AI-enabled Automation for Completeness Checking of Privacy Policies

Orlando Amaral, Sallam Abualhaija, *Member, IEEE,* Damiano Torre, *Member, IEEE,* Mehrdad Sabetzadeh, *Member, IEEE,* and Lionel C. Briand, *Fellow, IEEE,*

Abstract—Technological advances in information sharing have raised concerns about data protection. Privacy policies contain privacy-related requirements about how the personal data of individuals will be handled by an organization or a software system (e.g., a web service or an app). In Europe, privacy policies are subject to compliance with the General Data Protection Regulation (GDPR). A prerequisite for GDPR compliance checking is to verify whether the content of a privacy policy is complete according to the provisions of GDPR. Incomplete privacy policies might result in large fines on violating organization as well as incomplete privacy-related software specifications. Manual completeness checking is both time-consuming and error-prone. In this paper, we propose AI-based automation for the completeness checking of privacy policies. Through systematic qualitative methods, we first build two artifacts to characterize the privacy-related provisions of GDPR, namely a conceptual model and a set of completeness criteria. Then, we develop an automated solution on top of these artifacts by leveraging a combination of natural language processing and supervised machine learning. Specifically, we identify the GDPR-relevant information content in privacy policies and subsequently check them against the completeness criteria. To evaluate our approach, we collected 234 real privacy policies from the fund industry. Over a set of 48 unseen privacy policies, our approach detected 300 of the total of 334 violations of some completeness criteria correctly, while producing 23 false positives. The approach thus has a precision of 92.9% and recall of 89.8%. Compared to a baseline that applies keyword search only, our approach results in an improvement of 24.5% in precision and 38% in recall.

Index Terms—Requirements Engineering, Legal Compliance, Privacy Policies, The General Data Protection Regulation (GDPR), Artificial Intelligence (AI), Conceptual Modeling, Qualitative Research.

1 Introduction

Advances in information sharing technologies have raised concerns about protecting the privacy of individuals. In Europe and indeed worldwide, the General Data Protection Regulation (GDPR) [1] is widely viewed as a benchmark for data protection and privacy regulations. GDPR harmonizes data privacy laws across the European Economic Area (EEA), providing further protection to individuals for controlling their personal data in the face of new technological developments [2].

While undoubtedly beneficial to individuals in many ways, the reality is that organizations are having considerable difficulty complying with GDPR [3]. There is thus a pressing need for cost-effective methods that can help different organizations better deal with privacy considerations. This need has not gone unnoticed by the research community. For example, Perrera et al. [4] propose systematic guidance to help software engineers develop privacy-aware applications; Torre et al. [5], [6] propose the use of Model-driven Engineering as a basis for GDPR compliance automation; and Ayala-Rivera and Pasquale [7] present a step-wise approach for eliciting requirements related to GDPR compliance.

To comply with GDPR, organizations need to take into account the principles of personal data processing set out in the regulation, and to regularly review their measures, practices and processes related to the collection, use and protection of personal data. Compliance also entails that software systems storing or processing personal data should properly implement privacy-related GDPR requirements. Every organization, whether Europe-based or not, which is collecting, processing or in some way handling the personal data of European citizens and residents must comply with GDPR.

In this paper, we concern ourselves with GDPR *privacy policies*. A privacy policy can be viewed as a technical document stating the multiple privacy-related requirements that an organization (including processes, services, developed systems) should satisfy in order to help users make informed decisions about the data that this organization may collect and use. In other words, a privacy policy explains how an organization handles personal data and how it applies the principles of GDPR. Privacy policies are usually defined through natural-language statements. Natural language (NL) is an ideal medium for expressing privacy policies since it is flexible and universal [8]. Though NL is advantageous for establishing a common understanding, processing NL documents is challenging due to common

- *O.Amaral, S. Abualhaija, D. Torre, M. Sabetzadeh and L.C. Briand are with the SnT Centre for Security, Reliability, and Trust, University of Luxembourg, Luxembourg.*
  E-mail: {orlando.amaralcejas, sallam.abualhaija, mehrdad.sabetzadeh, lionel.briand}@uni.lu
- *M. Sabetzadeh and L.C. Briand are also affiliated with the school of Electrical Engineering and Computer Science, University of Ottawa, Canada.*
  E-mail: {m.sabetzadeh, lbriand}@uottawa.ca
- *D. Torre is also affiliated with the Department of Computer Information Systems, Texas A&M University – Central Texas, United States.*
  E-mail: damiano.torre@tamuct.edu

Manuscript received Month DD, 2021; revised Month DD, 2021.

quality issues such as ambiguity, incompleteness and inconsistency [9].

This paper tackles an important dimension of GDPR compliance checking for privacy policies. Specifically, in collaboration with legal experts from Linklaters (a multinational law firm), we develop an AI-enabled approach for checking whether a given privacy policy is "complete" according to the provisions of GDPR. We use the term "complete" rather than "compliant" to signify the fact that our approach can detect only the presence (or absence) of the information content types that GDPR envisages for privacy policies. A privacy policy is deemed "complete" according to GDPR if it explicitly contains certain information mandatory for ensuring data protection and privacy rights, e.g., about the rights individuals have over their personal data. We further clarify this concept with an example in the next section. According to our collaborating legal experts, *completeness checking* is an essential prerequisite for compliance checking. Manually checking completeness is however both time-consuming and error-prone. Providing automated support is thus desirable so that legal experts can focus their effort on more critical tasks.

1.1 Practical Scenario

In practice, completeness checking of privacy policies against GDPR can be beneficial to a diverse group of legal experts, software engineers, and other business stakeholders. The first step in completeness checking is to determine if GDPR-relevant information content is present or not in a given privacy policy. Based on the above analysis, the second step is then to map what is actually present in the privacy policy to what must be present according to the provisions of GDPR. In the rest of this paper, we will use the term *metadata type* to describe any information type which we extracted from the privacy-related provisions of GDPR. Some of these metadata types are mandatory and thus have a direct impact on completeness. We elaborate how we combine the metadata types for checking the completeness of a privacy policy in Sec. 4. A comprehensive description of these metadata types is provided in Sec. 3.

Examples of metadata types include: PROCESSING PURPOSES to characterize the purposes of the processing for which personal data is being collected, LEGAL BASIS to capture the legal basis for the processing of personal data, and DATA SUBJECT RIGHT to mark the clause(s) giving an individual the rights in relation with their personal data. Under DATA SUBJECT RIGHT, several specializations are listed to describe the different rights an individual has. For instance, DATA SUBJECT RIGHT.ACCESS is concerned with the right to request access to the personal data from the controller. The specializations of the metadata types are represented, throughout the paper, with a *dot*. Fig. 1 shows a complete privacy policy that is annotated with all metadata types (from Sec. 3). In the figure, we present the metadata types using numbers (further explained in the legend), and square brackets to delineate the text corresponding to the metadata types. For example, number **15** in Fig. 1 refers to the metadata type PROCESSING PURPOSES, numbers **20 – 25** refer to different specializations of LEGAL BASIS, and number **27** refers to DATA SUBJECT RIGHT.ACCESS.

To deem the example privacy policy in Fig. 1 complete, GDPR requires the presence of multiple mandatory metadata types, including the ones concerning CONTROLLER, i.e., the organization which collects personal data (GDPR, Art. 13 and Art. 14(f)). In particular, the policy should include the identity (i.e., CONTROLLER.IDENTITY) and contact details of the controller (i.e., CONTROLLER.CONTACT). As we see in Fig. 1, these two metadata types are mentioned respectively in number **1**, and numbers **4 – 6**. Verifying the presence of the metadata types about CONTROLLER is however not sufficient, and verification of other metadata types is needed in order to make the final decision as to whether the privacy policy is complete according to GDPR.

Legal provisions in GDPR can contain requirements which depend on one another. Consequently, the presence of certain metadata types in a privacy policy may necessitate the presence of certain other metadata types in the policy. For instance, if a privacy policy states that the legal basis for the processing of personal data is based on individual consent (i.e., LEGAL BASIS.CONSENT), then the right to withdraw this consent should be granted in the same policy (i.e., DATA SUBJECT RIGHT.WITHDRAW CONSENT). These metadata types correspond to two different GDPR articles, Art. 6.1(a) and Art. 13.2(c), respectively. Information related to these types can be found by reviewing paragraphs that are usually located in different parts of the privacy policy. In Fig. 1, LEGAL BASIS.CONSENT is mentioned in the text: *in accordance with applicable law, based in your consent [...]* (number **22**), while DATA SUBJECT RIGHT.WITHDRAW CONSENT is mentioned in: *the right to withdraw your consent [...]* (number **33**). If done manually, this back-and-forth reviewing of the text requires a considerable amount of effort and time in practice.

Checking the completeness of a given privacy policy according to GDPR is essential for ensuring the completeness of the privacy-related software requirements induced by the policy. To illustrate, consider our example in Fig. 1. Since the CONTROLLER (i.e., Hikari Bank Ltd – number **1**) is located in Japan, it is likely that the personal data of the bank's customers will be transferred outside the Europe. Articles 13.1(a), 13.1(f) and 14.1(f) in GDPR enforce requirements to ensure the protection of personal data, for example when transferred outside Europe. The implications of these articles are then two-fold. On one hand, the privacy policy must provide information about the IDENTITY and CONTACT details of the CONTROLLER REPRESENTATIVE (who has to be located in Europe) – as shown in numbers **2** and **3**, respectively. The policy must also state the legal agreement that is in place for transferring data to Japan such as the Act on the Protection of Personal Information (APPI) – provided in number **16**. On the other hand, the software developed to handle such personal data (e.g., the online banking service) has to comply with Japan's data protection law. APPI-compliant software should provide a response to the individuals' requests in relation with their personal data within two weeks. Otherwise, an individual can sue the controller. An incomplete privacy policy due to the missing metadata types related to the location where the controller is based (in our case, Japan) or to the legal agreement used for transferring data (i.e., APPI) fails to comply with GDPR. The missing metadata types will remain unknown for a

Hikari Bank Privacy Policy

By accepting this policy, you are providing personal data (as defined below) to 1 [Hikari Bank Ltd], represented by the 2 [Holding Bank Services], 3 [16, rue de Gasperich, L-5826 Hesperange, Grand Duchy of Luxembourg.] If you have questions or concerns about this policy, please contact us by post: 4 [20 Nihonbashi Honcho, Tokyo 103-8691, Japan]; 5 [by email: info@hikari.jp]; 6 [or by telephone: +81 3 36300941.]
We collect your personal data from: 7 [information you provide to us verbally, electronically or in writing]; 8 [information obtained from public bodies] 9 [that includes passport, identification card, tax identification number, national insurance number, social security number]; 10 [information obtained from third parties including your employer, credit reference agencies, law enforcement authorities]; 11 [or information obtained through cookies.] 9 [We will hold some or all of the following types of personal data: given name(s); gender; date of birth / age; marital status; social security number; passport number(s); nationality; images of passports; images of driving licences; images of signatures; authentication data (passwords, mother maiden name, face recognition, voice recognition)]
12 [in addition to some sensitive data, such as medical history, criminal convictions and religious beliefs.] 13 [Any personal data will be held for a period of up to 3 years after the termination of the relationship between you and the Hikari bank and in any event no longer than necessary with regard to the purpose of the data processing or as required by law.]14 [Your personal data might be disclosed to the tax authorities, or other third parties including legal or financial advisors, regulatory bodies, auditors and technology providers.] 15 [The purposes for which we may process personal data include processing subscription, redemption and conversion orders, as well as processing payments of dividends and other distributions.]
16 [We may also transfer your personal data to countries outside of European Union (including Japan) on the basis of: (i) European Commission's adequacy decisions, certified by the APPI Japan scheme]; 17 [(ii) our binding corporate rules]; 18 [(iii) suitable standard contractual clauses.] 19 [By accepting this policy, you expressly consent the processing of your personal data by Hikari Bank and any of the group companies if fits or applies to a vacancy outside the European Union.]
20 [The legal bases on which we may perform data processing, are: (i) For compliance with a legal obligation (e.g., to comply with our diversity reporting obligations)]; 21 [(ii) for the detection or prevention of crime (including the prevention of fraud) to the extent permitted by applicable law]; 22 [(iii) in accordance with applicable law, based on your consent prior to processing your sensitive personal data]; 23 [(iv) for reasons of substantial public interest and occurs on the basis of an applicable law that is proportionate to the aim pursued and provides for suitable and specific measures to safeguard your fundamental rights and interests]; 24 [(v) for protecting the vital interests of any individual]; 25 [or (vi) for issuing any contract that you may enter into with us, or to take steps prior to entering into a contract with us.]
26 [Where the processing of your personal information is for contractual purposes as outlined in this privacy notice, but you fail to provide us with the personal information required, then this may result in Hikari Bank not being able to offer you with our services.] 27 [Subject to applicable law, you have the following rights regarding the processing of your personal data: (i) the right to access your personal data] and 28 [the right to rectify any inaccuracies in the personal data we hold about you by making a request to us in writing]; 29 [(ii) the right to request erasure], 30 [restriction], 31 [portability] and 32 [to object to the processing of your personal data]; 33 [(iii) the right to withdraw your consent, where we process your personal data on the basis of consent]; 34 [(iv) the right to lodge a complaint with a data protection authority.] 35 [If you would like to contact the data protection officer, please send an email to dpo@management.com.]
36 [We have implemented appropriate technical and organizational security measures designed to protect your personal data against accidental or unlawful destruction, loss, alteration, unauthorised disclosure, unauthorised access, in accordance with applicable law.] 37 [We will take steps to limit the way in which the processing of personal data is carried out by automated means, to a reasonable and proportionate level.] 38 [The Bank's intention does not include holding the personal data of minors who may have access to its website.] 39 [However, since the Bank cannot feasibly ensure/confirm this, all minors who do use the website and send their personal data to the Bank via the website are obliged and expected to have obtained consent from the persons exercising parental care or from their guardians.]

1 Controller.Identity
2 Controller Representative.Identity
3 Controller Representative.Contact.Address
4 Controller.Contact.Legal Address
5 Controller.Contact.Email
6 Controller.Contact.Phone Number
7 PD Origin.Direct
8 PD Origin.Indirect.Publicly
9 PD Category
10 PD Origin.Indirect.Third Party
11 PD Origin.Indirect.Cookie
12 PD Category.Special
13 PD Time Stored
14 Recipients
15 Processing Purposes
16 Transfer Outside Europe.Adequacy Decision.Country
17 Transfer Outside Europe.Safeguards.Binding Corporate Rules
18 Transfer Outside Europe.Safeguards.EU Model Clauses
19 Transfer Outside Europe.Specific Derogation.Unambiguous Consent
20 Legal Basis.Legal Obligation
21 Legal Basis.Legitimate Interest
22 Legal Basis.Consent
23 Legal Basis.Public Function
24 Legal Basis.Vital Interest
25 Legal Basis.Contract.To Enter Contract
26 PD Provision Obliged
27 Data Subject Right.Access
28 Data Subject Right.Rectification
29 Data Subject Right.Erasure
30 Data Subject Right.Restriction
31 Data Subject Right.Portability
32 Data Subject Right.Object
33 Data Subject Right.Withdraw Consent
34 Data Subject Right.Complaint.SA
35 DPO.Contact.Email
36 PD Security
37 Auto Decision Making
38, 39 Children

Fig. 1: Example of a fully annotated privacy policy.

software developer and might lead to developing a non-compliant system. Consequently, the organization could bear significant fines for violating data-protection rules.

More precisely, a privacy policy can be considered as a form of legally binding requirements specification which describes some of the properties and functionalities of a system-to-be. Therefore, completeness checking of privacy policies, and identifying their metadata as a primary step, can be seen as part of a broader solution to ensure legal compliance in information systems. In the software engineering (SE) literature, there have been attempts at mapping the text of a privacy policy to the implementation of a given software application, as a method for detecting GDPR violations [10], [11]. For instance, based on what we argued earlier, the privacy-related requirement about answering an individual's request pertaining to their personal data has to be mapped onto some function in the developed software.

Similarly, other metadata types identified in privacy policies can play a major role in software development. Examples include PD SECURITY, PD TIME STORED, DATA SUBJECT RIGHT.ERASURE, LEGAL BASIS.CONSENT, and DATA SUBJECT RIGHT.WITHDRAW CONSENT. In response to PD SECURITY, the controller has to implement appropriate protection mechanisms during software development (e.g., using encryption) to avoid penalty charges for information leakage as stated in GDPR. Further, a software system has to automatically delete collected personal data according to the time limit specified in the privacy policy (PD TIME STORED) or upon an individual's request (DATA SUBJECT RIGHT.ERASURE). When the consent of an individual is required for processing personal data (LEGAL BASIS.CONSENT), a software system has to implement a clear request procedure for consent where the individual takes an action to provide consent, e.g., by checking an "I agree" checkbox. As stipulated by GDPR, the system would also have to provide individuals with the possibility to withdraw this consent (DATA SUBJECT RIGHT.WITHDRAW CONSENT). The above examples show the benefits of completeness checking in different scenarios. Since checking completeness manually is time-consuming and effort-intensive, computer-assisted support for this task is advantageous.

A naive completeness-checking solution is to automatically find certain metadata types in a privacy policy through searching for keywords that are commonly used to express these metadata types. Relying merely on keyword search is problematic due to several reasons. First, there are overlapping keywords among multiple metadata types. For example, the keyword "protect" can indicate three metadata types related to security, data protection office, and safeguards for transferring personal data outside of Europe. Second, some metadata types cannot be captured via keywords. For instance, the metadata type RECIPIENTS (i.e., the parties with which individual personal data is shared) is usually expressed in the privacy policy as a list of diverse organizations (number **14** in Fig. 1). Since each privacy policy can have a different list of RECIPIENTS, using keyword search is infeasible for identifying this metadata type. To illustrate, let us suppose that "third parties" is used as a keyword for identifying RECIPIENTS. Note that the same keyword can also be used to identify PD ORIGIN.INDIRECT.THIRD-PARTY. Searching for this keyword will result in missing all occurrences of RECIPIENTS that do not contain the keyword and falsely identifying some occurrences due to overlapping keywords. In addition to the limitations of keyword search, the problem of checking completeness raises several other challenges. A particular sentence can discuss one or more metadata types which can be described in a hierarchy based on the specializations introduced in GDPR. In other words, an automated solution should be able to predict multiple (hierarchical) labels (metadata types) for a given sentence in the privacy policy. Inter-dependent metadata types (e.g., CONSENT and WITHDRAW CONSENT – discussed earlier) do not always occur consecutively in the privacy policy. This means that successful completeness checking requires identifying all the related metadata types accurately.

1.2 Research Questions

The paper investigates the following six research questions (RQs):

RQ1: What are the metadata types required for checking the completeness of a privacy policy according to GDPR? We answer RQ1 by building a conceptual model that specifies GDPR's information requirements for privacy policies. Our conceptual model, comprised of 56 metadata types, was developed in close collaboration with subject-matter experts. The concepts in this model are described in a glossary and are further traceable to the articles of GDPR.

RQ2: What are the criteria for checking whether a privacy policy is complete according to GDPR? Drawing on our conceptual model, to answer RQ2, we define a set of 23 criteria specifying what in a privacy policy should be checked for completeness against GDPR. Violating any of these criteria might lead to an incomplete privacy policy.

RQ3: How can privacy policies be automatically checked for completeness against GDPR? To answer RQ3, we use a combination of NLP and ML methods based on word embeddings and semantic similarity to develop an AI-based approach. Our approach identifies the different metadata types (from our conceptual model in RQ1) that are present in a privacy policy (*metadata identification*), and then checks these metadata against the completeness criteria (derived in RQ2) using automated conditional expressions (*completeness checking*).

RQ4: How accurate is our proposed approach in identifying GDPR-relevant metadata in privacy policies? RQ4 examines the accuracy of our metadata identification approach. As we discuss in Sec. 6, we achieve an average precision of 92.1% and average recall of 95.3% on an evaluation set made up of 48 unseen privacy policies.

RQ5: How accurate is our approach in checking the completeness of privacy policies? In RQ5, we investigate the accuracy of our automated approach in checking the completeness of privacy policies according to the provisions of GDPR. Over the evaluation set, our approach successfully finds 300 out of 334 violations of the completeness criteria, while raising false alarms (false positives) in 23 cases. Our approach has thus a precision of 92.9% and a recall of 89.8%.

RQ6: Is our approach worthwhile compared to a simpler solution? In RQ6, we compare our AI-based approach to a baseline that uses only keyword search. Compared to this baseline and over our evaluation set, using AI-technologies improves the metadata identification by an

average precision of 26.9% and average recall of 5.2%. Our approach significantly improves the overall completeness checking of privacy policies by an average precision of 24.5% and average recall of 38%.

The research presented in this paper is an extension of a previous conference paper [12] published at the 28th IEEE International Requirements Engineering conference (RE'20). The current paper provides a much more extensive empirical investigation in terms of the research questions, privacy policies used for evaluation, and metadata types covered by these policies. In particular, (1) we provide, through a concrete and detailed example, different scenarios where automated completeness checking turned out to be useful to a diverse group of people including lawyers and software engineers; (2) we include two more research questions: RQ2 for addressing the qualitative methods leading to the derivation of completeness criteria and RQ6 for comparing our approach to a simple, intuitive baseline; (3) we apply our AI-based approach for identifying all the 56 metadata types in a given privacy policy; to put this into perspective, our earlier conference paper dealt with only 20 metadata types; and (4) we improve our validation method to empirically evaluate our approach on 48 unseen privacy policies (≈20% of the entire dataset), instead of only 24 policies as was the case in our earlier conference paper.

1.3 Contributions

This paper makes the following four contributions:

(1) We develop a conceptual model to characterize the content of privacy policies, as stated in the provisions of GDPR. This conceptual model provides an abstract and yet precise set of metadata types that one can expect to find in privacy policies according to GDPR.

(2) We create a set of completeness criteria that describe when a privacy policy is considered complete according to GDPR. For creating these criteria (and also the conceptual model in (1)), we use systematic qualitative methods, as will be further explained in the paper.

(3) We develop an automated completeness checking tool using AI technologies. Specifically, we devise an approach based on Natural Language Processing (NLP) and Machine Learning (ML) for automatically identifying the content of a given privacy policy. To do so, we rely on the metadata types in the conceptual model developed in (1) as classification types. Given the identified metadata, we subsequently use the completeness criteria created in (2) to automatically check whether a given policy meets the information requirements envisaged by GDPR.

(4) We empirically evaluate our approach using a dataset of 234 privacy policies. These policies collectively contain 19847 sentences manually assigned (when applicable) to one or more of the metadata types from our conceptual model. The large majority (87%) of these assignments have been made by independent, third-party annotators (non-authors). We use ≈80% of our dataset for developing our proposed solution and the remaining ≈20% for evaluation. On our evaluation set, our AI-based approach yields an average precision of 92.1% and average recall of 95.3% in automatically identifying the different metadata types. Our completeness checking yields an average precision of 92.9% and an average recall of 89.8%. Compared to a baseline that uses keyword search, our approach leads to an overall average improvement of 24.5% in precision and 38% in recall when checking the completeness of privacy policies.

1.4 Structure

Sec. 2 provides background information. Sec. 3 presents the qualitative study we conducted for building our privacy-policy conceptual model. Sec. 4 describes the methods we used to create a set of criteria for checking the completeness of privacy policies according to GDPR. Sec. 5 explains our proposed AI-based approach for checking the completeness of a given privacy policy. Sec. 6 discusses the empirical evaluation of our approach. Sec. 7 discusses threats to validity. Sec. 8 compares our contributions with related work. Sec. 9 describes how we envision our overall approach being replicated for other regulations and document types. Sec. 10 concludes the paper.

2 Background

In this section, we first briefly introduce GDPR. We then summarize the necessary background related to our technical approach.

2.1 GDPR

GDPR [1] is a complex regulation comprised of 173 recitals and 99 articles divided into 11 chapters. GDPR applies primarily to organizations within Europe. However, the regulation may also apply to organizations outside Europe, e.g., when these organizations offer goods or services to, or monitor individuals in Europe. If an organization is subject to GDPR, it has to identify itself as either a data controller or data processor. A controller determines the purpose and means of processing, whereas a processor acts on the instructions of the controller. The responsibilities of a given organization under GDPR vary depending on whether it is a processor or a controller. Processors notably have to: (1) implement adequate technical and organizational measures to keep personal data safe and secure, and, in cases of data breaches, notify the controllers; (2) appoint a statutory data protection officer (if needed) and conduct a formal impact assessment for certain types of high-risk processing; (3) keep records about their data processing; and (4) comply to GDPR restrictions when transferring personal data outside Europe. In comparison to processors, controllers are subject to more provisions. In particular, in addition to having to meet the obligations mentioned above, controllers have to: (1) adhere to six core personal data processing principles, namely, fair and lawful processing, purpose limitation, data minimization, data accuracy, storage limitation, and data security; (2) keep identifiable individuals informed about how their personal data will be used; and (3) preserve the individual rights envisaged by GDPR, e.g., the right to be forgotten and the right to lodge a complaint. GDPR includes some specific provisions in relation to privacy policies. Privacy policies play a major role in software development. For example, they refer to how a controller (i.e., the software) should ensure data security, how long the

collected data should be stored on the controller database, what the software needs to provide to the user if some specific user rights are in place (e.g., withdraw consent and data erasure), etc. We elaborate the GDPR provisions for privacy policies in Sec. 3.

2.2 Natural Language Processing

Natural language processing (NLP) is a sub-field of AI, which is used for automated processing of natural-language data. Examples of NLP applications include machine translation and information extraction [13], [14]. In our work, we apply the NLP pipeline depicted in Fig. 2. The pipeline combines six consecutive NLP modules divided in three categories.

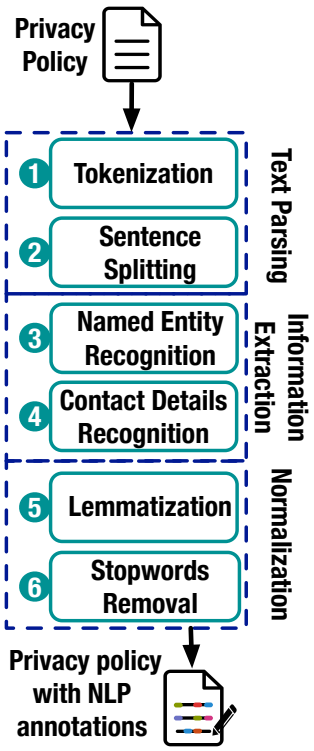


Fig. 2: NLP Pipeline.

The first category is aimed at parsing the text of a given privacy policy. This category includes *Tokenization* for separating out the words and punctuation marks from the running text and *Sentence Splitting* for decomposing the text into coherent sentences based on sentence boundary indicators such as periods, question and exclamation marks [14], [15].

The second category in the pipeline is concerned with extracting information from the text. The first step uses the *Named Entity Recognition (NER)*, which is the task of marking the mentions of named entities in a given text with their types [16], e.g., a country name like "Luxembourg" will be annotated with the type *location*. The entity types, in our work, are limited to *location* and *organization* since these two are expected to appear in a privacy policy. In addition to the NER module, we use regular expressions [17] to recognize the contact details that are mentioned in the input privacy policy, namely email and postal addresses, telephone numbers and websites. For example, the email address "info@hikari.jp" will be recognized as *email*.

The last category involves normalizing the text. In particular, different words in a text can be mapped to a single root form using the *Lemmatization* module, e.g., the words "deletion", "deleted" and "delete" will be lemmatized to the word "delete". Finally, we use the *Stopwords Removal* to remove stopwords, i.e., very frequent words such as prepositions (e.g., "in") and articles (e.g., "a" and "the"). Applying the NLP pipeline above results in adding various annotations to the input privacy policy.

2.3 Machine Learning

Machine learning (ML) is another sub-field of AI which describes the automated learning methods used for finding meaningful patterns in data [18]–[20]. Supervised ML assumes that training examples (input) are provided with their labels (output). Using these training examples, the machine then learns to predict the output of unseen examples. We will refer to the input and its associated output value as a *classification instance*. Text classification (also known as text categorization) is supervised learning for categorizing the text into a set of predefined groups [19], e.g., classifying the text of an email into *spam* and *not spam*.

In this work, we focus on *multiclass multilabel classification*. Multiclass classification is to classify the input examples into three or more predefined classes. A classical example in the ML literature is classifying an iris flower, given its sepal length and width and petal length and width, into one of the three possible types *setosa*, *versicolor*, or *virginica* [20]. Multilabel classification means that the same input example may belong to multiple classes, e.g., classifying movies into one or more genres based on the plot summary, where a movie can belong to *comedy* and *action* at the same time. The multi-label classification problem is often simplified into multiple binary classification problems [19]. A binary classification is a specific case of the multi-class classification with only two target classes. For example, a movie can be classified into genres using multiple learning algorithms, such that each learner predicts whether the movie is from a specific genre (e.g., *comedy*) or not from that genre (e.g., *not comedy*), and so on for the other genres.

2.4 Vector-space Representation of Text

Vectorization is a prerequisite step to text classification where the text has to be transformed into a set of feature vectors for describing the text under the different pre-defined classes [19], [21]. Each classification instance is represented by a feature vector. These features can be either manually crafted (e.g., the presence of first person pronouns like "we") or automatically generated using the words in the text. There are several models to perform vectorization, e.g., BoW (bags of words), TF/IDF (term frequency/inverse document frequency) and word embeddings. In our work, we use word embeddings. In particular, we utilize pre-trained word vectors from the GloVe model [22].

Word embeddings are representations of words as dense numerical vectors that capture the syntactic and semantic regularities [22]–[24]. Deriving these representations is based on the distributional hypothesis of Harris [25] which states that semantically-related words appear in similar contexts. Therefore, vectors that are close to each other in the vector space should represent words that are similar, e.g., the vector representing the word "frog" should be close to vectors of similar words such as "toad" and "lizard" [22]. Regularities are observed in the linear relations between word pairs. For example, if a word w is represented by the vector $\vec{w}$, then we observe the plural relation: $\vec{cat} - \vec{cats} \approx \vec{apple} - \vec{apples}$.

Out of the available methods for learning word embeddings, we use the pre-trained GloVe embeddings [22]. Pre-trained models are used to improve a range of NLP

tasks [26]–[33]. In modern NLP, pre-trained word embeddings perform better than those learned from scratch [34]. Compared to newer technologies for generating text representations like ELMo [35], OpenAI GPT [36] and BERT [37], GloVe provides context-independent word embeddings (i.e., one-to-one mapping between the words and their vectors) that can be directly used off-the-shelf. Despite being powerful, context-aware representations generated by (for example) BERT come with the cost of an extra step for training, or fine-tuning. Moreover, the GloVe pre-trained model achieves good results on NLP downstream tasks [38]. Compared to word2vec [23] and fasttext [39], which also provide pre-trained word vectors, GloVe learns words representations using both local and global context to better capture the semantics of words [40]. Global context is used to enrich the words representations by considering the co-occurrence counts of the words in a large corpus.

The GloVe pre-trained model, used in our work, uses 100-dimensional vectors generated by training on extensive text corpora from Wikipedia and the web. To illustrate, consider the text segment "Hikari Bank Privacy Policy" in Fig. 1. Using pre-trained word embeddings, each word is represented as a 100-dimensional vector, e.g., "hikari" is represented as $[0.42192, 0.41032, 0.23888, \ldots]_{100}$. Computing the vector representation of this text segment can then be performed by combining the word embeddings in the segment through different mathematical operations including summation and (simple or weighted) averaging [41]–[43]. In our approach, we use simple averaging because it proved to be effective in text-similarity-related tasks.

3 A Conceptual Model of Privacy-Policy Metadata (RQ1)

In this section, we present the following artifacts to answer RQ1: (1) a conceptual model specifying, in a comprehensive manner, the metadata types pertinent to GDPR privacy policies; and (2) a glossary defining all necessary terms to better understand the conceptual model with traceability to GDPR articles. The conceptual model (artifact 1) is shown in Fig. 3 and an excerpt of the glossary (artifact 2) is presented in Table 1. The complete glossary is provided as an online annex [44]. The above artifacts were built using an iterative and incremental method following three main steps (see Fig. 4): (1) reading the articles of GDPR that address privacy policies, (2) creating and refining the artifacts introduced above, and (3) validating these artifacts with legal experts. Building the artifacts took four iterations with each iteration requiring, on average, one month. We had several face-to-face and off-line validation sessions with legal experts. The sessions, which lasted between two and three hours each, collectively added up to approximately 30 hours.

We conducted our validation sessions with three legal experts, namely (a) a senior lawyer with more than 30 years of experience in European and international laws; (b) a mid-career lawyer with more than 10 years of experience in law with a focus on the data protection and financial domains; and (c) an IT professional with more than 10 years of experience in the legal domain. Each validation session was attended by at least two legal experts. The discussions continued until the experts in attendance agreed that the model correctly reflected their interpretation of GDPR. We observed that the differing viewpoints and thus the deliberations between the legal experts centered primarily around the specializations in the conceptual model (e.g., the sub-metadata types of Legal Basis.Contract) and about how different metadata types should be inter-related (e.g., how PD Origin.Indirect is related to PD Category.Type).

Initially, as suggested by our collaborating legal experts from Linklaters, we analyzed Art(icles) 13 and 14 of GDPR, i.e., the main GDPR articles targeting privacy policies. From these two articles, we extracted important concepts to create

TABLE 1: Glossary Excerpt.

Metadata (**Reference**[1])	Description
Controller (**Art. 13/14(f)**)	A natural or legal person, public authority, agency or any other body which, alone or jointly with others, determines the purposes and means of the processing of personal data where the purposes and means of such processing are determined by national or EU laws or regulations, the controller or the specific criteria for its nomination may be provided by national or EU law.
Identity (**Art. 13/14(f)**)	The legal name of the company/organization.
Contact (**Art. 13/14(f)**)	The method(s) with which the company/organization can be contacted.
Controller Representative (**Art. 13/14(f)**)	A natural or legal person established in the Union who is designated by the controller.
Data Protection Officer (DPO) (**Art. 13/14(f)**)	The one who is responsible for overseeing data protection strategy and implementation to ensure compliance with GDPR requirements.
Processing (**Art. 13/14(f)**)	Any operation performed on personal data, whether or not by automated means, including collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.
Personal Data (PD) (**Art.5(f)**)	Any information related to an identified or identifiable natural person.
Provision (**Art. 14(f)**)	The action of providing something (i.e., personal data) for use (i.e., to be processed).
PD Origin (**Art. 14.2(f)**)	From which source the personal data originates (i.e., direct or indirect), and if applicable, whether it came from a publicly and/or third-party and/or cookie sources.
Indirect (**Art. 14**)	When the personal data are not obtained from the data subject.
Third Party (**Art. 14**)	When the personal data are obtained from organisations external to the data controller.
Publicly (**Art. 14**)	When the personal data are obtained from public sources (i.e., from a public website).
Profiling (**Art. 4(f)**)	To analyze or predict aspects concerning a natural person's performance at work.

[1] GDPR-related articles

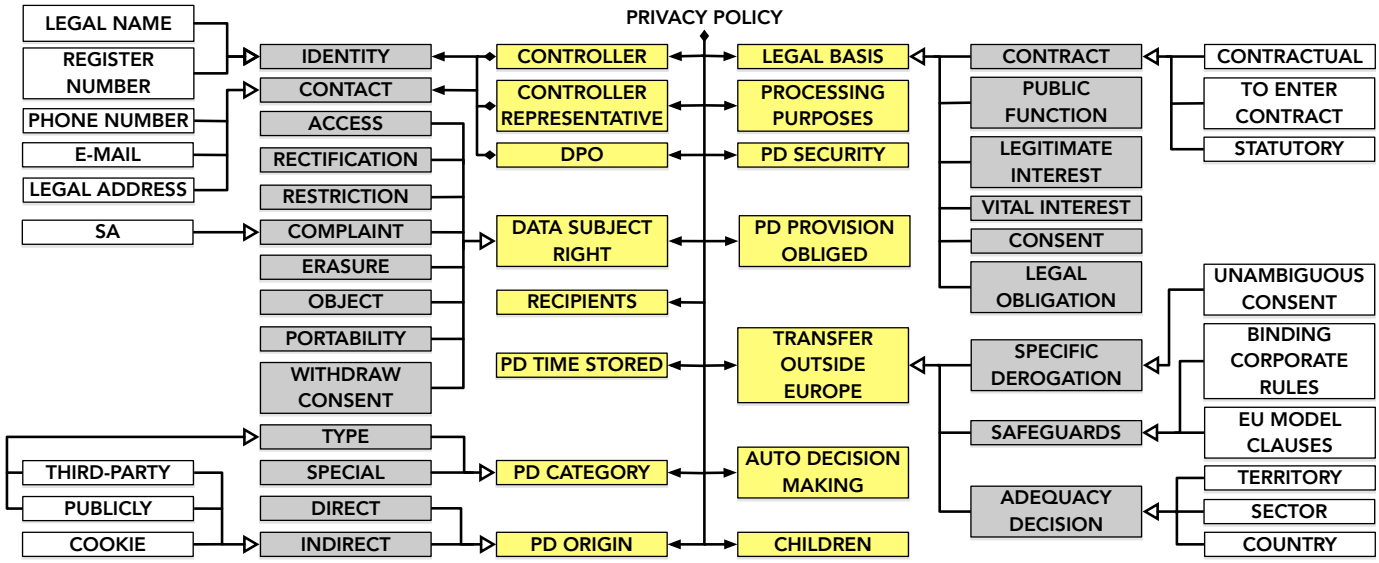


Fig. 3: Conceptual Model of Privacy-Policy Metadata.

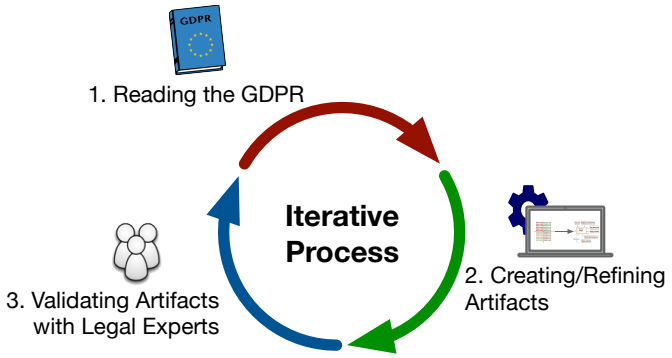


Fig. 4: Iterative Process.

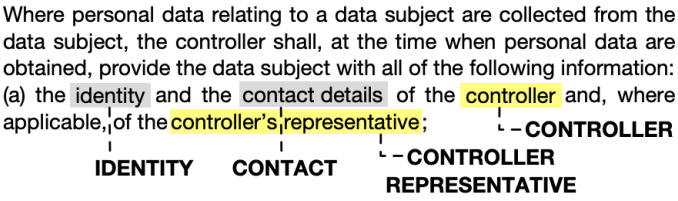


Fig. 5: Example of Coding in the Context of GDPR.

the metadata and the dependencies between them. Art. 13 focuses on personal data collected directly from a data subject (e.g., filling an online form or an interview), whereas Art. 14 focuses on personal data obtained indirectly from a data subject (e.g., obtained from a public website or public list). We observe that Art. 13.2(e) *(whether the provision of personal data is a statutory or contractual requirement, or a requirement necessary to enter into a contract, as well as whether the data subject is obliged to provide the personal data and of the possible consequences of failure to provide such data)* is related to the direct collection of personal data, while Art. 14.2(f) *(from which source the personal data originate, and if applicable, whether it came from publicly accessible sources)* deals with indirect collection. These observations were taken into consideration while building the two artifacts discussed above. Starting from Art. 13 and 14, as per the recommendation of legal experts, we also examined Art. 6, 9, 21, 37, 46, 47, 49, 55, and 56 by doing a snowball sampling from the cross-references in Art 13 and 14.

Fig. 5 illustrates an excerpt of Art. 13 from which we have inferred the hierarchical representation of four metadata types: CONTROLLER, CONTROLLER REPRESENTATIVE, and their descendants IDENTITY and CONTACT. These metadata types refer to four distinct concepts: (1) the identity of the data controller (CONTROLLER.IDENTITY), (2) the contact details of the data controller (CONTROLLER.CONTACT), (3) the identity of the data controller's representative (CONTROLLER REPRESENTATIVE.IDENTITY), and (4) the contact details of the data controller's representative (CONTROLLER REPRESENTATIVE.CONTACT). The metadata types IDENTITY and CONTACT were ultimately specialized with the inclusion of other sub-metadata types. The former, with LEGAL NAME and REGISTER NUMBER metadata types and the latter with EMAIL, LEGAL ADDRESS and PHONE NUMBER. Our conceptual model (depicted in Fig. 3) is organized into three hierarchical levels: **level-1**, shaded yellow, **level-2**, shaded grey, and **level-3**, shaded white. The colors were introduced to make the model more readable to annotators and legal experts. As presented in Fig. 6, the methodology we used for identifying the metadata types from GDPR and building the conceptual model involved three types of coding: *in-vivo coding*, *hypothesis coding* and *subcoding* [45].

1. In-vivo coding: we use this type of coding to identify the core concepts in GDPR and create an initial set of codes. In-vivo coding emphasizes the actual words in the text – in our case the text of GDPR – in order to create codes. The in-vivo approach allowed us to derive the names of the metadata types directly from the text of GDPR (i.e., the meta documents). Those metadata types are then used to characterize GDPR-related text found in privacy policies. A metadata type, representing a code, is a short phrase that symbolically assigns a summative, salient, essence-capturing, and/or evocative attribute to a particular text in a given privacy policy [45]. For example, the metadata type CONTROLLER in Fig. 5 refers to the text in a given privacy policy that discusses a natural or legal person,

public authority, agency or any other body which, alone or jointly with others, determines the purposes and means of personal data processing (see Art. 13.1(a) of GDPR).

2. Hypothesis coding: this type of coding refers to the application of a predetermined set of codes to qualitative data in order to assess researcher-generated hypotheses. The codes are developed from a prediction – in our case, the initial set of codes identified from GDPR with in-vivo coding – about what one would find in the actual data – in our case, privacy policies – before the data was collected and analyzed. Usually, the application of this coding methodology can range from simple frequency counts to more complex multivariate analyses. In our context, we are interested in the presence or absence of metadata types in a given privacy policy in order to check its completeness against GDPR. In particular, with the help of legal experts, we manually applied this initial set of codes (obtained with in-vivo coding) via hypothesis coding over 30 privacy policies (a subset of our training set) in order to ensure that our in-vivo codes are sufficient and at the right level of abstraction.

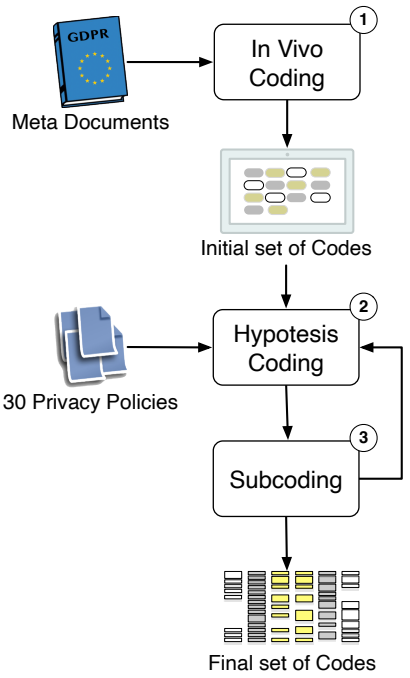


Fig. 6: Coding methodology.

While applying hypothesis coding to the privacy policies, for each metadata type, we collected the keywords that made us decide to associate a given sentence with a specific metadata. For example, the combination of keywords *"right to access"* was extracted from sentence number **27** of Fig. 1 and included in the list of of keywords associated with DATA SUBJECT RIGHTS.ACCESS. At the end of hypothesis coding, we obtained a list of keywords for each metadata type as shown in Fig. 3.

3. Subcoding: in addition to hypothesis coding, we also use subcoding, which refers to sub-codes as a second-order tag assigned after a primary code, in order to enrich our metadata types in terms of specificity. For example, in Fig. 3, the metadata type PD ORIGIN (in yellow) is specialized into two sub-metadata types: DIRECT and INDIRECT (in gray). Then, INDIRECT is further specialized into: THIRD-PARTY, PUBLICLY and COOKIE (in white). The use of subcoding ultimately contributed to the final set of codes represented by the conceptual model of Fig. 3.

Based on our interpretation and understanding of GDPR articles, we created an initial version of the metadata conceptual model along with their definitions. We kept track of GDPR articles to ensure traceability in our glossary (artifact 2). Table 1 presents an excerpt of our glossary.

These (interim) artifacts were then presented to legal experts for feedback. In addition to pointing out issues and omissions, the experts were encouraged to bring to our attention any GDPR article or external documentation/information needed to be considered in the context of privacy policies.The feedback obtained from legal experts was, by and large, concerned with information that was not explicitly included in GDPR (e.g., the European Working Party [46]). For example, Art. 13.1(f) states that "*the controller intends to transfer personal data to a third country or international organization and the existence or absence of an adequacy decision by the Commission, or [...] appropriate or suitable safeguards [...]*". This article is addressed in Fig. 3 by the metadata type TRANSFER OUTSIDE EUROPE.ADEQUACY DECISION. In response to the legal experts' feedback and by following the external source[1] recommended by them, we created the sub-metadata types of ADEQUACY DECISION that are not discussed in GDPR. In particular, three sub-metadata types were added to the conceptual model in Fig. 3, namely, TERRITORY, SECTOR, and COUNTRY. These metadata types refer to the adequacy decisions between the EU and a territory (e.g., Andorra, the Bailiwick of Jersey, etc.), specific sectors (e.g., the commercial organizations from Canada, Argentina, etc.), and a country (i.e., Japan, New Zealand, etc.), respectively. In the same manner, we used another external source[2] to create the level-3 sub-metadata type EU MODEL CLAUSES.

Once the conceptual model converged to a stable state, we put together a general report including the conceptual model and the glossary table. The conceptual modeling step terminated when the general report was approved by legal experts. The final version of the conceptual model, with a total of 56 metadata types (see Fig. 3), along with a complete glossary table of 60 entries, are provided in an online annex [44].

4 COMPLETENESS CHECKING CRITERIA FOR PRIVACY POLICIES (RQ2)

In this section, we answer RQ2 by presenting the criteria we use to check the completeness of privacy policies according to GDPR. In particular, we discuss our method for creating a set of 23 criteria for checking the completeness of privacy policies by analyzing GDPR articles. In order to identify these criteria, we used an iterative three-step method similar to the one we used to create the other two artifacts mentioned in Sec. 3 (see Fig.4). We obtained the final set of completeness criteria in six iterations, with each iteration requiring, on average, 15 days. During this process, we combined bi-weekly face-to-face validation sessions and off-line interactions with legal experts. The face-to-face sessions, which lasted between 2 to 3 hours each, collectively added up to approximately 15 hours, plus an additional five hours for off-line interactions.

1. EU Adequacy Decisions – https://bit.ly/38ciwPU (January 2021)
2. EU Standard Contractual Clauses – https://bit.ly/3nd6JFt (January 2021)

4.1 Transforming GDPR Articles into Criteria

The complete set of criteria discussed in this section uses the metadata types identified in the conceptual model of Fig. 3. We note that some metadata types are *inter-dependent*, meaning that the presence of a metadata type requires the presence of another metadata type. For example, if a privacy policy requires individuals to provide consent for collecting their personal data, then the policy shall also allow individuals to withdraw their consent, i.e., LEGAL BASIS.CONSENT and DATA SUBJECT RIGHT.WITHDRAW CONSENT are inter-dependent. Most of the criteria were extracted from the same GDPR articles from which the metadata types were also identified (see the external online annex [44] for criteria traceability to the GDPR articles). Based on our interpretation and understanding of these GDPR articles, we identified an initial set of criteria that we formulated as pseudo-code. Each pseudo-code statement is composed of two main parts: 1) a *precondition* (if any) about the identification of one or more metadata types in a privacy policy or other GDPR-related conditions proposed by the legal experts, and 2) a *postcondition* asserting the identification of one or more metadata types (different from the one(s) in the *precondition*) in a privacy policy. We use the following template [*precondition*], <*postcondition*> and show below examples of criteria written in pseudo-code; these are derived from the excerpt of Art. 13.1(a) shown in Fig. 5:

- C1 [ ], <CONTROLLER.IDENTITY.{REGISTER NUMBER *or* LEGAL NAME} must be identified>.
- C2 [ ], <CONTACT.{EMAIL or PHONE or LEGAL ADDRESS} must be identified>.
- C3 [*if* CONTROLLER is located outside of Europe], <*then* CONTROLLER REPRESENTATIVE.IDENTITY.{REGISTER NUMBER or LEGAL NAME} must be identified>.
- C4 [*if* CONTROLLER is located outside of Europe], <*then* CONTROLLER REPRESENTATIVE.CONTACT.{EMAIL or PHONE or LEGAL ADDRESS} must be identified>.

In this step, we transform the text of the relevant GDPR provisions into completeness criteria. For example, considering Fig. 5, the word *shall* is translated into a mandatory requirement for including the CONTROLLER.IDENTITY (C1) and CONTROLLER.CONTACT details (C2). On the other hand, the combination of the words *where* and *applicable* suggests that a given criterion should be enforced only if certain precondition(s) are met: the CONTROLLER REPRESENTATIVE.IDENTITY (C3) and CONTROLLER REPRESENTATIVE.CONTACT (C4) need to be checked only if the CONTROLLER is located outside of Europe.

While defining the criteria from the GDPR articles, we realized that some of them should not always be checked. Articles 13.1(a,e,f), 13.2(e), 14.1(a,d,e,f), and 14.2(f) are GDPR articles that apply to privacy policies only in specific situations. After reviewing these articles with legal experts, they asked us to create a questionnaire that would help them specify, under various situations, the exact content of a given privacy policy for completeness checking. The person who should ideally provide answers to the questionnaire should have expertise in the legal domain as well as extensive knowledge about the company for which the privacy policy analysis is being performed. For example, to determine whether C3 and C4 above should be checked, it is important to know beforehand from the questionnaire that the CONTROLLER is located outside Europe.

The questionnaire contains a set of critical questions whose answers depend on context and are often left tacit in privacy policies. Nevertheless, these answers carry important implications on what needs to be explicitly covered in privacy policies, and hence on completeness checking. The questionnaire includes the following six questions:

- **Q1** Who is the CONTROLLER in charge of data processing? *Write name*.
- **Q2** Do you plan to transfer the collected personal data outside Europe? *Yes/No*.
- **Q3** Will there be other recipients of the collected personal data besides you? *Yes/No*.
- **Q4** What is the core of your activities?
  ☐ The processing of personal data is carried out by a public authority or body (except for courts acting in their judicial capacity).
  ☐ The processing of operations which, by nature, scope and/or purposes, require regular and systematic monitoring of data subjects on a large scale.
  ☐ The processing, on a large scale, of personal data relating to sensitive categories (e.g., racial or ethnic origin, political opinions, or religious or philosophical beliefs) or to criminal convictions and offenses.
- **Q5** Where will the activities carried out by your organization take place?
  ○ *Inside Europe*
  ○ *Outside Europe* – if selected, then write the name of CONTROLLER REPRESENTATIVE: . . . . . .
- **Q6** How will the personal data of the data subject be collected? ○ DIRECT ○ INDIRECT ○ *Both*

Question **Q1** is not intended to trigger the checking of any criterion. This first question is used to facilitate the identification of the metadata type CONTROLLER.IDENTITY. The main objective of the remaining questions is to determine whether some context-related criteria should be checked. In particular, each of the other five questions (**Q2-Q6**) triggers the search for one or more metadata types. This leads to checking some specific criteria. A positive answer to question Q2 triggers the verification of criteria C10 – C14. If the answer to Q3 is yes, then criterion C19 is verified. Similarly, Q4 will trigger the verification of criterion C23, if any of the optional answers to this question is checked. The answer to question Q5 activates the verification of criteria C3 and C4, if the activities carried out by the CONTROLLER take place outside Europe. Finally, answering Q6 as "DIRECT" activates checking criterion C22; "INDIRECT" activates checking criteria C15 – C18; and "*Both*" requires checking all the above criteria (i.e., C15 – 18 and C22).

The remaining criteria (C1, C2, C5 – C9, C20 and C21) are always verified because they refer to metadata types that, according to GDPR, must be present in every privacy policy.

At the end of the first step, we created a table with all the information about the criteria set, including an identifier (ID) for each criterion (first column), preconditions (middle column) and postconditions (last column), as shown in Table 2. Since C1, C2, C5, C20 and C21 are not triggered by any preconditions, they always need to be checked. The rest of

TABLE 2: Completeness Criteria according to GDPR.

ID	Criteria	
	Precondition[1]	Postcondition[2]
C1	-	CONTROLLER.IDENTITY
C2	-	CONTROLLER.CONTACT.{LEGAL ADDRESS, EMAIL, *or* PHONE NUMBER}
C3	**A5** is a country *outside the EU*	CONTROLLER REPRESENTATIVE.IDENTITY
C4	**A5** is a country *outside the EU*	CONTROLLER REPRESENTATIVE .CONTACT.{LEGAL ADDRESS, EMAIL, *or* PHONE NUMBER}
C5	-	DATA SUBJECT RIGHT.{ACCESS, COMPLAINT, RECTIFICATION, *and* RESTRICTION}
C6	DATA SUBJECT RIGHT.COMPLAINT	DATA SUBJECT RIGHT.COMPLAINT.SA
C7	LEGAL BASIS.CONTRACT	DATA SUBJECT RIGHT.PORTABILITY
C8	LEGAL BASIS .{LEGITIMATE INTEREST *or* PUBLIC FUNCTION}	DATA SUBJECT RIGHT.OBJECT
C9	LEGAL BASIS.CONSENT	DATA SUBJECT RIGHT.{ERASURE, OBJECT, PORTABILITY, *and* WITHDRAW CONSENT}
C10	**A2** is *Yes*	TRANSFER OUTSIDE EUROPE
C11	TRANSFER OUTSIDE EUROPE	TRANSFER OUTSIDE EUROPE.{ADEQUACY DECISION, SAFEGUARDS, *or* SPECIFIC DEROGATION}
C12	TRANSFER OUTSIDE EUROPE .ADEQUACY DECISION	TRANSFER OUTSIDE EUROPE.ADEQUACY DECISION .{COUNTRY, SECTOR, *or* TERRITORY}
C13	TRANSFER OUTSIDE EUROPE .SAFEGUARDS	TRANSFER OUTSIDE EUROPE.SAFEGUARDS.{EU MODEL CLAUSES, *or* BINDING CORPORATE RULES}
C14	TRANSFER OUTSIDE EUROPE .SPECIFIC DEROGATION	TRANSFER OUTSIDE EUROPE.SPECIFIC DEROGATION.UNAMBIGUOUS CONSENT
C15	**A6** is INDIRECT or *Both*	PD ORIGIN.INDIRECT
C16	PD ORIGIN.INDIRECT	PD ORIGIN.INDIRECT.{THIRD PARTY, *or* PUBLICLY}
C17	**A6** is INDIRECT or *Both*	PD CATEGORY
C18	PD ORIGIN.INDIRECT .{THIRD PARTY, *or* PUBLICLY}	PD CATEGORY.TYPE
C19	**A3** is *Yes*	RECIPIENTS
C20	-	PD TIME STORED
C21	-	PROCESSING PURPOSES
C22	**A6** is DIRECT or *Both* **and** LEGAL BASIS .{CONTRACT.TO ENTER CONTRACT, *or* LEGAL OBLIGATION}	PD PROVISION OBLIGED
C23	At least one answer in **Q4** is selected	DPO.CONTACT.{LEGAL ADDRESS, EMAIL *or* PHONE NUMBER}

[1] Includes the answers to Q2 – Q6 (**A2 – A6**), or the metadata types that are present in a privacy policy.
[2] Metadata types that must / should be present.

the criteria are triggered by some precondition related to answers to questions Q2 – Q6 (referred to as A2 – A6) from the questionnaire or the presence of some metadata in the privacy policy.

Table 2 presents criteria that *should* be satisfied according to GDPR (ID highlighted in orange) and may lead to a warning, and other criteria that *must* always be satisfied (ID highlighted in red) and may lead to a violation. We further discuss the difference between warnings and violations in the next subsection.

4.2 Evaluating the Criteria with Legal Experts

To facilitate the validation of the criteria presented in Table 2 with legal experts, we decided to capture them as activity diagrams, following the observation by Soltana et al. [47] that legal experts can understand activity diagrams with relative ease given some basic training. With the help of legal experts, we created a final set of 23 criteria to capture the mechanisms necessary to check the completeness of privacy policies according to GDPR. Among the 16 criteria shown in Fig. 7, C3, C4, C15, C17, C19, and C23 depend on the answers to the questionnaire. Fig. 7 and Fig. 8 show the 23 criteria to check the completeness of a privacy policy with respect to the metadata types of the conceptual model presented earlier. Fig. 7 contains every possible violation in privacy policies and Fig. 8 all the possible warnings.

The completeness criteria in Fig. 7 and Fig. 8 use in general three shapes to represent different types of actions

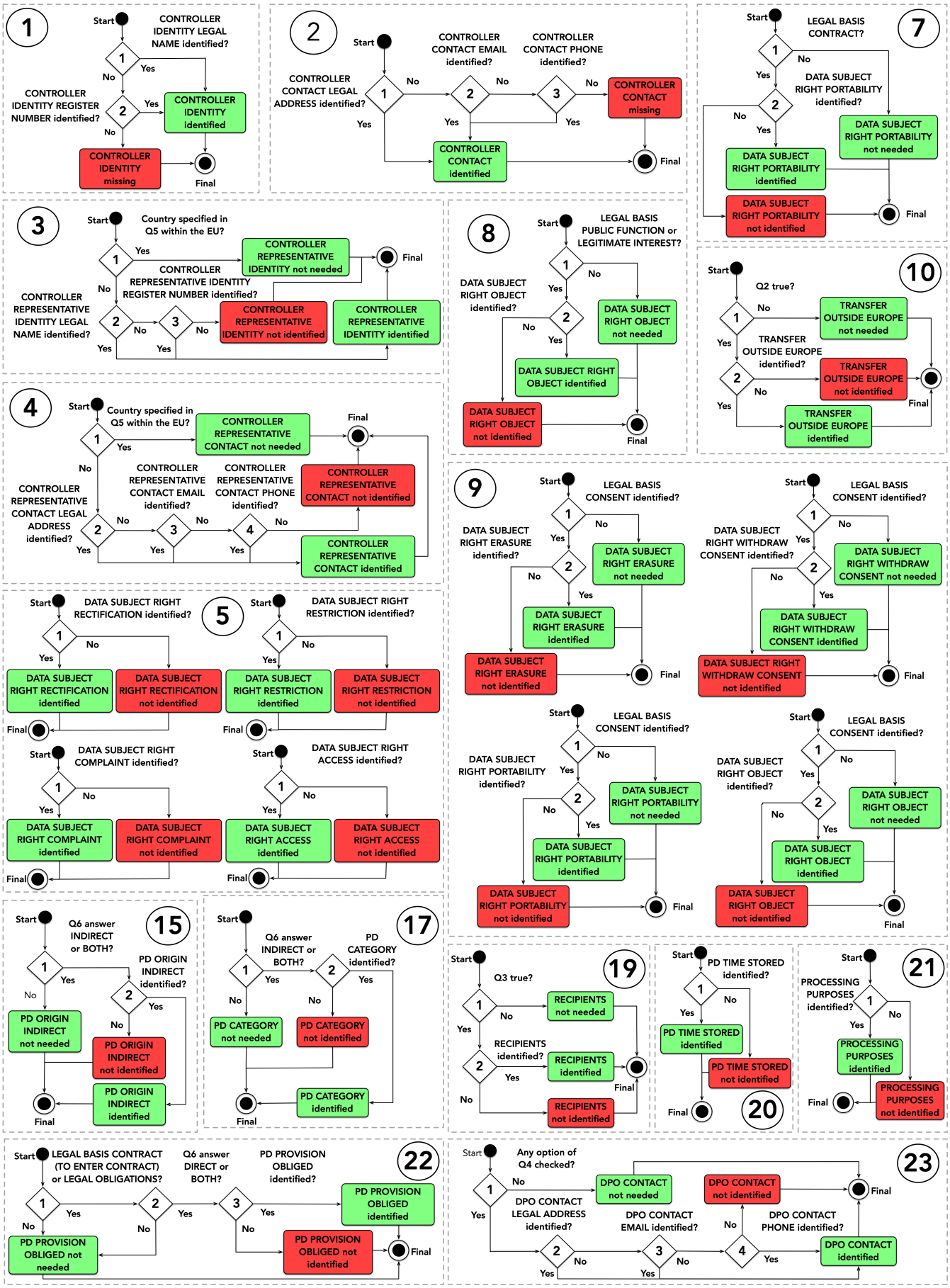


Fig. 7: GDPR Completeness Criteria Represented as Activity Diagrams (Violations).

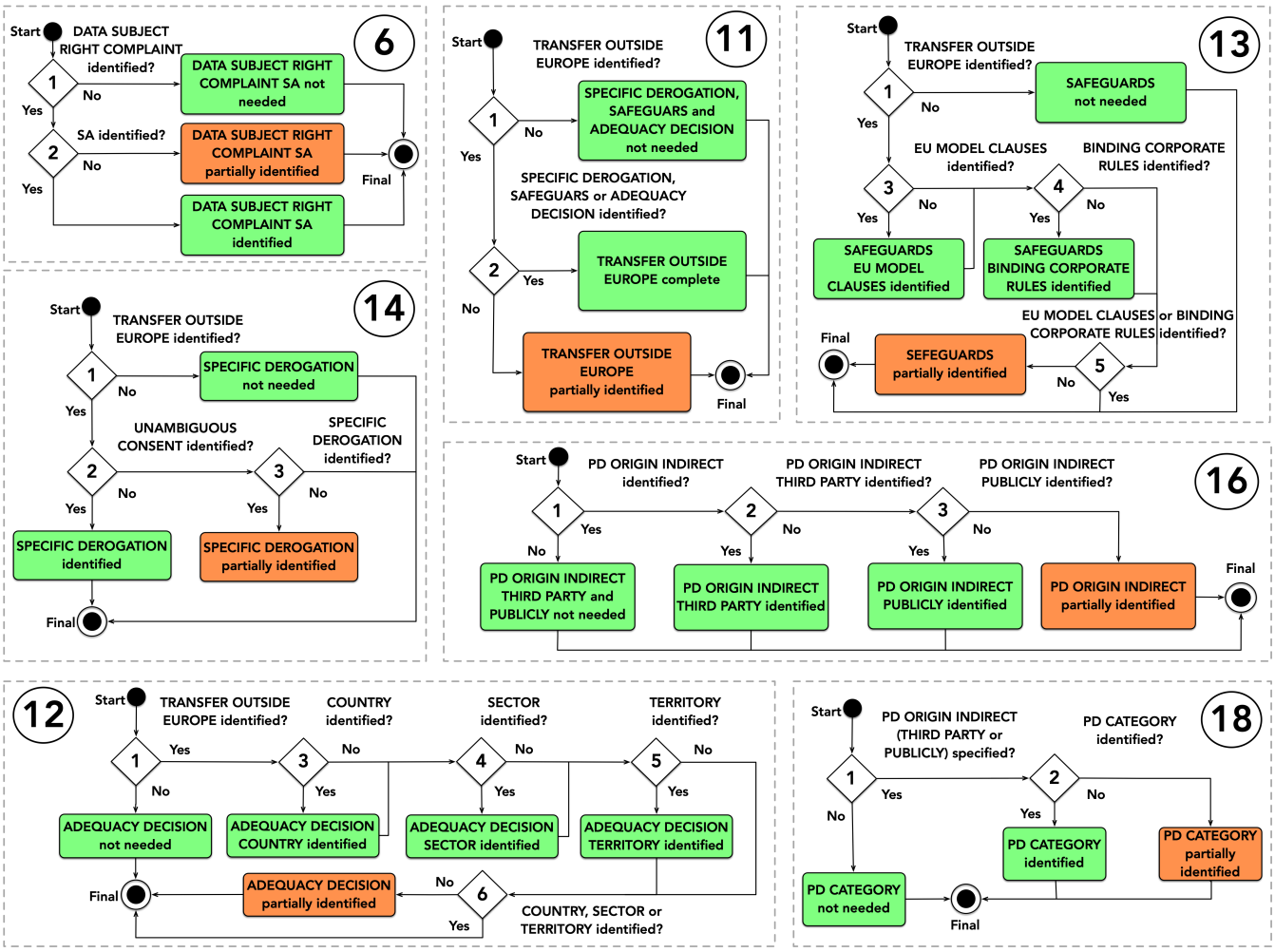


Fig. 8: GDPR Completeness Criteria Represented as Activity Diagrams (Warnings).

or steps in a process: (1) a circle represents the start and endpoint, (2) a diamond indicates a decision, and (3) a rectangle stands for an action representing that (3.1) a metadata type was correctly identified or not needed in a privacy policy (in green), (3.2) a mandatory metadata type was entirely missing in a privacy policy (referred to as *violation* – highlighted in red), and (3.3) a metadata type was only partially identified or, in other words, a metadata type was identified but some related information is missing (referred to as *warning* – highlighted in orange). An *incompleteness issue* is raised when a criterion returns a violation or warning. A violation corresponds to a direct breach of GDPR, whereas a warning leads to further assessment by the legal expert to finally decide whether there is a breach of GDPR.

Below, we illustrate two criteria, C15 and C16, derived from Art. 14.2(f) of the GDPR (see Fig. 7 and Fig. 8). These criteria check the completeness of a privacy policy with respect to the metadata type PD ORIGIN.INDIRECT. C15 is meant for identifying a violation:

- **(1)** If the answer to **Q6** is INDIRECT or *Both* (recall the questionnaire presented in Section 4.1), then go to **(2)**; otherwise PD ORIGIN.INDIRECT is *not needed*.
- **(2)** If the indirect origin of the personal data is mentioned, then PD ORIGIN.INDIRECT is *identified*; otherwise PD ORIGIN.INDIRECT is *missing* – **Violation**.

Criterion C16 is meant for identifying a warning:

- **(1)** If PD ORIGIN.INDIRECT is identified in C15, then go to **(2)**; otherwise PD ORIGIN.INDIRECT is *not needed*.
- **(2)** If the indirect origin of personal data is from a third-party, then PD ORIGIN.INDIRECT.THIRD-PARTY is *identified*; otherwise go to **(3)**.
- **(3)** If the indirect origin of personal data is from public sources, then PD ORIGIN.INDIRECT.PUBLICLY is *identified*; otherwise PD ORIGIN.INDIRECT is *partially identified* – **Warning**.

Note that C16 in Fig. 8 does not refer to COOKIE although COOKIE is a subtype of PD ORIGIN.INDIRECT in the conceptual model of Fig. 3. The above-shown criterion strictly follows GDPR, which does not regulate cookies. However, our collaborating legal experts suggested the inclusion of COOKIE in our conceptual model since cookies are often mentioned in privacy policies and they may become relevant to GDPR in the future.

5 AI-BASED APPROACH FOR COMPLETENESS CHECKING (RQ3)

In this section, we address RQ3 and present our AI-based approach for **Comp**liance checking of privacy policies using

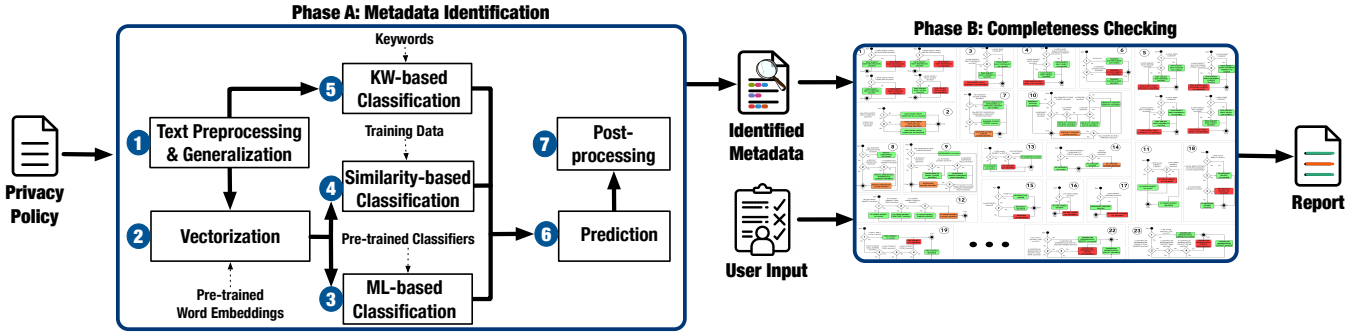


Fig. 9: Overview of the Completeness Checking Approach (CompAI).

Artificial **I**ntelligence against GDPR (thereafter referred to as *CompAI*). *CompAI* does not use deep learning (DL) architectures (e.g., LSTM [48]), since we do not have enough data for developing such models with enough accuracy. We leave investigating the possibility of using DL for future work. CompAI, shown in Fig. 9, is composed of two main phases. Phase A, *metadata identification*, takes as an input a privacy policy, and returns the metadata types that are present in this policy as an intermediary output. More precisely, Phase A results in a binary decision for each metadata type regarding whether or not it is present in the input privacy policy. Phase B, *completeness checking*, takes as an input the identified metadata types from Phase A and the user input based on the questionnaire (explained in Sec. 4). Phase B then returns a detailed report about whether the input privacy policy is complete according to GDPR. We elaborate these phases next.

5.1 Metadata Identification (Phase A)

Phase A uses a combination of NLP and ML to identify the metadata types that are present in a given privacy policy. Our metadata identification approach aims to solve a hierarchical, multi-label and multi-class classification problem. The nature of the problem is visible from the conceptual model in Fig. 3, where most level-1 metadata types are further specialized into sub-metadata types (level-2 and level-3). Multi-label classification reflects the fact that a sentence in the privacy policy can discuss one or more metadata types. Therefore, our solution can predict one or more potential labels (metadata types) for each sentence in the input privacy policy. Our approach considers a *sentence* as the unit of analysis. A sentence refers to the textual entity that results from applying the sentence splitting module in the NLP pipeline (Fig. 2), irrespective of whether the sentence identified by this module corresponds to a grammatical sentence. The rationale behind using sentences rather than phrases is that the former are more likely to contain the context necessary for understanding their meaning [49] and thus lead to more accurate classification results.

Phase A is further composed of seven steps. In the first two steps, the text of the input privacy policy is preprocessed, generalized and transformed into a mathematical representation (vectors). In steps 3-5, we classify the sentences of the input privacy policy into one or more metadata types using three classification methods based on ML, semantic similarity, and keywords. As we will see, relying on these complementary methods is necessary to overcome the complexity of the hierarchical classification problem. In step 6, we combine the results of steps 3, 4, and 5 to predict metadata types for each sentence in the input privacy policy. In the last step, we refine the results through post-processing. We explain these steps in detail next.

5.1.1 Text Preprocessing and Generalization (Step 1)

In step 1, we apply the NLP pipeline (Fig. 2) to parse the input privacy policy and obtain the sentences. Using the annotations produced by the NLP pipeline, we generalize the text in each sentence by replacing specific textual entities with more generic ones. Specifically, we replace named entities (as identified by the named entity recognition module) with their types. For example, the entities "Japan" and "Hikari Bank Ltd" in Fig. 10 will be replaced with the types *location* and *organization*, respectively. Similarly, we generalize emails, postal addresses, telephone numbers, and websites, e.g., "info@hikari.jp" is replaced with *email*. The intuition behind generalization is to normalize the text such that, despite significant diversity across the privacy policies used for training (e.g., the mention of different locations), the approach can still learn common patterns and accurately predict metadata types. The generalized sentences are further normalized through lemmatization and stopword removal, e.g., in Fig. 10 "accepting" becomes "accept" and stopwords like "by" are removed.

Original Text

By accepting this policy, you are providing personal data (as defined below) to Hikari Bank Ltd, represented by the Holding Bank Services, 16, rue de Gasperich, L-5826 Hesperange, Grand Duchy of Luxembourg. If you have questions or concerns about this policy, please contact us by post: 20 Nihonbashi Honcho, Tokyo 103-8691, Japan; by email: info@hikari.jp; or by telephone: +81 3 36300941.

Preprocessed and Generalized Text

Sentence-1 accept policy provide personal data define organization represent organization address location. Sentence-2 question concern policy please contact post Nihonbashi Honcho location location email email telephone phone.

Fig. 10: Example of Text Preprocessing and Generalization.

5.1.2 Vectorization (Step 2)

Step 2 transforms the textual sentences resulting from step 1 into embeddings. To do this, we utilize the pre-trained word-vector model of 100-dimensional vectors from GloVe [22] (introduced in Sec. 2.4). Using off-the-shelf, pre-trained and context-independent (i.e., one vector per word regardless of context) word vectors increases the applicability of our approach by making it

directly applicable for analyzing new document types. For computing the sentence embedding, we first retrieve the corresponding embedding for each word in the sentence as given by the pre-trained model. Then, we average over all the word embeddings to get a single vector representing the sentence embedding. For example, the embedding of the sentence "data privacy policy" in Fig. 11 is the average of the word embeddings in that sentence, such that the first entry in the sentence embedding (i.e., -0.14414) corresponds to the average of the first entries in the word embeddings of "data", "privacy", and "policy" (i.e., -0.47099, 0.099115, and -0.060532), respectively. The objective of the vectorization step is to achieve a representation for measuring text similarity that is effective and fast to train and test. Driven by this objective, we use simple averaging of embeddings because doing so has proven to be efficient for generating sentence embeddings across a broad range of different domains and NLP tasks, including text similarity [41], [43].

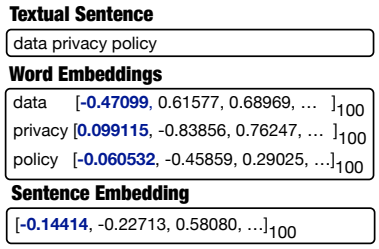


Fig. 11: Example of Vectorization.

5.1.3 ML-based Classification (Step 3)

In this step, we attempt to solve the multi-class, multi-label classification problem by transforming it into multiple binary classification problems (as explained in Sec. 2.3). To do so, we apply the pre-trained ML classifiers for predicting the presence of level-1 and level-2 metadata types in each sentence of the input privacy policy. We restrict the use of ML to level-1 and level-2 metadata types because the number of positive examples we have in our training set for level-3 metadata types is not sufficient for building accurate ML classifiers at that level.

Our classifiers are trained on a feature matrix in which each row corresponds to a sentence and the columns are the 100-dimensional sentence embedding computed in step 2. The prediction class for each classifier indicates the presence of a level-1 or level-2 metadata type in the sentence. For example, the sentence: *Your personal data might be disclosed to the tax authorities, or other third parties including legal or financial advisors, regulatory bodies, auditors and technology providers.* (number **14** in Fig. 1) is predicted as RECIPIENTS. We train the classifiers with positive examples representing the sentences that have been annotated with a particular metadata type (e.g., DATA SUBJECT RIGHT) and negative examples annotated with any other metadata type *at the same level* (i.e., all but DATA SUBJECT RIGHT). In most of the cases, we obtained imbalanced datasets with positive examples being under-represented. Inspired by Wang and Manning [50], we use a support-vector machine (SVM) classifier with its default hyper-parameters for sentence classification. SVM is widely used for text classification [51]. We address the imbalance problem in our work using under-sampling over negative examples [20]. Our preliminary experiments suggested that using both SVM for text classification and under-sampling for handling imbalanced datasets outperformed alternatives, e.g., using Naïve Bayes classifier or minority over-sampling. Further, as we will discuss in Sec. 6, the high accuracy obtained by our current solution alleviates the need to empirically examine alternatives.

Step 3 uses one pre-trained binary classifier for each level-1 and level-2 metadata type in the model of Fig. 3. The classifier predicts for each sentence in the input privacy policy, using its embedding vector as features, whether it should be labelled with the metadata type on which the classifier has been trained. For example, the sentence: *the right to request erasure, restriction, portability, and to object to the processing of your personal data* (numbers **29** – **32** in Fig. 1) is predicted by the corresponding binary classifiers as the level-2 metadata types DATA SUBJECT RIGHT.{ERASURE, RESTRICTION, PORTABILITY, AND OBJECT}. If a metadata type is not predicted by any binary classifier to be present in a sentence, then this metadata type is deemed as absent. The resulting classifications are passed on to step 6 (Prediction).

5.1.4 Similarity-based Classification (Step 4)

In this step, we classify each sentence of the input privacy policy based on how similar it is to the group of sentences, in the training set, that are annotated with a certain level-1 or level-2 metadata type. Restricting the use of this classification to level-1 and level-2 metadata types is due to the same reason explained in step 3.

Step 4 creates one group for each level-1 and level-2 metadata type. Similar to step 3, this step characterizes a sentence using the vector representation built in step 2. Since an individual sentence can have multiple metadata-type annotations, the same sentence embedding can be part of several groups. Each group is represented by a single vector which is computed as the average of all sentence embeddings in that group. To predict whether a sentence (S) should be annotated with a certain metadata type (t), we compute the *cosine similarity* between the sentence embedding ($\vec{S}$) and the vector capturing the average embedding of the group of sentences annotated by t (i.e., $\vec{t}$) in the training set. If the cosine similarity is above a pre-specified threshold, we predict t to be a metadata type for S. We set the value of this threshold to 0.9. This value was empirically obtained by evaluating the accuracy of the prediction using a range of similarity threshold values between 0.5 and 0.9, with a step of 0.01, on a subset of the privacy policies in the training set. Threshold values less than 0.5 are not considered because they fail to capture similarity.

To illustrate, consider our example in Fig. 1. The cosine similarity between the group of sentences annotated with PD ORIGIN.INDIRECT (t) and the vector representation ($\vec{S}$) of the sentence: *information obtained from third parties including [...]* (number **10**) is 0.91, while the cosine similarity with $\vec{S'}$ of the sentence: *Your personal data might be disclosed to the tax authorities, or other third parties [...]* (number **14**) is 0.43. As a result, S is classified as t while S' is not. The results of this step are passed on to step 6 (Prediction).

5.1.5 Keyword-based Classification (Step 5)

In this step, we conduct a keyword search (from a predefined list) over the (textual) sentences in the input privacy policy. If a sentence S contains one or more of the keywords

associated with metadata type t, then we predict that S should be annotated with t. For example, the sentence (number **16** in Fig. 1): *We may also transfer your personal data to countries outside the European Union (including Japan) on the basis of: European Commission's adequacy decisions, certified by the **APPI Japan Scheme**.* will be predicted as TRANSFER OUTSIDE EUROPE.ADEQUACY DECISION.COUNTRY, since it contains keywords indicating this metadata type (highlighted in bold). Another important point in relation to keywords is that, the text generalization performed in step 1 improves the efficacy of keyword search. For example, number **5** in Fig. 1 represents an email address that is generalized with *email*. Thus, including *email* as a keyword enables identifying the metadata type CONTACT.EMAIL. We have collected a list of keywords covering all of the metadata types in Fig. 3. We elaborate in Sec. 6 on how we obtain these keywords. The results of this step are passed on to the next step (Prediction).

5.1.6 Prediction (Step 6)

This step combines the classification results produced based on ML (step 3), semantic similarity (step 4) and keyword search (step 5) to produce a final recommendation about which metadata types should be ascribed to a given sentence.

The reason why we use three different classifiers is to overcome the complexity of the hierarchical multi-class classification problem and hence improve the accuracy of predicting the potential labels for each sentence in the privacy policy. Each method alone has some limitations. On the one hand, relying only on keyword search is not sufficient because of the limitations discussed in Sec. 1. ML-based and similarity-based classifications, on the other hand, are restricted to level-1 and level-2 metadata types and are further more accurate for the former since the number of datapoints gets much smaller at level-2. Thus, ensembling the three classifiers yields accurate predictions as we will show in our empirical evaluation (Sec. 6).

Our strategy for combining the above classification methods is elaborated in Algorithm 1. The algorithm applies ML-based and similarity-based classifiers for predicting both level-1 and level-2 metadata types. Despite having keywords for all metadata types, the use of keyword search in our approach is limited. We use keywords to predict level-3 that is specializing an already-predicted (level-2) metadata type or to provide supporting evidence for predicting a level-2 metadata type in case its level-1 cannot be predicted.

The algorithm starts with an initially empty set of labels ($\mathcal{M}$) – line 1. A label can be represented as *level-1.level-2.level-3* for specialized metadata, e.g., DATA SUBJECT RIGHT.COMPLAINT.SA. A partial label can also be predicted, e.g., DATA SUBJECT RIGHT.COMPLAINT or CHILDREN, in case the metadata has no specialization or there is no evidence that supports predicting a specialization.

Level-1 and Level-2 Metadata. The algorithm predicts a level-1 metadata type and its corresponding level-2 specializations in two cases, Case 1 (lines 4 – 10) and Case 2 (lines 11 – 17). Case 1 applies when some level-1 metadata type can be predicted; the algorithm then attempts to predict its level-2 type. Case 2 applies when Case 1 fails to predict a level-1 type but there is strong support for predicting its level-2 type. The rationale behind Case 2 is that when two classifiers jointly predict a level-2 metadata type (as we elaborate next), then their predictions should compensate for the absence of a level-1 prediction in Case 1. If Case 2 leads to a prediction of a certain level-2 metadata type (e.g., VITAL INTEREST), then this will be considered as an indirect indication for predicting the level-1 of that metadata type (e.g., LEGAL BASIS).

Algorithm 1 Metadata Prediction for a Sentence S

Require: $\vec{S}$: vector representation of S; cf_1, cf_2: binary classifiers trained on level-1 and level-2 metadata types for $\vec{S}$, respectively; $\vec{av}(t)$: average vector for the group of sentences annotated with metadata type t; $\mathcal{K}$: set of metadata types predicted based on keyword search in S; C_{ID}, CR_{ID}: the values of CONTROLLER.IDENTITY and CONTROLLER REPRESENTATIVE.IDENTITY, respectively.
Output: $\mathcal{M}$: a set of metadata types predicted for S

1: $\mathcal{M} \leftarrow \emptyset$
2: Let $\mathcal{L}_1$ be the set of level-1 metadata types
3: **for** $\ell_i \in \mathcal{L}_1$ **do**
4: **if** cf_1 predicts ℓ_i **or** $\texttt{sim}(\vec{S}, \vec{av}(\ell_i)) \geq 0.9$ **then** // *Predict level-1 & level-2 (Case 1)*
5: Add ℓ_i to $\mathcal{M}$
6: **for** ℓ_j s. t. ℓ_j is a (level-2) specialization of ℓ_i **do**
7: **if** cf_2 predicts ℓ_j **or** $\texttt{sim}(\vec{S}, \vec{av}(\ell_j)) \geq 0.9$ **then**
8: Add $\ell_i.\ell_j$ to $\mathcal{M}$
9: **end if**
10: **end for**
11: **else** // *Predict level-1 & level-2 (Case 2)*
12: **for** ℓ_j s. t. ℓ_j is a (level-2) specialization of ℓ_i **do**
13: **if** (cf_2 predicts ℓ_j **and** $\texttt{sim}(\vec{S}, \vec{av}(\ell_j)) \geq 0.9$) **or** ($cf_2$ predicts ℓ_j **and** $\ell_j \in \mathcal{K}$) **or** ($\texttt{sim}(\vec{S}, \vec{av}(\ell_j)) \geq 0.9$ **and** $\ell_j \in \mathcal{K}$) **then**
14: Add $\ell_i.\ell_j$ to $\mathcal{M}$
15: **end if**
16: **end for**
17: **end if**
18: **end for**
19: **if** S contains C_{ID} **then**
20: Add CONTROLLER.IDENTITY to $\mathcal{M}$
21: **else if** S contains CR_{ID} **then**
22: Add CONTROLLER REPRESENTATIVE.IDENTITY to $\mathcal{M}$
23: **end if**
24: **for** $\ell_i.\ell_j \in \mathcal{M}$ **do** // *Predict level-3*
25: **for** ℓ_q s. t. ℓ_q is a (level-3) specialization of ℓ_j **do**
26: **if** $\ell_q \in \mathcal{K}$ **then**
27: Add $\ell_i.\ell_j.\ell_q$ to $\mathcal{M}$
28: **end if**
29: **end for**
30: **end for**

Case 1: If a level-1 metadata type (ℓ_i) is predicted for the sentence (S) via the (level-1) ML-based classifier (cf_1) or by the similarity-based classifier (line 4), then ℓ_i is added to $\mathcal{M}$ (Line 5). If the predicted ℓ_i has any specialization, the algorithm attempts to further predict its level-2 metadata type (ℓ_j). If ℓ_j is predicted by the (level-2) ML-based (cf_2) or similarity-based classifiers (line 7), then the annotation $\ell_i.\ell_j$ is added to $\mathcal{M}$. Since ℓ_i has been confirmed earlier, it is sufficient to get ℓ_j predicted by one classifier (excluding

keyword-based for the reasons mentioned earlier). Regardless of whether or not the algorithm succeeds to predict ℓ_j, ℓ_i is still added to $\mathcal{M}$ (line 5). The rationale is that pinpointing the sentence that discusses ℓ_i helps the legal experts easily locate ℓ_j which is expected to appear in the following sentences.

Case 2: If the level-1 metadata type (ℓ_i) cannot be directly predicted, the algorithm checks whether its level-2 (ℓ_j) can still be predicted. Case 2 requires ℓ_j to be predicted by two classifiers (line 18). Specifically, the label $\ell_i.\ell_j$ is added to $\mathcal{M}$ if at least one of the following three preconditions is satisfied: ℓ_j is predicted by the (level-2) ML-based (cf_2) and similarity-based classifiers (line 13 – first condition). Alternatively, ℓ_j is predicted by either cf_2 or the similarity-based classifier, and ℓ_j is further predicted by keyword search (line 13 – second two conditions). In Case 2, ℓ_i is automatically added to the set of annotations to get the hierarchical label, since there is enough evidence to support the prediction of ℓ_j. To obtain a joint prediction by the three classifiers in Case 2, we considered all possible combinations as described in the set of rules – line 13. These rules include combining the predictions of (i) ML-based with similarity-based, (ii) ML-based with keyword-based, and (iii) similarity-based with keyword-based.

The level-2 metadata types CONTROLLER.IDENTITY (C_{ID}) and CONTROLLER REPRESENTATIVE.IDENTITY (CR_{ID}) are provided by the user through the questionnaire explained in Sec. 4 (as answers to Q1 and Q5). If C_{ID} (or CR_{ID}) occurs in the sentence S, then CONTROLLER.IDENTITY (or CONTROLLER REPRESENTATIVE.IDENTITY) is added to $\mathcal{M}$ (lines 19 – 23).

Level-3 Metadata. Recall that ML-based and similarity-based classifiers are not applicable to level-3 metadata types due to the lack of positive examples in our training data. Therefore, we use keyword-based classification only. The algorithm attempts to predict level-3 metadata types based on any already predicted level-1.level-2 annotation. Specifically, the algorithm considers all level-3 metadata types that specialize some level-2 metadata type already in $\mathcal{M}$. For each level-2 metadata type that is predicted, if its level-3 is predicted by keyword search, then level-3 metadata type is added to $\mathcal{M}$ (line 27).

5.1.7 Post-processing (Step 7)

In the seventh and final step of our metadata identification approach, we refine the results of step 6 by considering the metadata types predicted for the sentences surrounding a given sentence. The intuition behind this step is the observation that specializations of certain metadata types are discussed in consecutive sentences of privacy policies. Based on this observation, when a sentence S is predicted as having a specific metadata type t, the surrounding context, specifically the preceding and succeeding sentences, can provide a confirmatory measure about whether t is a reliable prediction for S.

We employ several such context-based heuristics for post-processing DATA SUBJECT RIGHT, TRANSFER OUTSIDE EUROPE, and LEGAL BASIS, since these types are often discussed in consecutive sentences in the privacy policy. The heuristic states that if some level-2 metadata type (ℓ_j) is predicted for a sentence (S), then we look at the n preceding and n succeeding sentences, such that n equals the number of the metadata types at the same level of ℓ_j. The number n accounts for the possibility to discuss the level-2 of a metadata type each in a separate sentence. For example, eight sentences before and after a sentence are considered to belong to the context for the level-2 metadata type DATA SUBJECT RIGHT.PORTABILITY, where the level-2 metadata types of DATA SUBJECT RIGHT can be listed in eight sentences at most.

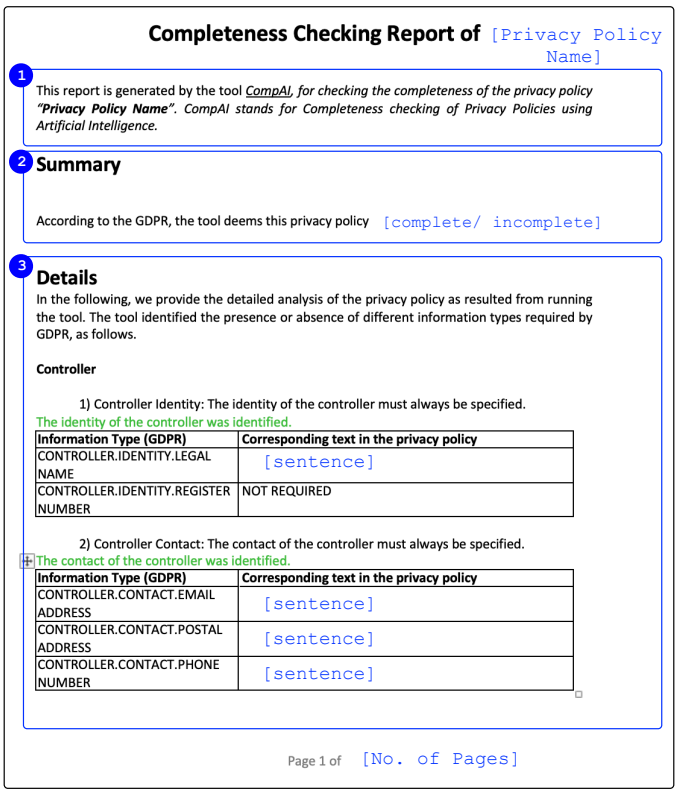
If none of these surrounding sentences are predicted to discuss a metadata type relevant to ℓ_j, then we remove from the annotations for S the predicted label that includes ℓ_j. This is because the context around S lends no support to ℓ_j being a correct annotation for S. A metadata ℓ'_j is said to be relevant to ℓ_j if it belongs to the same level-1 metadata type. To illustrate, let S be sentence number **16** in Fig. 1. This sentence can be falsely classified as DATA SUBJECT RIGHT.PORTABILITY, because of the misleading words ("transfer", "personal", "data"). In post-processing, we look at the metadata types predicted for the *eight* preceding (i.e., **8 – 15**) and *eight* following sentences (i.e., **17 – 24**) to decide if there is enough support to confirm the prediction of S. If none of the predicted metadata types in the context is relevant to DATA SUBJECT RIGHT.PORTABILITY, then we filter out this prediction assuming it is false.

5.2 Completeness Checking (Phase B)

Phase B takes as an input: (1) the output of Phase A representing the predicted metadata types in the input privacy policy, and (2) the answers of the user to the six questions discussed in Sec. 4. Phase B then returns a detailed report on completeness analysis as the final output of our overall approach. Fig. 12 shows the template of the report which CompAI generates. The first part is a preamble including the name of the privacy policy. The second part presents a summary about the final decision regarding completeness. The third part shows the details of the identified metadata types under each completeness criterion. If the metadata type is not identified in any sentence, the report will show "NOT FOUND" and indicate a violation or warning accordingly. If the metadata type is not required because the presence of another metadata type is sufficient, or if the criterion is not applicable based on the answers to the questionnaire, then the report will reflect this through the respective statements "NOT REQUIRED" or "NOT APPLICABLE".

This phase implements the completeness criteria shown in Fig. 7 and Fig. 8. Using our running example in Fig. 1, the expected answers to the questionnaire are the following. The CONTROLLER.IDENTITY is *Hikari Bank Ltd* (**Q1**), personal data will likely be transferred outside the EU (**Q2**), there will be recipients other than the CONTROLLER (**Q3**), the core activities include processing special categories (**Q4**), processing of personal data will take place in Europe (**Q5**), and finally the personal data will be collected both directly and indirectly (**Q6**). The answer to **Q5** requires an additional input from the user about the CONTROLLER REPRESENTATIVE.IDENTITY which is *the Holding Bank Services*.

Based on the answers given above, all completeness criteria (see Sec. 4) need to be checked in this step. For example, the criterion **C22** states that PD PROVISION OBLIGED should


Completeness Checking Report of [Privacy Policy Name]

1 This report is generated by the tool *CompAI, for checking the completeness of the privacy policy "**Privacy Policy Name**". CompAI stands for Completeness checking of Privacy Policies using Artificial Intelligence.*

2 **Summary**

According to the GDPR, the tool deems this privacy policy [complete/ incomplete]

3 **Details**

In the following, we provide the detailed analysis of the privacy policy as resulted from running the tool. The tool identified the presence or absence of different information types required by GDPR, as follows.

Controller

1) Controller Identity: The identity of the controller must always be specified.

The identity of the controller was identified.

Information Type (GDPR)	Corresponding text in the privacy policy
CONTROLLER.IDENTITY.LEGAL NAME	[sentence]
CONTROLLER.IDENTITY.REGISTER NUMBER	NOT REQUIRED

2) Controller Contact: The contact of the controller must always be specified.

The contact of the controller was identified.

Information Type (GDPR)	Corresponding text in the privacy policy
CONTROLLER.CONTACT.EMAIL ADDRESS	[sentence]
CONTROLLER.CONTACT.POSTAL ADDRESS	[sentence]
CONTROLLER.CONTACT.PHONE NUMBER	[sentence]

Page 1 of [No. of Pages]

Fig. 12: Template of Completeness Analysis Report.

be present in a privacy policy when the answer to Q6 is either PD ORIGIN.DIRECT or both, and at the same time the legal basis of processing personal data is either LEGAL BASIS.LEGAL OBLIGATION or LEGAL BASIS.CONTRACT.TO ENTER CONTRACT. A violation of this criterion raises an incompleteness issue. In our example in Fig. 1, both the above-mentioned metadata types are found in the privacy policy, in sentences **20** and **25**, respectively. As a result, we have to find the metadata type PD PROVISION OBLIGED in the same policy; this comes in sentence **26**. Had this sentence not been correctly identified by phase A, either due to inaccurate prediction or because it is actually missing in the policy, then this criterion would have been violated. The result of Phase B is a set of detected violations and warnings for the 23 criteria due to missing metadata in the input privacy policy.

6 EMPIRICAL EVALUATION

6.1 Implementation

We have implemented our approach using Java. The implementation has ≈ 7500 lines of code excluding comments and third-party libraries. For the basic NLP pipeline, we use the DKPro toolkit [52]. For text generalization, we use regular expressions available in Java. We transform words into embeddings by utilizing the publicly available pre-trained word embeddings from GloVe [22]. Noting that our implementation is Java-based, we perform operations on word embeddings using Deeplearing4j [53]. Our metadata identification approach uses ML-based classification. For classification and handling imbalance in our dataset, we employ WEKA [54], [55]. For computing similarity between two textual entities in the similarity-based classification, we use Cosine Similarity [56].

6.2 Data Collection Procedure

Our data collection aimed at collecting and annotating privacy policies according to the conceptual model of Fig. 3. Specifically, we collected from the fund domain a total of 234 privacy policies, of which about 60% were provided to us by Linklaters. For the remaining 40%, we downloaded privacy policies from companies in the fund registry of Luxembourg, which has a substantial footprint in fund management [57]. We chose the fund domain because it is one of the main domains in which Linklaters is active. Focusing on the fund domain has an impact on the external threat to validity, as we will elaborate in Sec. 7. Nonetheless, the conceptual model described in Sec. 3 is domain-agnostic, noting that it was derived from GDPR and the (domain-independent) knowledge of legal experts about privacy policies.

Our data collection was performed in two steps. In the first step, a batch of 30 policies was annotated by the third author of this paper who has acquired domain expertise through close interaction with Linklaters. For annotating this first batch, hypothesis coding was applied (as explained in Sec. 3). During this step, we also drafted detailed guidelines with illustrative examples to explain the annotation process. These guidelines were then shared with the external annotators.

The second batch (204 policies) was annotated by four third-party annotators (non-authors). Three of these individuals are graduate students in social sciences; they are native English speakers with considerable prior exposure to legal documents. The fourth annotator is a computer-science graduate student with an excellent command of English and six months of prior internship experience on legal text processing in industry. All four annotators attended two four-hour training sessions, focused on GDPR concepts and the definitions of our metadata types. The annotators were further provided with the guidelines drafted in the first step, during which the conceptual model was also refined. To obtain an unbiased evaluation, the conceptual model was frozen before the second step started since a subset of the annotations is used for evaluating our approach as we explain later in this section. Thus, the two steps of our annotation process are performed in a strict sequence. During the entire annotation process, the annotators kept track of the keywords that were frequently used to express certain metadata types.

The annotators were asked to annotate each sentence in the privacy policies with the metadata types that they deemed to be present in the sentence. When no metadata was present, they classified the sentence as *no metadata identified*. To illustrate, consider the example in Fig. 1. Numbers **27 – 34** represent one sentence that includes multiple metadata types related to DATA SUBJECT RIGHT. The annotators would then annotate the sentence with *all* the metadata types that are present in the sentence. To measure the quality of our dataset, we computed the interrater agreement using Cohen's Kappa (κ) [58]. Specifically, we selected 24 privacy policies (≈10% of our dataset) using random stratification to

TABLE 3: Document Collection Results.

		Total		Training Data (T)		Test Data (E)	
Level	**Metadata**	**Manifestations**	**Sentences**	**Manifestations**	**Sentences**	**Manifestations**	**Sentences**
L1	**CONTROLLER**	-	-	-	-	-	-
L2	IDENTITY	221	799	175	415	46	384
L2	CONTACT	151	652	117	466	34	186
L1	**CONTROLLER REPRESENTATIVE**	-	-	-	-	-	-
L2	IDENTITY	19	36	16	33	3	3
L2	CONTACT	21	63	18	60	3	3
L1	**DPO**	-	-	-	-	-	-
L2	CONTACT	125	462	104	404	21	58
L1	**DATA SUBJECT RIGHT**	224	3352	180	2651	44	701
L2	ACCESS	209	378	167	304	42	74
L2	RECTIFICATION	211	345	169	267	42	78
L2	RESTRICTION	170	311	137	247	33	64
L2	COMPLAINT	172	286	137	219	35	67
L3	SA	172	264	138	217	34	47
L2	ERASURE	196	386	157	296	39	90
L2	OBJECT	181	484	145	373	36	111
L2	PORTABILITY	163	263	131	210	32	53
L2	WITHDRAW CONSENT	169	395	136	322	33	73
L1	**LEGAL BASIS**	231	4511	185	3238	46	1273
L2	CONTRACT	161	553	127	366	34	187
L3	CONTRACTUAL	123	275	89	183	34	92
L3	TO ENTER CONTRACT	73	105	18	30	55	75
L3	STATUTORY	20	25	13	16	7	9
L2	PUBLIC FUNCTION	73	122	51	84	22	38
L2	LEGITIMATE INTEREST	214	2424	170	1846	44	578
L2	VITAL INTEREST	17	24	10	15	7	9
L2	CONSENT	180	554	141	423	39	131
L2	LEGAL OBLIGATION	200	1028	155	704	45	324
L1	**TRANSFER OUTSIDE EUROPE**	178	823	148	707	30	116
L2	ADEQUACY DECISION	47	76	64	109	4	4
L3	COUNTRY	47	76	45	74	2	2
L2	SAFEGUARDS	136	280	113	230	23	50
L3	BINDING CORPORATE RULES	50	64	46	58	4	6
L3	EU MODEL CLAUSES	96	129	76	100	20	29
L2	SPECIFIC DEROGATION	17	20	10	13	7	7
L3	UNAMBIGUOUS CONSENT	16	18	10	12	6	6
L1	**PD ORIGIN**	216	1904	310	125	45	356
L2	DIRECT	165	436	125	310	40	126
L2	INDIRECT	209	1356	164	1129	45	227
L3	THIRD PARTY	113	294	80	206	33	88
L3	PUBLICLY	79	127	55	80	24	47
L3	COOKIE	155	668	123	597	32	71
L1	**PD CATEGORY**	228	2209	182	1860	46	349
L2	SPECIAL	98	265	69	198	29	67
L2	TYPE	33	71	24	60	9	11
L1	**RECIPIENTS**	209	1599	167	1369	42	230
L1	**PD TIME STORED**	200	873	162	738	38	135
L1	**PD PROVISION OBLIGED**	128	281	102	230	26	51
L1	**PROCESSING PURPOSES**	158	1422	112	1099	46	323
L1	**PD SECURITY**	182	883	140	717	42	166
L1	**AUTO DECISION MAKING**	84	295	63	224	21	71
L1	**CHILDREN**	24	70	19	58	5	12

ensure that this subset covers some annotations from each of the four third-party annotators. The annotated sentences in the 24 privacy policies were independently checked by the first author who had done more than half a year of training on the completeness checking of privacy policies before validating the annotations. The interrater agreement is computed for level-1 metadata types only. The agreement obtained at a sentence-level is on average $\kappa = 0.71$ indicating "moderate agreement" [59]. We observed that most of the disagreements occurred over the identification of PROCESSING PURPOSES and LEGAL BASIS. Since these two metadata types are usually related and often span multiple sentences in a privacy policy, such disagreements are expected. LEGAL BASIS ensures that the processing of personal data for certain purposes (i.e., PROCESSING PURPOSES) is lawful when some conditions are satisfied in line with the sub-metadata of LEGAL BASIS. At a privacy-policy level, we obtained an average κ score of 0.87, indicating

"strong agreement" [59]. This suggests that the annotators strongly agreed on which metadata types are present in a given privacy policy. We believe this agreement is acceptable in our context given that our automated solution aims at identifying metadata at a privacy-policy level.

Table 3 shows the results of our document collection. Following best practices, the entire document collection (234 privacy policies) is split randomly into two subsets containing about 80% and 20% of the policies, respectively used for training and development (186 policies) and for evaluation (48 policies). The first batch used in our annotation for finalizing the model is included in the training set. Hereafter, we refer to the dataset used for training as *T*, and to that used for testing as *E*. We use *E* for answering the research questions (RQs).

The table provides statistics about the entire dataset, *T* and *E*. Specifically, we provide per metadata type t: the number of manifestations of t in our document collection, and the number of sentences that are annotated with t. A *manifestation* of t is counted once per privacy policy when t appears in that privacy policy (i.e., there is at least one sentence annotated with t). We compute the number of manifestations of t in our document collection as the sum of the manifestations of t across the privacy policies. For example, the number of manifestations of DATA SUBJECT RIGHT in Table 3 is 224 (i.e., DATA SUBJECT RIGHT appears in 224 out of 234 privacy policies). Further, this metadata type was annotated in a total of 3352 sentences across the privacy policies in our collection. We note that none of the sentences in our dataset is annotated with CONTROLLER, CONTROLLER REPRESENTATIVE or DPO as separate labels. These metadata types always appear with their specializations, e.g., CONTROLLER.IDENTITY or CONTROLLER.CONTACT.

6.3 Evaluation Procedure

We answer RQ4 – RQ6 by conducting the experiments explained next.

EXPI. This experiment answers RQ4. We assess the accuracy of our metadata identification approach. To do so, we run our approach and compare the results against manual annotations of the privacy policies in the test set *E* (defined in Sec. 6.2). We evaluate, in EXPI, the manifestations of metadata types detected by our approach in a given privacy policy. We recall that a manifestation of a metadata type is counted only once per privacy policy, even if the metadata type appears in multiple sentences. Designing our evaluation around manifestations (instead of actual sentences) is driven by our objective, which is completeness checking. To verify the completeness criteria, presented in Sec. 4, one needs to ascertain whether or not a manifestation of the metadata type exists in the privacy policy. To illustrate, consider criterion **C6** as an example. In C6, we need to find DATA SUBJECT RIGHT.PORTABILITY in the privacy policy, when the legal ground is based on CONTRACT. If our approach is able to identify manifestations of CONTRACT and PORTABILITY, then the completeness of the policy can be properly checked.

Let a manifestation of the metadata type t_i be represented, in a privacy policy, by $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$, such that $\mathcal{S}$ is the set of sentences that are annotated with t_i according to our ground truth. Our approach deems a manifestation of t_i as present, if t_i is predicted as a label for at least one sentence in the privacy policy. Following this, we define a *True Positive (TP)* when the approach correctly identifies a manifestation of t_i, i.e., the approach finds at least one sentence $s_j \in \mathcal{S}$. A *False Positive (FP)* is when the approach falsely identifies a manifestation of t_i, i.e., the approach finds a group of sentences $\mathcal{S}'$ to be about t_i, such that either $\mathcal{S}' \nsubseteq \mathcal{S}$ or there is no manifestation of t_i in the privacy policy according to our ground truth. A *False Negative (FN)* is when the approach misses a manifestation of t_i, i.e., the approach does not find any $s_j \in \mathcal{S}$.

In EXPI, we report the overall *Accuracy (A)*, *Precision (P)*, *Recall (R)*, and the harmonic mean *F-measure (F_β)* for each metadata type across *E*. We compute these metrics as $A = (TP + FN)/(TP + FP + FN + TN)$, $P = TP/(TP + FP)$, $R = TP/(TP + FN)$, and $F_\beta = (1 + \beta^2) * (P * R)/(\beta^2 * P + R)$. For metadata identification, recall is more important than precision, since the metadata types identified by the approach will be used to check completeness of privacy policies. This means that, if a metadata type is falsely introduced by the approach, it can be reviewed and filtered out by an analyst, whereas missing metadata types will require the analyst to review the entire privacy policy. In EXPI, we report the *F2-measure* (i.e., β = 2) to show the evaluation in favor of recall. We choose F2-measure for two reasons: First, values of $\beta \geq 2$ do not change the reasoning about our evaluation. Second, despite recall being more important, precision still has a great value, a very low precision (too many false positive errors) will require more time and effort in filtering the erroneous findings.

EXPII. This experiment answers RQ5. We evaluate the accuracy of checking the completeness criteria on our test set. The unit of evaluation in EXPII is *an incompleteness issue* resulting from an unsatisfied criterion in a given privacy policy. Correspondingly, we redefine a *TP* as an incompleteness issue found correctly by our approach, a *FP* as an incompleteness issue found by our approach when the criterion is satisfied, a *FN* as an incompleteness issue missed by our approach, and a *TN* when our approach correctly concludes that there is no incompleteness issue in the privacy policy. Similar to EXPI, we report *A*, *P*, *R*, and *F2-measure*.

EXPIII. This experiment answers RQ6. We compare our *AI-based* approach for completeness checking to a simple approach that uses keyword search (hereon, referred to as *KW-based*). The latter predicts a manifestation of a certain metadata type t_i, in a given privacy policy, if at least one keyword associated with t_i is present in any sentence in this policy. We note that keyword search is introduced as one of the classifiers in our AI-based approach (Step 5 in Fig. 9). To have a fair comparison, the list of keywords used in our approach is the same one used in the baseline. In EXPIII, we compare our approach against KW-based using the same evaluation metrics defined in EXPI for metadata identification, and in EXPII for completeness checking.

6.4 Results and Discussion

In this section, we describe the results and answer the RQs stated in Sec. 1.

RQ4. How accurate is our proposed approach in identifying GDPR-relevant metadata in privacy policies?

Table 4, on the left-hand side, shows the results of EXPI. As explained in Sec. 6.3, the results are obtained by running our metadata identification approach on the test set (E) which is comprised of 48 privacy policies. The table reports the accuracy, precision, recall and F2-measure computed on manifestations of the metadata. We show in Table 3 the total number of manifestations of each metadata in E. Out of the 56 metadata types (see Fig. 3), we exclude the evaluation of CONTROLLER.IDENTITY and CONTROLLER REPRESENTATIVE.IDENTITY because they are given as input by the user, and are looked up in the privacy policy rather than being identified like the other metadata types (see Algorithm 1). We also exclude TRANSFER OUTSIDE EUROPE.ADEQUACY DECISION.{TERRITORY *and* SECTOR} because we have no examples in our experimental material (both for training or testing). To summarize the different metrics, we report the *micro average* across the different metadata types, by computing the metrics on all TPs, FPs, FNs, TNs found across all metadata types.

Accuracy evaluates how the metadata identification approach performs in correctly predicting the manifestations of metadata in the privacy policies. Apart from the excluded metadata types (explained above), the table shows that the presence or absence of all metadata types are identified with an accuracy greater than 80%. The relatively low accuracy in the case of PD ORIGIN.DIRECT and PD ORIGIN.INDIRECT.THIRD PARTY is due to eight manifestations which our approach identifies, but the sentences representing these manifestations predicted by the approach were not the same as the ones in the ground truth. We thus counted these manifestations as both FPs and FNs. As for PD PROVISION OBLIGED, the approach produces 12 errors and achieves a relatively low accuracy: ≈75.5%. We note that this metadata is usually expressed in a conditional statement, e.g., if the individual fails to provide the personal data as needed, there will be consequences. Conditional sentences in English take multiple forms. Thus, semantic analysis would be required to improve the accuracy of identifying this metadata.

Precision reflects how many actual manifestations of the metadata are correctly identified by the approach out of the total number of identified manifestations. Our approach achieves a precision greater than 80% for 51 out of the 54 metadata types. In the case of metadata type LEGAL BASIS.CONTRACT.TO ENTER CONTRACT, at level L3, the reason for achieving low precision is the reliance on keyword search. Keywords can easily introduce false positives as we elaborate later in our analysis under RQ6. The same reasons mentioned above for the low accuracy of PD PROVISION OBLIGED and PD ORIGIN.INDIRECT.THIRD PARTY can also explain the low precision of these two metadata types. In total, our approach introduces 119 false positives out of 1449 identified manifestations.

Recall assesses how many actual manifestations of metadata types in the privacy policies are also correctly identified. The table shows that we achieve a high recall for all metadata types, except for CONTROLLER REPRESENTATIVE.CONTACT and PD PROVISION OBLIGED. Note that, in E, we only have three manifestations of CONTROLLER REPRESENTATIVE.CONTACT, and the low recall (66.7%) is due to missing only one manifestation. In total, our approach missed 68 manifestations from a total of 1448 actual manifestations of metadata type in E.

The answer to RQ4 is that our metadata identification approach achieves an average accuracy, precision, recall and F2-measure of 93.4%, 92.1%, 95.3% and 94.9%, respectively.

RQ5. How accurate is our approach in checking the completeness of privacy policies?

Table 5, on the left-hand side, shows the results of EXPII. We evaluate in RQ5 how well our completeness criteria (see Sec. 4) can detect incompleteness issues, given the metadata identified by the approach and evaluated in RQ4. An incompleteness issue can be either a violation or a warning (as defined in Sec. 4). The table reports the number of TPs, FPs, FNs and TNs (redefined for EXPII in Sec. 6.3) in addition to the evaluation metrics, namely accuracy (A), precision (P), recall (R) and F2-measure (F2).

We note that seven criteria lead to warnings, namely C6, C11 – C14, C16 and C18. The remaining criteria lead to violations. We also note that C1, C2, C5, C20, and C21 are concerned with the unconditional presence of mandatory metadata types, whereas the criteria C3, C4, C10 – C19, C22 and C23 need to be checked only in specific situations based on the answers provided on the questionnaire (explained in Sec. 4). For the latter set of criteria, we assume in our evaluation that they *always* need to be checked. The criteria C6 – C9, C11 – C14, C16, C18 and C22 are concerned with metadata types that need to be present only if some other metadata types are also present in the same privacy policy.

Violations. Out of 285 genuine violations in the test privacy policies, our completeness criteria correctly detect 261, while introducing 16 false positives. This results in a precision of 94.2% and a recall of 91.6%.

Table 5 shows that the approach introduces 40 errors (16 FPs and 24 FNs) that led to violations. We analyzed the reason for having these errors. Out of the 40 errors, 26 are originated from false positives in the metadata identification results. The remaining 14 errors are due to missed metadata types (FNs) across seven criteria. Specifically, one or two missed metadata types yielded errors in **C02**, **C04**, **C08**, **C09** and **C21**, whereas five missed metadata types yielded errors in **C22**. The low precision and recall values for the completeness checking of **C22** are in part due to the accuracy of identifying PD PROVISION OBLIGED. As we explained in RQ4, this metadata type requires further analysis to capture the variations of how it is expressed.

Warnings. Out of 49 genuine warnings in the test privacy policies, our completeness criteria correctly detect 39, while introducing seven false positives. This results in a precision of 84.4% and a recall of 79.6%. A total of 17 errors resulted in warnings, including seven FPs and 10 FNs. All of these errors are due to FPs from metadata identification, except one that is due to a missed metadata in one privacy policy.

Our completeness checking approach generates a report, as an output, that is shared with the analyst. The report includes not only the final decision regarding whether a privacy policy is or is not complete according to GDPR, but also a structured summary of the identified metadata types. Specifically, the report lists under each criterion the

TABLE 4: Results of Metadata Identification.

		AI-based solution (**RQ4**)				KW-based solution (**RQ6**)			
Level	**Metadata**	**A (%)**	**P (%)**	**R (%)**	**F2 (%)**	**A (%)**	**P (%)**	**R (%)**	**F2 (%)**
L1	**CONTROLLER**	-	-	-	-	-	-	-	-
L2	CONTACT	91.8	94.1	94.1	94.1	71.4	71.7	97.1	90.7
L3	PHONE NUMBER	95.8	92.9	92.9	92.9	85.4	68.4	92.9	86.7
L3	EMAIL	85.7	90.9	80.0	82.0	62.5	58.1	100	87.4
L3	LEGAL ADDRESS	86.0	88.9	85.7	86.3	49.1	51.1	82.1	73.2
L1	**CONTROLLER REPRESENTATIVE**	-	-	-	-	-	-	-	-
L2	CONTACT	97.9	100	66.7	71.4	08.2	04.3	66.7	17.2
L3	LEGAL ADDRESS	97.9	100	66.7	71.4	10.2	04.4	66.7	17.5
L1	**DATA SUBJECT RIGHT**	97.9	97.8	100	99.5	91.7	91.7	100	98.2
L2	ACCESS	88.0	90.9	95.2	94.3	84.0	87.0	95.2	93.5
L2	RECTIFICATION	100	100	100	100	70.4	81.0	81.0	81.0
L2	RESTRICTION	95.8	94.3	100	98.8	93.8	91.7	100	98.2
L2	COMPLAINT	100	100	100	100	87.5	85.4	100	96.7
L3	SA	100	100	100	100	60.4	67.6	73.5	72.3
L2	ERASURE	100	100	100	100	100	100	100	100
L2	OBJECT	97.9	97.3	100	99.4	97.9	97.3	100	99.4
L2	PORTABILITY	100	100	100	100	95.9	96.9	96.9	96.9
L2	WITHDRAW CONSENT	95.8	100	93.9	95.1	95.8	94.3	100	98.8
L1	**LEGAL BASIS**	97.9	97.9	100	99.6	95.8	95.8	100	99.1
L2	CONTRACT	85.4	82.9	100	96.0	70.8	70.8	100	92.4
L3	CONTRACTUAL	86.0	88.6	91.2	90.6	83.7	90.6	85.3	86.3
L3	TO ENTER CONTRACT	83.3	70.8	94.4	88.5	79.2	64.3	100	90.0
L3	STATUTORY	97.9	100	85.7	88.2	83.3	45.5	71.4	64.1
L2	PUBLIC FUNCTION	83.7	81.8	81.8	81.8	45.8	45.8	100	80.9
L2	LEGITIMATE INTEREST	84.0	92.9	88.6	89.4	91.7	91.7	100	98.2
L2	VITAL INTEREST	100	100	100	100	14.6	14.6	100	46.1
L2	CONSENT	93.8	95.0	97.4	96.9	81.3	81.3	100	95.6
L2	LEGAL OBLIGATION	93.9	95.7	97.8	97.3	93.8	93.8	100	98.7
L1	**TRANSFER OUTSIDE EUROPE**	97.9	96.8	100	99.3	73.5	70.7	96.7	90.1
L2	ADEQUACY DECISION	97.9	80.0	100	95.2	66.7	20.0	100	55.6
L3	COUNTRY	100	100	100	100	79.2	10.0	50.0	27.8
L2	SAFEGUARDS	97.9	95.8	100	99.1	97.9	95.8	100	99.1
L3	BINDING CORPORATE RULES	100	100	100	100	100	100	100	100
L3	EU MODEL CLAUSES	97.9	95.2	100	99.0	97.9	95.2	100	99.0
L2	SPECIFIC DEROGATION	97.9	87.5	100	97.2	60.4	20.0	57.1	41.7
L3	UNAMBIGUOUS CONSENT	97.9	85.7	100	96.8	58.3	15.0	50.0	34.1
L1	**PD ORIGIN**	93.8	93.8	100	98.7	93.8	93.8	100	98.7
L2	DIRECT	76.5	80.4	92.5	89.8	83.3	83.3	100	96.2
L2	INDIRECT	93.9	95.7	97.8	97.3	93.8	93.8	100	98.7
L3	THIRD PARTY	71.7	76.5	78.8	78.3	45.8	51.6	48.5	49.1
L3	PUBLICLY	89.6	82.8	100	96.0	83.3	75.0	100	93.8
L3	COOKIE	95.8	94.1	100	98.8	89.9	90.9	93.8	93.2
L1	**PD CATEGORY**	93.9	95.7	97.8	97.4	62.0	93.5	63.0	67.4
L2	SPECIAL	95.8	93.5	100	98.6	95.8	93.5	100	98.6
L2	TYPE	81.6	50.0	77.8	70.0	81.3	0.0	0.0	0.0
L1	**RECIPIENTS**	93.8	93.3	100	96.6	29.0	41.9	42.9	42.7
L1	**PD TIME STORED**	91.7	94.7	94.7	94.7	86.0	87.8	94.7	93.3
L1	**PD PROVISION OBLIGED**	75.5	76.9	76.9	76.9	54.2	54.2	100	85.5
L1	**PROCESSING PURPOSES**	84.3	91.3	91.3	91.3	40.0	58.1	54.3	55.1
L1	**PD SECURITY**	82.7	88.4	90.5	90.0	83.7	85.4	97.6	94.9
L1	**AUTO DECISION MAKING**	93.8	95.0	90.5	91.3	64.6	55.3	100	86.1
L1	**CHILDREN**	95.8	80.0	80.0	80.0	77.1	31.3	100	69.4
L1	**DPO**	-	-	-	-	-	-	-	-
L2	CONTACT	97.9	95.5	100	99.1	47.9	45.7	100	80.8
L3	PHONE NUMBER	100	100	100.0	100.0	62.5	5.6	50.0	19.2
L3	EMAIL	97.9	95.0	100	99.0	50.0	44.2	100	79.8
L3	LEGAL ADDRESS	91.8	81.3	92.9	90.3	20.4	17.8	57.1	39.6
	Summary	93.4	92.1	95.3	94.6	70.7	65.2	90.1	83.7

set of sentences describing the identified manifestations of the metadata types or "not found" in case no manifestations are found by our approach. The analyst can then review this summary instead of analyzing a privacy policy in its entirety. The criteria that erroneously result in violations or raise warnings due to false positives in the metadata

TABLE 5: Results of Completeness Checking.

	AI-based approach (**RQ5**)								KW-based baseline (**RQ6**)							
Criteria	**TPs**	**FPs**	**FNs**	**TNs**	**A (%)**	**P (%)**	**R (%)**	**F2 (%)**	**TPs**	**FPs**	**FNs**	**TNs**	**A (%)**	**P (%)**	**R (%)**	**F2 (%)**
C01	2	0	0	46	100	100	100	100	2	0	0	46	100	100	100	100
C02	13	1	1	33	95.8	92.9	92.9	92.9	2	0	12	34	75.0	100	14.3	17.2
C03	45	0	0	3	100	100	100	100	42	1	3	2	91.7	97.7	93.3	94.2
C04	45	1	0	2	97.9	97.8	100	99.6	2	0	43	3	10.4	100	04.4	05.5
C05	36	0	4	152	97.9	100	90.0	91.8	24	2	15	152	91.1	92.3	61.5	65.9
C07	7	4	0	37	91.7	63.6	100	89.7	7	9	0	32	81.3	43.8	100	79.5
C08	7	1	3	37	91.7	87.5	70.0	72.9	9	2	1	36	93.8	81.8	90.0	88.2
C09	31	1	1	159	99.0	96.0	96.9	96.9	30	19	2	141	89.1	61.2	93.8	84.7
C10	17	0	1	30	97.9	100	94.4	95.5	7	0	11	30	77.1	100	38.9	44.3
C15	2	0	1	45	97.9	100	66.7	71.4	0	0	3	45	93.8	n/a	0.0	n/a
C17	1	0	1	46	97.9	100	50.0	55.6	2	15	0	31	68.8	11.8	100	40.0
C19	3	0	3	42	93.8	100	50.0	55.6	2	3	4	39	85.4	40.0	33.3	34.5
C20	8	2	2	36	91.7	80.0	80.0	80.0	7	0	3	38	93.8	100	70.0	74.5
C21	1	1	1	45	95.8	50.0	50.0	50.0	1	4	1	42	89.6	20.0	50.0	38.5
C22	17	5	5	21	79.2	77.3	77.3	77.3	0	0	22	26	54.2	n/a	0.0	n/a
C23	26	0	1	21	97.9	100	96.3	97.0	2	0	25	21	47.9	100	07.4	9.1
C06	1	0	0	47	100	100	100	100	0	4	1	43	89.6	0.0	0.0	n/a
C11	6	0	0	42	100	100	100	100	1	0	5	42	89.6	100	16.7	20.0
C12	2	1	0	45	97.9	66.7	100	90.9	2	4	0	42	91.7	33.3	100	71.4
C13	3	0	0	45	100	100	100	100	3	0	0	45	100	100	100	100
C14	1	0	0	47	100	100	100	100	0	0	1	47	97.9	n/a	0.0	n/a
C16	6	1	3	38	91.7	85.7	66.7	69.8	5	4	4	35	83.3	55.6	55.6	55.6
C18	20	5	7	16	75.0	80.0	74.1	75.2	24	15	3	6	62.5	61.5	88.9	81.6
Summary	300	23	34	1035	95.9	92.9	89.8	90.4	174	82	159	977	82.7	68.0	52.3	54.8

identification (33 in total) can be easily filtered out by the analyst. If the analyst reviews the incompleteness issues only instead of the summary, our approach still fares well in identifying about 90% of the actual violations and warnings. In practice, the accuracy of our approach is sufficient to be used by diverse users, including software engineers who might lack legal expertise or legal experts who need assistance to optimize their time and effort.

Based on our assumption that all criteria need to be satisfied in order for a privacy policy to be complete according to GDPR, all of the 48 policies in the test set are incomplete. Our approach is able to correctly identify all the privacy policies that have some incompleteness issue.

The answer to RQ5 is that our completeness checking approach can detect incompleteness issues in privacy policies with an average accuracy, precision, recall and F2-measure of 95.9%, 92.9%, 89.8% and 90.4%, respectively.

RQ6. Is our approach worthwhile compared to a simpler solution?

Tables 4 and 5 show the results of EXPIII including, on the left-hand side, the results of our AI-based approach as discussed in RQ4 and RQ5, and on the right-hand side the results of the KW-based approach that we introduced in Sec. 6.3.

Metadata identification. The table suggests that there are two disadvantages of using the KW-based approach. First, not all of the metadata types can be accurately identified using keywords, e.g., RECIPIENTS. Recall our discussion in Sec. 1 about this metadata type that can include a list of organizations. A finite list of predefined keywords cannot possibly cover all organization names that might appear in RECIPIENTS or capture the diverse PROCESSING PURPOSES of personal data. Our ground truth contains a total of 42 actual manifestations of RECIPIENTS and 46 of PROCESSING PURPOSES. We note that 21 manifestations of RECIPIENTS and 17 manifestations of PROCESSING PURPOSES predicted by KW-based are counted as both FPs and FNs, because none of the identified sentences associated with these manifestations are matching the ones in the ground truth. In contrast, our AI-based approach finds only three manifestations of PROCESSING PURPOSES with irrelevant sentences. This shows that our approach is more reliable in finding the correct sentences related to the identified manifestations (35 less such errors).

The second disadvantage of KW-based is that, though it achieves a relatively good recall, this comes at the cost of precision. For example, the recall for identifying the metadata type TRANSFER OUTSIDE EUROPE.ADEQUACY DECISION using keywords is 100% but the precision is only 20%. Despite such high recall, our AI-based approach achieves an overall better F2-measure, namely +11%. To summarize, our AI-based approach misses in total 68 manifestations of metadata types (FNs) and introduces 119 FPs, whereas KW-based misses 144 FNs and produces 697 FPs (76 more FNs and 578 more FPs than our approach). As a result, we achieve a gain of ≈5% in recall and ≈27% in precision.

Completeness checking. The difference in performance becomes even clearer in the completeness checking task, which depends largely on the accuracy of metadata identification. The KW-based approach has a respective precision and

recall of 71.6% and 48.9% for detecting violations, and 55.7% and 69.4% for detecting warnings. In comparison with the total of 57 errors produced by our approach for violations and warnings, KW-based produces 241 errors (i.e., 184 more errors). Of these 241 errors, 45 are due to missed manifestations of metadata types (FNs), 15 are caused by PD CATEGORY (which is hard to capture via keywords), and the remaining 196 are originating from false positives in metadata identification. Filtering so many cases out is, compared to the 33 FPs introduced by our approach, much more time-consuming for the analyst. Our approach is therefore advantageous over the KW-based solution in terms of both precision and recall. Specifically, using a combination of NLP and ML leads to a significant improvement of ≈23% in precision and ≈43% in recall for detecting violations. The overall improvement, considering both warnings and violations, is ≈25% higher precision, ≈38% higher recall, and ≈36% higher F2-measure.

The answer to RQ6 is that our AI-based approach presents a significant improvement over merely using keywords, both in metadata identification and in completeness checking. In metadata identification, our approach outperforms the KW-based solution by an average of 22.7% in accuracy, 26.9% in precision, 5.2% in recall and 11% in F2-measure. This leads to a significant follow-on gain in completeness checking, where our approach outperforms the baseline by 24.5% in precision and 38% in recall.

7 Threats to Validity

Below, we discuss threats to the validity of our empirical results and what we did to mitigate these threats.

Internal Validity. Bias was an important concern in relation to internal validity. To mitigate bias, we curated most (≈90%) of the manual annotations through third-parties (non-authors). Another potential threat to internal validity is that the authors interpreted the text of GDPR provisions in order to create the privacy policy conceptual model presented in Fig. 3. To minimize the threat posed by a subjective interpretation, this phase was done in close collaboration with three independent legal experts from Linklaters, who have expertise in European and international laws with a focus on the data protection and financial domains. While we cannot entirely rule out subjectivity, we provide our interpretation in a precise and explicit form. In addition, our model is publicly available and thus open to scrutiny. Another threat to internal validity is our reliance on a static set of keywords. Changing this set might have an impact on the results of our automated solution. However, we believe that our set of keywords is reasonably adequate and complete since we manually created the keywords during our qualitative study in close collaboration with legal experts.

External Validity. The qualitative study through which we built our conceptual model of privacy policies is domain-agnostic: the study was rooted in GDPR and further enhanced by feedback from legal experts who had familiarity with data protection in a variety of domains. This provides a fair degree of confidence about our conceptual model being generalizable. As for our evaluation of automation accuracy of our completeness checking approach (see Sec. 6.4), and more specifically, whether the accuracy levels observed would generalize beyond the fund domain, we note that certain metadata types were rare in privacy policies from the fund domain. Furthermore, we have not yet conducted a multi-domain evaluation of our metadata identification and completeness checking approaches. For these reasons, it would be premature to make claims about how our accuracy results would carry over to other domains. That said, we believe that the core components of our automation approach, notably, our hybridized use of word embeddings, ML-based classification, similarity analysis and keyword search, provides a versatile basis for the future development of a more broadly applicable solution to check the completeness of privacy policies.

8 Related Work

Our proposed approach for checking the completeness of privacy policies spans three different tasks. The first task involves the elicitation of privacy-related requirements for GDPR compliance. The second task covers the completeness checking of privacy policies (with a GDPR focus). The last task is concerned with checking the data handling practices and privacy compliance of software against their associated privacy policies. This last task enables an implicit compliance checking of the software against the privacy-requirements stated in the provisions (GDPR, in our case). Our work concentrates on providing automation for the first two tasks, noting that the results from these two tasks also serve as an input to the third task, which we do not directly address in this article. Below, we position our work against the related work on (i) identifying privacy-policy requirements, (ii) completeness checking of privacy policies, and (iii) completeness/compliance checking of software against data protection regulations.

8.1 Elicitation of Privacy-related Requirements

Vanezi et al. [60] propose a graphical modeling language for GDPR privacy policies and a methodology for transforming such graphically-defined privacy policies into formal definitions. This work focuses on one (namely, PROCESSING PURPOSES) out of the 56 metadata types we consider in our work. Caramujo et al. [8] target privacy policies for the web and mobile applications, and propose a domain-specific language along with model transformations for specifying privacy-policy models. Similarly, Pullonen et al. [61] present a multi-level model to be used as an extension of the Business Process Model and Notation to enable the visualization, analysis, and communication of the privacy-policy characteristics of business processes. Finally, Kumar and Shyamasundar [62] explore the suitability of information-flow controls as a tool for specifying and enforcing privacy-policy requirements. These existing works address a rather small subset of the privacy-policy metadata types considered in this article. In addition, excluding [60], all of the above-mentioned papers focus on providing guidelines that are not strictly based on GDPR. In contrast, we systematically identify the requirements that, according to GDPR, must be met by privacy policies for completeness.

8.2 Completeness Checking of Privacy Policies

Sánchez et al. [63] check the compliance of privacy policies with respect to six data protection goals as stated by GDPR, including lawfulness, purpose limitation, data minimisation, accuracy, storage limitation, and integrity and confidentiality. The authors use four privacy policies to train binary classifiers for deciding whether a privacy policy is compliant with respect to each of the six goals. These goals cover only 15 out of the 56 metadata types we handle.

Nejad et al. [64] present three different models for classifying the paragraphs of privacy policies into pre-defined categories using supervised machine learning. To train their models, the authors use a dataset containing 115 privacy policies from various US companies. The authors consider 12 high-level categories for their classification. All these categories are included in our set of metadata types.

Tesfay et al. [65] propose a ML-based method for classifying the content of privacy policies across 10 categories using predefined keywords. Those categories are all covered by our metadata types, except for one category, Policy Change, which is orthogonal to our purposes.

Bhatia et al. [66] develop a semi-automated framework for extracting privacy goals from privacy policies through crowdsourcing and NLP. Similar crowdsourcing initiatives have been proposed by others as well, e.g., by Liu et al. [67] and Wilson et al. [68], where privacy policies are manually annotated in order to match their text segments against privacy issues of interest. Guerriero et al. [69] propose a framework for specifying, enforcing and checking privacy policies in data-intensive applications. Bhatia et al. [70] present a semantic frame-based representation for privacy statements that can be used to identify incompleteness in four categories of data action: collection, retention, usage, and transfer. Lippi et al. [71] present 33 metadata types for GDPR privacy policies and provide automatic support for vagueness detection based on manually crafted rules and ML classifiers built using the exact terminology of the policies as learning features.

In summary, in comparison to the above-cited works, we have a different analytical focus, namely completeness checking. In terms of the metadata types, our proposed 56 types cover all the ones identified by others, except – as noted before – one metatadata type, Policy Change [65], which is orthogonal to completeness checking. Furthermore, the existing approaches outlined above rely to a large extent on the exact phrasing of the policies to be able to extract and classify information. They do not present a thorough conceptualization of the content expected in privacy policies. The scope of application of these approaches is thus limited and, where automation is provided, the accuracy is not high enough for industrial use. In this article, we addressed the above limitations by considering a wider set of metadata types and using a combination of advanced NLP and ML for automated support.

8.3 Compliance Checking of Software

Fan et al. [72] check for the compliance of mobile health applications against GDPR. To do so, the authors propose an automated system for detecting three types of violations: incompleteness of the app privacy policy, inconsistency of data collection, and non-secure data transmission. For incompleteness checking of privacy policies, the authors define six categories of privacy-related information that need to be present in a privacy policy. They apply ML-based binary classifiers on bag-of-words representations of the sentences in a given policy to predict whether any of the categories is present in the policy. Specifically, they apply random forest (RF), decision trees (DT) and naïve Bayes (NB). Based on 10-fold cross-validation over 100 privacy policies (1,284 labelled sentences), RF performed the best in four of the six categories, and DT performed the best in the other two. The best reported precision and recall are on average 92.5% and 93.3%, respectively. In contrast with our work, the authors consider only six out of the 56 metadata types that we present in this article. Moreover, and from an ML standpoint, our solution architecture is different: we use embeddings to create the representations of the text in a given policy and apply an ensemble classification approach.

The COVID-19 pandemic has heightened privacy concerns for individuals, as seen, for example, in the analysis of app reviews for COVID-19 contact-tracing apps [73]. Hatamian et al. [74] analyze the privacy and security aspects of COVID-19 contact tracing apps. In their analysis, they consider 12 metadata types derived from different GDPR articles, including children protection, data retention and others. The authors collect the data access intentions from the permissions an Android app is given, e.g., the access to data such as call logs and contact lists. Through manual incompleteness checking of the privacy policies of 28 COVID-19 contact tracing apps, the authors assess the extent to which the policies cover the 12 GDPR principles. Subsequently, the authors check whether the apps fulfill the provisions in their respective privacy policies. Nine of the privacy principles addressed by Hatamian et al. are pertinent to privacy-policy completeness checking and are thus tackled in our work. However, we provide a more elaborate treatment of these principles (metadata types). Moreover, we devise an AI-based solution to automatically identify these metadata types in the privacy policies and thereby analyze incompleteness.

Kununka et al. [75] assess the compliance of Android and iOS apps with their privacy policies. The basis for selecting which apps to analyze is the number of third-party domains that the apps transfer sensitive data to. In total, 30 apps are selected. The authors first analyze the categories of personal data transferred to a third-party. They then manually identify metadata types in the privacy policies of these apps, focusing only on the collection, use and transfer of personal data. Finally, the authors check for the compliance of the data practices in the apps against what is stated in the policies. Compared to our automated approach, their metadata identification from both the apps and their privacy policies, as well as the compliance checking, are done entirely manually. Further, they consider only two metadata types (i,e., PD CATEGORY and its specialization SPECIAL) from what we present in this article.

9 Replicating our Methodology

Our proposed completeness checking process is not limited to privacy policies and GDPR, and can be instantiated for checking the completeness of any given document type (D) according to any given regulation (R). In our context, D represents a privacy policy, and R is GDPR. Reusing our approach can be done by replicating the same methodology as described in this paper. Specifically, one must first conduct a qualitative study over those of R's provisions that are relevant to checking the completeness of D. Such a qualitative study should aim at building a conceptual model and a set of completeness checking criteria. Subsequently, one must develop automation for completeness checking. When supervised machine learning is used for automation, one will need to (manually) create a labeled dataset covering a relatively large number of documents of type D. This will be followed by the development of classification methods, potentially alongside prediction rules and post-processing steps.

The effort required to replicate our methodology for other regulations and document types depends on several factors, including the number and complexity of the provisions that need to be considered in the qualitative study, the background and expertise in conceptual modeling and AI, the size of the evaluation data and the complexity of the classification algorithms used in the work. Based on our experience, we anticipate that 30-40% of the effort would go towards building a conceptual model and completeness criteria and the remaining 60-70% would go towards developing an automated solution.

10 Conclusion

In this paper, we proposed an AI-enabled approach for completeness checking of privacy policies according to the General Data Protection Regulation (GDPR). We first developed a conceptual model aimed at providing a thorough characterization of the content of privacy policies. Based on this conceptual model, we devised criteria describing how a privacy policy should be checked for completeness against GDPR. Second, using Natural Language Processing and Machine Learning, we developed automated support for classifying the content of privacy policies and thus identifying the metadata types necessary for checking privacy-policy completeness.

We curated a considerable number of annotated privacy policies (234 policies in total), with the majority of the annotation work performed by third-parties. We evaluated our approach on a test set of 48 privacy policies. Our metadata identification approach achieved an average precision of 92% with an average recall of 95% for identifying the manifestations of all metadata types across the test privacy policies. We ran the completeness criteria over the identified metadata. Our completeness checking approach was able to detect 300 out of 334 incompleteness issues in the real-world privacy policies we used for validation. The approach also generated 23 false positives. Our completeness checking approach thus had a precision of 93% and a recall of 90% over our test set. Compared to an intuitive automated solution based on keyword search, our AI-based approach leads to a significant improvement in precision and recall of 27% and 5% for metadata identification and of 24.5% and 38% in completeness checking, respectively.

In the future, we plan to enhance our completeness criteria so that they consider not only the presence/absence of metadata but also the meaning of the sentences containing the metadata. Another important direction for future work is to go beyond our current case-study domain (funds) in order to assess the generalizability of our approach.

Acknowledgments

This paper was supported by Linklaters, Luxembourg's National Research Fund (FNR) under grant BRIDGES/19/IS/13759068/ARTAGO, and NSERC of Canada under the Discovery, Discovery Accelerator and CRC programs.

References

[1] European Union, "General data protection regulation," *Official Journal of the European Union*, 2018.
[2] EU-GDPR. (2019) EU GDPR portal. [Online]. Available: https://eugdpr.org
[3] C. Tankard, "What the GDPR means for businesses," *Network Security*, vol. 6, 2016.
[4] C. Perera, M. Barhamgi, A. K. Bandara, M. Ajmal, B. A. Price, and B. Nuseibeh, "Designing privacy-aware internet of things applications," *Inf. Sci.*, vol. 512, no. 1, 2020.
[5] D. Torre, G. Soltana, M. Sabetzadeh, L. C. Briand, Y. Auffinger, and P. Goes, "Using models to enable compliance checking against the GDPR: an experience report," in *22nd ACM/IEEE International Conference on Model Driven Engineering Languages and Systems, MODELS 2019, Munich, Germany, September 15-20, 2019*, 2019.
[6] D. Torre, M. Alférez, G. Soltana, M. Sabetzadeh, and L. C. Briand, "Model driven engineering for data protection and privacy: Application and experience with GDPR," *CoRR*, vol. abs/2007.12046, 2020.
[7] V. Ayala-Rivera and L. Pasquale, "The grace period has ended: An approach to operationalize GDPR requirements," in *Proceedings of 31st IEEE International Conference on Requirements Engineering (RE'18)*, 2018.
[8] J. Caramujo, A. Rodrigues da Silva, S. Monfared, A. Ribeiro, P. Calado, and T. Breaux, "RSL-IL4Privacy: A domain-specific language for the rigorous specification of privacy policies," *Requirements Engineering*, vol. 24, no. 1, 2019.
[9] J. Bhatia and T. D. Breaux, "Semantic incompleteness in privacy policy goals," in *26th IEEE International Requirements Engineering Conference, RE 2018, Banff, AB, Canada, August 20-24, 2018*, 2018.
[10] R. Slavin, X. Wang, M. B. Hosseini, J. Hester, R. Krishnan, J. Bhatia, T. D. Breaux, and J. Niu, "Toward a framework for detecting privacy policy violations in android application code," in *Proceedings of the 38th International Conference on Software Engineering, ICSE 2016, Austin, TX, USA, May 14-22, 2016*, 2016.
[11] M. Fan, L. Yu, S. Chen, H. Zhou, X. Luo, S. Li, Y. Liu, J. Liu, and T. Liu, "An empirical evaluation of gdpr compliance violations in android mhealth apps," in *2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE)*, 2020.
[12] D. Torre, S. Abualhaija, M. Sabetzadeh, L. C. Briand, K. Baetens, P. Goes, and S. Forastier, "An ai-assisted approach for checking the completeness of privacy policies against GDPR," in *28th IEEE International Requirements Engineering Conference, RE 2020, Zurich, Switzerland, August 31 - September 4, 2020*, 2020.
[13] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, 2015.
[14] D. Jurafsky and J. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, 2nd ed. Prentice Hall, 2009.
[15] M. Rodrigues, A. Teixeira *et al.*, *Advanced applications of natural language processing for performing information extraction*. Springer, 2015.
[16] A. Mikheev, M. Moens, and C. Grover, "Named entity recognition without gazetteers," in *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, 1999.

[17] J. E. Friedl, *Mastering regular expressions*. " O'Reilly Media, Inc.", 2006.

[18] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science*, vol. 349, no. 6245, 2015.

[19] C. C. Aggarwal, *Machine Learning for Text*, 1st ed. Springer Publishing Company, Incorporated, 2018.

[20] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Morgan Kaufmann, 2016.

[21] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge University Press Cambridge, 2008.

[22] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[23] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[24] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2014.

[25] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[26] A. Joshi, V. Tripathi, K. Patel, P. Bhattacharyya, and M. Carman, "Are word embedding-based features useful for sarcasm detection?" in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016.

[27] J. Mu, S. Bhat, and P. Viswanath, "All-but-the-top: Simple and effective postprocessing for word representations," *ArXiv*, vol. abs/1702.01417, 2018.

[28] L.-C. Yu, J. Wang, K. R. Lai, and X. Zhang, "Refining word embeddings for sentiment analysis," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017.

[29] D. Ghosh, W. Guo, and S. Muresan, "Sarcastic or not: Word embeddings to predict the literal or sarcastic meaning of words," in *proceedings of the 2015 conference on empirical methods in natural language processing*, 2015.

[30] M. Iyyer, V. Manjunatha, J. Boyd-Graber, and H. Daumé III, "Deep unordered composition rivals syntactic methods for text classification," in *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*, 2015.

[31] Y. Qi, D. Sachan, M. Felix, S. Padmanabhan, and G. Neubig, "When and why are pre-trained word embeddings useful for neural machine translation?" in *NAACL*, 2018.

[32] Y. Kim, "Convolutional neural networks for sentence classification." 2014.

[33] D. Chen and C. D. Manning, "A fast and accurate dependency parser using neural networks," in *EMNLP*, 2014.

[34] J. Turian, L. Ratinov, and Y. Bengio, "Word representations: a simple and general method for semi-supervised learning," in *Proceedings of the 48th annual meeting of the association for computational linguistics*, 2010.

[35] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[36] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding with unsupervised learning," *Technical report, OpenAI*, 2018.

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

[38] K. Ethayarajh, "How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings," *arXiv preprint arXiv:1909.00512*, 2019.

[39] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, 2017.

[40] E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng, "Improving word representations via global context and multiple word prototypes," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2012.

[41] W. Blacoe and M. Lapata, "A comparison of vector-based representations for semantic composition," in *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, 2012.

[42] X. Zhu, T. Li, and G. De Melo, "Exploring semantic properties of sentence embeddings," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018.

[43] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Towards universal paraphrastic sentence embeddings," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[44] O. Amaral, D. Torre, S. Abualhaija, M. Sabetzadeh, and L. C. Briand, *Glossary and Completeness Criteria traceability to the GDPR articles*, available at https://tinyurl.com/t3h8e75z, May 2021.

[45] J. Saldana, *The Coding Manual for Qualitative Researchers*. SAGE Publishing, 2016.

[46] European Union, "Article 29 working party - guidelines on data protection officers (dpos)," *Justice and Consumers*, 2018.

[47] G. Soltana, N. Sannier, M. Sabetzadeh, and L. C. Briand, "Model-based simulation of legal policies: Framework, tool support, and validation," *Software & Systems Modeling*, vol. 17, no. 3, 2018.

[48] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, 1997.

[49] L. Michaelis, "Word meaning, sentence meaning, and syntactic meaning," *Cognitive approaches to lexical semantics*, vol. 23, 2003.

[50] S. Wang and C. D. Manning, "Baselines and bigrams: Simple, good sentiment and topic classification," in *Proceedings of the 50th annual meeting of the association for computational linguistics: Short papers-volume 2*, 2012.

[51] A. Komninos and S. Manandhar, "Dependency based embeddings for sentence classification tasks," in *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, 2016.

[52] R. Eckart de Castilho and I. Gurevych, "A broad-coverage collection of portable NLP components for building shareable analysis pipelines," in *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT'14)*, 2014.

[53] E. D. D. Team, "Deeplearning4j: Open-source distributed deep learning for the jvm, apache software foundation license 2.0," 2020, last accessed: January 2020. [Online]. Available: http://deeplearning4j.org

[54] J. H. Hayes, W. Li, and M. Rahimi, "Weka meets tracelab: Toward convenient classification: Machine learning for requirements engineering problems: A position paper," in *Artificial Intelligence for Requirements Engineering (AIRE), 2014 IEEE 1st International Workshop on*, 2014.

[55] F. Eibe, M. Hall, and I. Witten, "The weka workbench. online appendix for data mining: Practical machine learning tools and techniques," *Morgan Kaufmann*, 2016.

[56] C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, 2008.

[57] ALFI. (2019) Association of the luxembourg fund industry - 946 member funds. Last accessed: March 2019. [Online]. Available: https://www.alfi.lu/Alfi/media/Members/Member\%20Company\%20Directory/Membres-ALFI-Fonds-par-nom.pdf

[58] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement (EPM)*, vol. 20, no. 1, pp. 37–46, 1960.

[59] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica (BM)*, vol. 22, no. 3, pp. 276–282, 2012.

[60] E. Vanezi, G. M. Kapitsaki, D. Kouzapas, A. Philippou, and G. A. Papadopoulos, "Diálogop - A language and a graphical tool for formally defining GDPR purposes," in *Research Challenges in Information Science - 14th International Conference, RCIS 2020, Limassol, Cyprus, September 23-25, 2020, Proceedings*, 2020.

[61] P. Pullonen, J. Tom, R. Matulevicius, and A. Toots, "Privacy-enhanced BPMN: Enabling data privacy analysis in business processes models," *Software & Systems Modeling*, vol. 18, no. 6, 2019.

[62] N. V. N. Kumar and R. K. Shyamasundar, "Realizing purpose-based privacy policies succinctly via information-flow labels," in *2014 IEEE Fourth International Conference on Big Data and Cloud Computing, BDCloud 2014, Sydney, Australia, December 3-5, 2014*, 2014.

[63] D. Sánchez, A. Viejo, and M. Batet, "Automatic assessment of privacy policies under the gdpr," *Applied Sciences*, vol. 11, no. 4, 2021.

[64] N. Mousavi Nejad, P. Jabat, R. Nedelchev, S. Scerri, and D. Graux, "Establishing a strong baseline for privacy policy classification," in *ICT Systems Security and Privacy Protection*, M. Hölbl, K. Rannenberg, and T. Welzer, Eds., 2020.

[65] W. B. Tesfay, P. Hofmann, T. Nakamura, S. Kiyomoto, and J. Serna, "Privacyguide: Towards an implementation of the EU GDPR on internet privacy policy evaluation," in *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics, IWSPA@CODASPY 2018, Tempe, AZ, USA, March 19-21, 2018*, 2018.

[66] J. Bhatia, T. D. Breaux, and F. Schaub, "Mining privacy goals from privacy policies using hybridized task recomposition," *ACM Trans. Softw. Eng. Methodol.*, vol. 25, no. 3, 2016.

[67] F. Liu, R. Ramanath, N. M. Sadeh, and N. A. Smith, "A step towards usable privacy policy: Automatic alignment of privacy statements," in *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, 2014.

[68] S. Wilson, F. Schaub, R. Ramanath, N. M. Sadeh, F. Liu, N. A. Smith, and F. Liu, "Crowdsourcing annotations for websites' privacy policies: Can it really work?" in *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, 2016.

[69] M. Guerriero, D. A. Tamburri, and E. D. Nitto, "Defining, enforcing and checking privacy policies in data-intensive applications," in *Proceedings of the 13th International Conference on Software Engineering for Adaptive and Self-Managing Systems, SEAMS@ICSE 2018, Gothenburg, Sweden, May 28-29, 2018*, 2018.

[70] J. Bhatia, M. C. Evans, and T. D. Breaux, "Identifying incompleteness in privacy policy goals using semantic frames," *Requir. Eng.*, vol. 24, no. 3, 2019.

[71] M. Lippi, P. Pałka, G. Contissa, F. Lagioia, H.-W. Micklitz, G. Sartor, and P. Torroni, "Claudette: an automated detector of potentially unfair clauses in online terms of service," *Artificial Intelligence and Law*, vol. 27, no. 2, 2019.

[72] M. Fan, L. Yu, S. Chen, H. Zhou, X. Luo, S. Li, Y. Liu, J. Liu, and T. Liu, "An empirical evaluation of GDPR compliance violations in android mhealth apps," *CoRR*, vol. abs/2008.05864, 2020.

[73] M. Bano, D. Zowghi, and C. Arora, "Requirements, politics, or individualism: What drives the success of covid-19 contact-tracing apps?" *IEEE Software*, vol. 38, no. 1, pp. 7–12, 2021.

[74] M. Hatamian, S. Wairimu, N. Momen, and L. Fritsch, "A privacy and security analysis of early-deployed COVID-19 contact tracing android apps," *Empir. Softw. Eng.*, vol. 26, no. 3, 2021.

[75] S. Kununka, N. Mehandjiev, and P. Sampaio, "A comparative study of android and ios mobile applications' data handling practices versus compliance to privacy policy," in *Privacy and Identity Management. The Smart Revolution - 12th IFIP WG 9.2, 9.5, 9.6/11.7, 11.6/SIG 9.2.2 International Summer School, Ispra, Italy, September 4-8, 2017, Revised Selected Papers*, 2017.

Orlando Amaral Cejas obtained his B.Sc. in Automation Engineering (2012) and his M.Sc. in Digital Systems (2017) at the Technological University of Havana (Cuba). During that time, he participated in the adaptation and integration of the HT2000-B Chinese traffic light controller to the Cuban intelligent transport system. Currently, he is working as a doctoral researcher at SnT centre for Security, Reliability and Trust at the University of Luxembourg. He is doing his PhD as part of an industrial project on compliance checking of legal text according to the GDPR. His main research interests include: regulatory compliance, model-driven engineering, applied natural language processing, text mining and machine learning.

Sallam Abualhaija is a research scientist at SnT centre for Security, Reliability, and Trust, University of Luxembourg. She did her PhD in Computer Science (2016) at Hamburg University of Technology (Germany) on solving Word Sense Disambiguation using bio-inspired optimisation methods. She is currently leading two industrial research projects in close engagement with industry partners form diverse sectors. Her main research interests combine AI with software engineering, with a focus on: natural language processing, information extraction, text mining, machine learning, requirements engineering, legal and regulatory compliance.

Damiano Torre is an Associate Research Scientist in the Department of Computer Information Systems, Texas A&M University Central Texas, United States. He is involved in a research projects with U.S. agencies. His research interests are focused on computer science, and more specifically on software engineering, cybersecurity, artificial intelligence, model-driven engineering, and empirical software engineering. Prior to coming to the United States, he was a research associate at the University of Luxembourg from 2018 to 2021. He received his B.Sc. from the University of Bari (Italy), M.Sc. from the University of Castilla-La Mancha (Spain) and Ph.D. from Carleton University (Canada) in 2009, 2011 and 2018, respectively. He regularly serves on the organizing / program committees of ISSRE and QRS, and satellite events of EMSE, ICSE, and ASE. He is an IEEE member.

Mehrdad Sabetzadeh is an Associate Professor at the School of Electrical Engineering and Computer Science of the University of Ottawa and a part-time Faculty Member at the Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg. Previously, Sabetzadeh worked as a permanent member of the research staff at Simula Research Laboratory (Norway), and as an NSERC postdoctoral fellow at University College London (UK). Sabetzadeh received his Ph.D. in Computer Science from the University of Toronto. His main research interests are in software engineering with an emphasis on requirements engineering, model-based development, and regulatory compliance. Sabetzadeh is passionate about fostering stronger ties between academia and industry; in the past decade, he has conducted most of his research in close collaboration with industry partners. His experience spans several sectors, including government, finance, legal services, telecommunications, maritime, energy, aerospace, railways, and automotive. Sabetzadeh has co-authored more than 70 scientific papers and secured more than $6M of research funding as lead investigator. He regularly serves on the organizing / program committees of several international conferences such as RE, ICSE, ESEC/FSE, and MODELS.

Lionel C. Briand is professor of software engineering and has shared appointments between (1) School of Electrical Engineering and Computer Science, University of Ottawa, Canada and (2) The SnT centre for Security, Reliability, and Trust, University of Luxembourg. He is the head of the SVV department at the SnT Centre and a Canada Research Chair in Intelligent Software Dependability and Compliance (Tier 1).

He holds an ERC Advanced Grant, the most prestigious European individual research award, and has conducted applied research in collaboration with industry for more than 25 years, including projects in the automotive, aerospace, manufacturing, financial, and energy domains. He was elevated to the grades of IEEE and ACM fellow, granted the IEEE Computer Society Harlan Mills award (2012) and the IEEE Reliability Society Engineer-of-the-year award (2013) for his work on software verification and testing. His research interests include: Testing and verification, search-based software engineering, model-driven development, requirements engineering, and empirical software engineering.