISSRE5003.2020.00032
ID773	I. Fouad et al.: On Compliance of Cookie Purposes with the Purpose Specification Principle, 5th IEEE EuroS&P Workshops, 2020. doi: 10.1109/EuroSPW51379.2020.00051
ID783	L. Verderame et al.: On the (Un)Reliability of Privacy Policies in Android Apps, Int. Joint Conf. on Neural Networks, 2020. doi: 10.1109/IJCNN48605.2020.9206660
ID796	V. Bannihatti Kumar et al.: Finding a Choice in a Haystack: Automatic Extraction of Opt-Out Statements from Privacy Policy Text, World Wide Web Conf., 2020. doi: 10.1145/3366423.3380262
ID804	B. Andow et al., Actions speak louder than words: Entity-sensitive privacy policy and data flow analysis with policheck, 29th USENIX Sec. Symp., 2020
ID805	M. Keymanesh et al.: Toward domain-guided controllable summarization of privacy policies, Nat. Legal Lang. Processing Workshop, 2020
ID810	A. Ravichander et al.: Question answering for privacy policies: Combining computational and legal perspectives, Conf. on Empirical Methods in Nat. Lang. Processing, 2020. doi: 10.18653/v1/d19-1500
ID815	L. Lebanoff and F. Liu: Automatic detection of vague words and sentences in privacy policies, Conf. on Empirical Methods in Nat. Lang. Processing, 2020. doi: 10.18653/v1/d18-1387
ID848	L. Yu et al.: Enhancing the Description-to-Behavior Fidelity in Android Apps with Privacy Policy, IEEE Trans. Softw. Eng. 44:9, 2018. doi: 10.1109/TSE.2017.2730198
ID885	S. Zimmeck et al.: Automated analysis of privacy requirements for mobile apps, Netw. and Distributed System Sec. Symposium, 2017. doi: 10.14722/ndss.2017.23034
ID886	K.M. Sathyendra et al.: Automatic extraction of opt-out choices from privacy policies, in AAAI Fall Symposium, 2016
ID978	K. M. Sathyendra et al.: Identifying the provision of choices in privacy policy text, Conf. on Empirical Methods in Nat. Lang. Processing, 2017. doi: 10.18653/v1/d17-1294
ID983	L. Yu et al.: Can we trust the privacy policies of android apps?, 46th Annual IEEE/IFIP Int. Conf. on Dependable Systems and Netw., 2016. doi: 10.1109/DSN.2016.55
ID989	R. Slavin et al.: Toward a framework for detecting privacy policy violations in android application code, 38th Int. Conf. on Softw. Eng., 2016. doi: 10.1145/2884781
ID993	F. Liu et al.: Analysing vocabulary intersections of expert annotations and topic models for data practices in privacy policies, in AAAI Fall Symposium, 2016
ID996	F. Liu et al.: Modeling language vagueness in privacy policies using deep neural networks, in AAAI Fall Symposium - Technical Report, 2016
ID1018	F. Liu et al.: A step towards usable privacy policy: Automatic alignment of privacy statements, 25th Int. Conf. on Computational Linguistics, 2014

Paper ID	Reference
ID1044	E. Costante et al.: What websites know about you: Privacy policy analysis using information extraction, LNCS 7731, 2013
ID1045	A.K. Massey et al.: Automated text mining for requirements analysis of policy documents, 21st IEEE Int. Req. Eng. Conf., 2013. doi: 10.1109/RE.2013.6636700

References

1. Morel V, Pardo R (2020) SoK: three facets of privacy policies. In: Workshop on privacy in the electronic society, Virtual, France. https://hal.inria.fr/hal-02267641
2. Wilson S et al (2016) The creation and analysis of a website privacy policy corpus. In: Proceedings of the 54th annual meeting of the association for computational linguistics, pp 1330–1340
3. Acar G et al (2014) The web never forgets: persistent tracking mechanisms in the wild. In: Proceedings of the 2014 ACM SIGSAC conference on computer and communications security, pp 674–689
4. Fielding R, Singer D (2019) Tracking preference expression (DNT). W3C note, W3C (January). https://www.w3.org/TR/2019/NOTE-tracking-dnt-20190117/
5. McDonald AM, Cranor LF (2008) The cost of reading privacy policies. Isjlp 4:543
6. Regulation GDP (2016) Regulation EU 2016/679 of the European Parliament and of the Council of 27 April 2016
7. Indurkhya N, Damerau FJ (2010) Handbook of natural language processing. Chapman & Hall/CRC, Cambridge
8. Moisl H (2000) Nlp based on artificial neural networks: introduction. In: Dale R, Moisl HSH (eds) Handbook of natural language processing. Marcel Dekker, New York, pp 655–713
9. Sun S, Luo C, Chen J (2017) A review of natural language processing techniques for opinion mining systems. Inf Fus 36:10–25
10. Kreimeyer K et al (2017) Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. J Biomed Inform 73:14–29
11. Nazir F, Butt WH, Anwar MW, Khattak MAK (2017) The applications of natural language processing (NLP) for software requirement engineering-a systematic literature review. In: International conference on information science and applications. Springer, Berlin, pp 485–493
12. Kang Y et al (2020) Natural language processing (NLP) in management research: a literature review. J Manag Anal 7(2):139–172
13. Guamán DS, Del Alamo JM, Caiza JC (2020) A systematic mapping study on software quality control techniques for assessing privacy in information systems. IEEE Access 8:74808–74833
14. Murmann P, Fischer-Hübner S (2017) Tools for achieving usable ex post transparency: a survey. IEEE Access 5:22965–22991
15. Becher S, Gerl A, Meier B (2020) Don't forget the user: from user preferences to personal privacy policies. In: 10th international conference on advanced computer information technologies. IEEE, pp 774–778
16. Kirrane S, Villata S, d'Aquin M (2018) Privacy, security and policies: a review of problems and solutions with semantic web technologies. Semantic Web 9(2):153–161
17. Petersen K, Vakkalanka S, Kuzniarz L (2015) Guidelines for conducting systematic mapping studies in software engineering: an update. Inf Softw Technol 64:1–18
18. Mongeon P, Paul-Hus A (2016) The journal coverage of web of science and scopus: a comparative analysis. Scientometrics 106(1):213–228
19. Wohlin C (2014) Guidelines for snowballing in systematic literature studies and a replication in software engineering. In: Proceedings of the 18th international conference on evaluation and assessment in software engineering, pp 1–10
20. Krippendorff K (2004) Reliability in content analysis: some common misconceptions and recommendations. Hum Commun Res 30(3):411–433
21. Antón AI et al (2007) Hipaa's effect on web site privacy policies. IEEE Secur Privacy 5(1):45–52

22. Slankas J, Xiao X, Williams L, Xie T (2014) Relation extraction for inferring access control rules from natural language artifacts. In: Proceedings of the 30th annual computer security applications conference, pp 366–375
23. Breaux TD, Hibshi H, Rao A (2014) Eddy, a formal language for specifying and analyzing data flow specifications for conflicting privacy requirements. Requir Eng 19(3):281–307
24. Cunningham H, Maynard D, Tablan V (1999) Jape: a java annotation patterns engine
25. Woodward M (2021) GDPR has inspired different privacy legislations worldwide. https://securityscorecard.com/blog/countries-with-gdpr-like-data-privacy-laws