A Combined Rule-Based and Machine Learning Approach for Automated GDPR Compliance Checking

Rajaa EL HAMDANI
HEC Paris
France

Majd Mustapha
EURA NOVA
Belgium

David Restrepo Amariles
HEC Paris
France

Aurore Troussel
Steptoe & Johnson LLP
Belgium

Sébastien Meeùs
HEC Paris
France

Katsiaryna Krasnashchok
EURA NOVA
Belgium

ABSTRACT

The General Data Protection Regulation (GDPR) requires data controllers to implement end-to-end compliance. Controllers must therefore ensure that the terms agreed with the data subject and their own obligations under GDPR are respected in the data flows from data subject to controllers, processors and sub processors (i.e. data supply chain). This paper seeks to contribute to bridge both ends of compliance checking through a two-pronged study. First, we conceptualize a framework to implement a document-centric approach to compliance checking in the data supply chain. Second, we develop specific methods to automate compliance checking of privacy policies. We test a two-modules system, where the first module relies on NLP to extract data practices from privacy policies. The second module encodes GDPR rules to check the presence of mandatory information. The results show that the text-to-text approach outperforms local classifiers and enables the extraction of both coarse-grained and fine-grained information with only one model. We implement full evaluation of our system on a dataset of 30 privacy policies annotated by legal experts. We conclude that this approach could be generalized to other documents in the data supply as a means to improve end-to-end compliance.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Computing methodologies** → **Information extraction**; **Machine learning**; • **Social and professional topics** → **Privacy policies**; • **Security and privacy** → *Usability in security and privacy*.

ACM Reference Format:
Rajaa EL HAMDANI, Majd Mustapha, David Restrepo Amariles, Aurore Troussel, Sébastien Meeùs, and Katsiaryna Krasnashchok. 2021. A Combined Rule-Based and Machine Learning Approach for Automated GDPR Compliance Checking. In *Eighteenth International Conference for Artificial Intelligence and Law (ICAIL'21), June 21–25, 2021, São Paulo, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3462757.3466081

ICAIL'21, June 21–25, 2021, São Paulo, Brazil

ACM ISBN 978-1-4503-8526-8/21/06…$15.00
https://doi.org/10.1145/3462757.3466081

1 INTRODUCTION

It has become widely acknowledged that complying with data protection laws, particularly the European General Data Protection Regulation (GDPR), is the most difficult compliance challenge organizations face today across industries [1]. Moreover, technology became the most important compliance cost for organizations, as they turned to specialized technologies to carry out compliance tasks such as document review, regulatory checks and operational audits [1]. Despite the increasing interest in technology as a compliance tool in data protection, there is not yet a comprehensive conceptual framework characterizing the tasks required to verify compliance in the entire data supply chain (i.e., end-to-end compliance), and how current methods can contribute to address them.

This paper intends to contribute to both of these issues. First, we lay down a framework to implement and monitor GDPR compliance in the data supply chain through a document-centric approach. We define three key tasks of compliance checking based on the function and content of the document: (1) document to regulation, (2) document to document, and (3) document to operations. We apply this framework to analyze the compliance function of privacy policies in the data supply chain and define the tasks required to determine if a privacy policy is compliant with the GDPR.

Second, we develop and test several methods to verify compliance of privacy policies to the GDPR by leveraging the advantages of both machine learning and rule-based approaches. In particular, we build a two-modules system to verify the completeness of privacy policies with regards to mandatory information. The first module automatically extracts coarse-grained and fine-grained data practices, while the second module analyzes extracted data practices and checks the presence of mandatory information according to the provisions of the GDPR. We make use of the OPP-115 dataset [48] for training and evaluation of our models. We treat the extraction of data practices as a Hierarchical Multi-label Classification (HMTC) task and experiment with two different approaches: local classifiers and text-to-text. Our proposed text-to-text method has several advantages over local classifiers, including extraction of additional information and better scalability.

Our contributions are the following:

- we present a theoretical framework to implement automated GDPR compliance in the data supply chain;
- we provide formal and substantive approach to verify compliance of privacy policies with the GDPR;

we develop a combined rule-based and machine learning approach to experiment with automated formal compliance checking of privacy policies with the GDPR;
we propose a text-to-text approach for HMTC using transfer learning and multi-task learning ;
additionally, we extract the span of text from a privacy policy corresponding to the ne-grained data practices, which provides better explainability of the classi cation results;
we further annotate a dataset of 30 privacy policies with the presence of mandatory information, in order to fully evaluate our two-modules system of compliance checking;
nally, we release a public repository of the dataset, implementations and the ne-tuned T5-11B model on OPP-115[1]

This paper is organized as follows. Section 2 provides an overview of the related work. Section 3 lays down a general compliance framework to automate GDPR compliance and applies it to privacy policies. Section 4 describes the development of data practices extraction algorithms and its evaluation results. Section 5 presents the rule-based system, which veri es the presence of mandatory information, and its evaluation on an annotated dataset.

2 RELATED WORK

The Information Commissioner's O ce (ICO), the Data Protection Authority of the United Kingdom, observes that overall compliance with the GDPR is jeopardized by the "opaque nature" of the data supply chain, which is poorly documented, and generally fails to comply with the GDPR's accountability principle [21]. The ICO further points out that controllers and processors should not limit compliance to entering into contract and producing legal documents. They must monitor processing activities, and conduct audit to ensure that appropriate technical and organizational measures are in place throughout the data supply chain [2].

Most research in AI and law studying GDPR compliance has focused on the relations between data subject and data controllers, rather than on the compliance challenges in the data supply chain [3, 6, 11, 54]. More recently, the BPR4GDPR project started working on a compliance ontology speci cation that supports end-to-end compliance [37] and which can contribute to address some of the operational challenges raised by the ICO. A compliance framework is provided in [40] for speci c documents in the supply chain such as the data protection impact assessment (DPIA).

2.1 Model-based Compliance Checking

The AI and law research community has developed model-based methods for automated compliance checking, such as legal ontologies that support legal reasoning via logic programming [9, 11, 12, 13, 15, 16, 19, 30, 44, 45]. However, even though most legal rules can be described in logic programming, these methods face two main challenges when applied to real-life cases:

Knowledge acquisition bottleneck [17]: Logic programming requires the encoding of facts into predicate form, but such an encoding would be very cumbersome for each data protection documents in the data supply chain.

Open texture of legal language [46]: Privacy concepts can be quite abstract, and their evaluation arduous, as it is impossible to de ne a nite set of rules for all possible applications. For example, the storage limitation principle states that personal data must not be kept "longer than is necessary for the purposes for which the personal data are processed."(Art. 5.1(e)). This principle does not de ne speci c time limits that can be evaluated easily.

To solve the aforementioned problems, we employ natural language processing (NLP) techniques to automatically extract information from the data supply chain documents. Likewise, authors of [44] suggest using NLP to extract information from legal documents according to their UML model of the GDPR, to automatically construct their model-based representation. A large and growing body of literature uses machine learning and NLP algorithms to extract privacy information from legal documents, however, it only considers privacy policies and does not yet deal with other documents of the data supply chain, such as DPA and DPIA.

2.2 Information extraction from privacy policies

Privacy policies are long documents that are di cult to read for data subjects. Empirical studies have been conducted to study and measure the ambiguity and vagueness of privacy policies [5, 20, 24, 36]. E orts have been made to decrease these de ciencies [42, 51], by using classi cation techniques to extract data practices from the text and represent them in a user-friendly interface. Other approaches focus on the extraction of one speci c type of data practice, such as "opt-out choice" in [38].

More and more methods are developed to analyze the compliance of privacy policies with the GDPR. A variety of NLP tools such as word embeddings are used in [7, 43] to verify the completeness of privacy policies according to the rules set out by the GDPR. The CLAUDETTE project [6] extracts clauses that are problematic with respect to the GDPR. In a di erent project [23], privacy policies were analyzed in a large-scale setting to study the e ect of the GDPR on their provisions. For instance, the comparison between pre-GDPR and post-GDPR versions of 6,278 English privacy policies showed that the GDPR caused textual changes of the privacy policies, such that their appearance improved, their length increased, and that they cover more data practice categories.

A great deal of previous research on privacy policies relies on supervised machine learning methods, which require datasets of annotated privacy policies. However, there are very few such publicly released datasets. In our work, we use OPP-115 [45] [a corpus of 115 privacy policies annotated with both coarse-grained and ne-grained data practices. Several works used OPP-115 to train machine learning algorithms on the task of extracting data practices [14, 25, 28, 29, 32]. PrivacyQA [35] is another publicly available dataset of 1,750 questions about the privacy policies of mobile applications, which is used to train question-answering algorithms. A serious limitation of the publicly available datasets is that their annotation schemes contain few concepts aligned with the GDPR. A connection between the data practices identi ed in OPP-115 and the GDPR was presented in [31], which revealed that the principle of accountability of Article 5 is absent from OPP-115 concepts.

[1] https://github.com/smartlawhub/Automated-GDPR-Compliance-Checking

2.3 Transfer learning in NLP

Until very recently, in computer science research, general NLP tasks such as text classification were commonly handled with architectures based on word embeddings [37], Convolutional Neural Networks (CNN) [21] and Recurrent Neural Networks (RNN) [18]. RNN-based solutions, which achieved state-of-the-art results at the time, are, however, limited when it comes to dealing with long text, due to their sequential nature. Moreover, this nature stands in the way of most parallel training methods, therefore limiting the ability to decrease training time. Encoder-decoder architectures commonly used for sequence to sequence problems have inspired the attention-based transformers [47], and these architectures managed to go beyond the limitations of sequential networks.

The use of transformers is gradually becoming the state of the art of NLP. Indeed, they are effective in the variety of tasks, including masked token prediction, next sentence prediction, question answering, machine translation, summarization, sentiment analysis and classification [10, 33]. The fact that transformers are trained in an unsupervised way, which reduces the reliance on labeled data and allows the use of a larger pool of text, explains the increased performance. Transformers are very effective in transfer learning, allowing researchers to pretrain them with large amounts of general-purpose texts and then to fine-tune them for their specific tasks with good results, less effort, and less labeled data. Several works explored the potential of fine-tuning transformers on OPP-115 to extract coarse-grained data practices [28, 29].

Similarly to [14], we extract both coarse-grained and fine-grained data practices from privacy policies. And similarly to [28, 29], we use variations of the transformer architecture as base for fine-tuning on OPP-115. Furthermore, our best-performing model is able to extract the fine-grained data practices with corresponding spans of text from the policy, thus improving explainability of the results.

3 A FRAMEWORK FOR COMPLIANCE CHECKING IN THE DATA SUPPLY CHAIN

New data protection and privacy regulations around the world have empowered data subjects vis-à-vis data controllers (EU fundamental rights, California Privacy laws, Canada PIPEDA, Brazil, etc.). As mentioned earlier, the data supply chain (i.e., the way data flows from data subject to controllers, processors and sub processors) is opaque and remains hard to monitor, ultimately hindering the effectiveness of data-subject rights.

At the operation level, the data supply chain is characterized by the high volume of data flows across processors and jurisdictions for tasks as varied as storage, pre-processing, producing analytics, implementing AI-methods, generating visualizations, etc. In addition to regulations, these flows are regulated by fragmented legal artifacts such as contracts, data protection addendum (DPA), technical and operational measures (TOMs), etc., which are often of restricted access, lack interoperability and have low operational efficiency, ultimately obstructing compliance with the GDPR.

To help data controllers use technological tools to improve compliance with the GDPR, we draw a general conceptual framework to automate compliance checking in the data supply chain based on a document-centric approach and lay down the tasks required to achieve end-to-end compliance.

3.1 Compliance Checking Framework: A Document-centric Approach

The starting point of our compliance framework is the set of documents establishing the conditions and processing activities that data controllers and processors are intended to and effectively carry out (hereafter compliance documents), such as privacy policies, contracts, DPAs, TOMs, DPIAs, etc. This is a reasonable starting point as the GDPR establishes a document-centric approach to compliance, entrusting documents with different functions in relation to the obligations of data controllers and processors. First, it requires data controllers to properly demonstrate compliance with the Regulation ("principle of accountability", Art. 5.2). Second, data controllers shall inform data subjects about the processing activities they carry out (Art. 12, 13, and 14) and document instructions they give to data processors (Art. 29). Third, sub-processors must obtain a written authorization from controllers to further subcontract a processing activity, and must also document their instructions (Art. 28.2 and 28.3). Lastly, controllers and processors must maintain a record of processing activities (ROPA) (Art. 30).

Hence, we define GDPR compliance checking as the assessment of the provisions of a compliance document in relation to: (1) a regulation, (2) another document in the supply chain, and (3) an operation. These dimensions define the three key tasks of compliance checking in the data supply chain under the GDPR.

(1) Document to Regulation Compliance: The provisions of a compliance document are assessed against the regulation. For example, according to the GDPR, privacy policies must provide certain mandatory information to data subjects (Art. 12, 13 and 14). We further divide this compliance task into two sub-tasks:
   - Formal Compliance Checking: whether the document fulfills its informative function, i.e., is the mandatory information included in the document. For example, this sub-task could consist of the verification of the existence of information about data retention (Art. 13.2(a)).
   - Substantive Compliance Checking: whether the document fulfills its accountability function, i.e., does the information provided comply with the GDPR. For example, this sub-task could consist of the verification that the data retention period is lawful, i.e., it does not exceed the necessary time as required by Article 5.1(e).

(2) Document Chain Compliance: The supply chain documents are assessed against the main contractual standards or against documents with a higher hierarchy in the chain. Using the previous example on data retention, two assessments could be done on (i) the contract between the data controllers and the data processors to verify if the agreed data retention period corresponds to the one provided in the privacy policy, and on (ii) the ROPAs of both the data controllers and processors that contain the effective date of erasure (Art. 30.1).

(3) Operational Compliance: This task consists of assessing the adequacy between operations, documents, and regulations. For the previous example, this task would imply verifying if the data has been effectively deleted from the servers of data processors and controllers when the retention period ends.

3.2 The compliance of privacy policies

Privacy policies are the compliance documents that appear at the top of the data supply chain. Hence, we rst apply our framework to analyze privacy policies' compliance in the data supply chain. This paper seeks to de ne the tasks required to determine a privacy policy's formal and substantive compliance with the GDPR.

The GDPR does not explicitly mention privacy policies, but data controllers widely use them. Their main function is to provide mandatory information to data subjects according to Articles 13 and 14. The absence of any part of this information renders the privacy policy non-compliant. Moreover, the GDPR requires privacy policies to use plain and clear language so individuals can understand how their personal data are processed, provide their consent and exercise their rights. Consequently, the tasks to verify the formal compliance of privacy policies are:

> Check the presence of each mandatory information according to Articles 13 and 14.
> Check the readability and clarity of the language used in the privacy policy.

The substantive compliance checking of privacy policies consist of verifying that the data processing complies with the data protection rules (e.g, fair and transparent processing of Art. 5, and lawfulness of the processing of Art. 6). For example, a privacy policy must specify the legal basis of the processing according to Article 13.1(c), but it must also demonstrate that this legal basis complies with Article 6 requirements on the lawfulness of the processing (e.g., if the legal basis is consent, the controller must ensure that consent has been given for one or more speci c purposes.).

In this study, we automate the rst task of formal compliance checking of privacy policies. In the following sections, we describe how we combined rules and machine learning to check the presence of mandatory information in privacy policies automatically. The end-users of such a system would be lawyers or data protection o cers who review large numbers of privacy policies to check their compliance with the GDPR. Another type of end-user would be project managers in small companies who lack the legal knowledge to ensure privacy policies' compliance.

4 EXTRACTION OF DATA PRACTICES FROM PRIVACY POLICIES

Ensuring both the compliance of documents and data processing activities is becoming more burdensome to companies due to several challenges. We focus on the challenge posed by the large number of documents that data protection o cers need to review to guarantee compliance. We suggest using natural language processing technologies to assist data protection o cers in performing the compliance checking tasks. NLP could help to extract compliance information from unstructured compliance documents and save it into structured formats such as XML or RDF to unlock use cases such as automated compliance checking.

In this paper, we describe our experiment in automating formal compliance of privacy policies. We rst train a machine learning algorithm to extract from privacy policies information describing the company's data practices. We then use the extracted information as input to a rule-based system that encode Articles 13 and 14.

4.1 OPP-115 : Training Dataset of Online Privacy Policies

For our task we make use of the Usable Privacy Policy Project's Online Privacy Policies (OPP-115) corpus, introduced by [48], which contains detailed annotations made by Subject Matter Experts (SMEs) for the data practices described in a set of 115 website privacy policies.

At a high level, annotations fall into one of ten data practice categories:

(1) 1BCParty Collection/Use: What, why and how information is collected by the service provider.
(2) 3A3 Party Sharing/Collection: What, why and how information shared with or collected by third parties.
(3) User Choice/Control: Control options available to users.
(4) User Access, Edit, & Deletion: If/how users can access, edit or delete information.
(5) Data Retention: How long user information will be stored.
(6) Data Security: Protection measures for user information.
(7) Policy Change: Informing users if policy information has been changed.
(8) Do Not Track: If and how DNT signals for online tracking and advertising are honored.
(9) International & Speci c Audiences: Practices pertaining to a speci c group of users.
(10) Other: General text, contact information or practices not covered by the other categories.

According to the dataset creators, the best agreement between SMEs was achieved on Do Not Track class with Fleiss' Kappa equal to 91%, whereas the most controversial class was Other, with only 49% of agreement [48]. We further decompose the latter category into its attributes Introductory/Generic , Privacy Contact Information and Practice Not Covered resulting in 12 categories.

Figure 1 depicts a fragment of OPP-115 taxonomy: for each class (grey shaded blocks), a set of lower-level privacy attributes is assigned (20 in total, dark blue shaded blocks), with speci c values corresponding to each attribute. For example, the attribute Personal Information Type designates the di erent types of personal information mentioned in the text, as can be seen from the annotations in Figure 2 from the IMDb policy[2], annotated with 1BCParty Collection/Use category.

OPP-115 comprises 3,792 segments, each segment labeled with one or more classes out of 12. The SMEs produced a total of 23K annotations of categories. In aggregate, these categories were associated with 128K values for attributes and 103K selected spans of policy text. To the extent of our knowledge, this is the rst e ort to leverage these spans to extract information from privacy policies.

We split the OPP-115 dataset on a policy-document level into 3 sets: 65 policies are used for training, 35 for validation and 30 policies are kept as a testing set.

4.2 Problem formulation

The taxonomy of data practices is organized in a class hierarchy that we model as a Directed Acyclic Graph (DAG) shown in Figure 3.

[2]To retrieve the exact source used:<https://web.archive.org/web/20200526092253if_/https://www.imdb.com/privacy#auto>(Automatic Information sub-section.)

Figure 1: The privacy taxonomy of [48]. The top level of the hierarchy (grey shaded blocks) de nes coarse-grained data practices or privacy categories. The lower level de nes a set of privacy attributes (blue shaded blocks), each assuming a set of values. We show a subset of the taxonomy for clarity and space considerations.

Figure 2: Annotated excerpt from IMDb privacy notice

We choose a DAG structure instead of a tree structure because some attributes are associated with more than one category, i.e., have more than one parent. For example, the attribute Personal information type belongs to both 1st Party Collection/Use and 3rd Party Collection/Use categories.

The training dataset is a corpus of N privacy policies: $\mathcal{D} = \{d_1, \ldots, d_N\}$. Each privacy policy d_i is a set of annotated segments: $d_i = \{(s_{i1}, y_{i1}), \ldots, (s_{iM}, y_{iM})\}$ where $y_{ij} \in Y = \{c_1, \ldots, c_L\}$ such that c_k is a path in the DAG that starts from category 1 and ends at leaf node L.

Figure 3: The DAG structure of the OPP-115 taxonomy

We treat predicting the categories of data practices and the values of each attribute as an HMTC task. There are three methods to solve hierarchical text classi cation tasks: at, local, or global methods [39]. The at method behaves like traditional classi cation algorithms by ignoring the labels' hierarchy and predicting only classes at the leaf nodes. Local methods take into account the hierarchy by training independent local classi ers. Global methods train a single classi er for all classes. In this paper we conduct two experiments, where we rst build a local multi-label classi er, and then cast the HMTC task to two text-to-text tasks.

4.3 Local classi ers approach

This approach is inspired by Polisis [14], where authors build a local multi-label classi er for the higher level categories, and one local multi-label classi er per attribute to predict their values. Predictions are made in a top-down order: once the categories of a segment are inferred, the second step predicts the values of attributes children of the predicted categories. For example, if the rst-level classi er predicts the Data Retention and Data Security categories, only the local classi ers corresponding to the attributes Retention Period , Retention Purpose , Personal Information Type , Security Measure are chosen in the second step.

Authors of [14] are using the same base classi er for all the multi-label classi ers. In this paper we reproduce their work by using CNN as the base classi er. We use the same architecture of CNN and hyperparameters. The CNN classi er is composed of one convolutional layer with a ReLU activation, followed by a dense layer and a ReLU activation. The last layer is a dense layer with a sigmoid activation. We tokenize segments using PENN Treebank tokenization in NLTK [41]. Tokens are mapped into a 300-dimensional space via an embedding Layer. We used FastText to train Word embeddings on 130,326 privacy policies [54].

Recently, the state-of-the-art results have been achieved by transformers. We reproduce the framework of Polisis, using XLNet [49] instead of CNN as a base classi er. XLNet is a transformer language model, which extends Transformer-XL [13]. It is an auto-regressive

language model, pretrained on all the permutations of the input sequence. We ne-tune the XLNet[3] on 21 tasks one task for predicting the categories, and the rest for each attribute's values.

4.4 Text-to-Text approach

In this section we explain how we use T5 [33] to solve HMTC. T5 is a pretrained language model based on the transformer architecture. T5 has two main di erences in comparison to XLNet. First, it is pretrained on a multi-task mixture of unsupervised and supervised tasks. Second, each task is converted into a text-to-text format. We adopt T5 both for its top results on NLP benchmarks and for its text-to-text nature.

The local classi ers approach has two main drawbacks. First, it trains the set of local classi ers independently. Second, the number of local classi ers grows linearly with the size of the label hierarchy. These limitations motivate this second approach, where we convert the HMTC task into two text-to-text tasks one for each level of the label hierarchy to better capture the dependencies of labels belonging to the same level. Moreover, by training one unique algorithm for each level, we ensure that the number of classi ers scales linearly with the hierarchy's depth.

Thanks to the text-to-text nature of T5, we can simplify HMTC into two text-to-text tasks shown in Figure 4. To prepare the task of categories prediction, we prepend the "categories prediction: " pre x to the text of segments and generate one sequence of categories separated by "; " as shown in Figure 5. The lists of categories were sorted in alphabetical order so that they have the same order across training examples, as advised by the authors of [33].

The second task's objective is to predict the values of the attributes of a category from an input segment, as well as to generate the spans of texts related to the predicted values. This task is similar to a reading comprehension task [34], where the question is "what is the value of the attribute?", and the context paragraph is the pair (segment, category). So we format it into a text-to-text task, similar to how the authors of T5 formatted the reading comprehension dataset SQuAD (see Figure 6).

Once we format the tasks, we ne-tune the largest available T5 of 11B parameters on these tasks. We try two ne-tuning methods: the rst method (advised by [33]) is to ne-tune on each task independently, and the second method is to ne-tune in a multi-task setting on a mixture of both tasks to capture the global labels hierarchy. We xed the hyperparameters of input sequence length and output sequence length to 512 and the batch size to 16, and performed a grid-search over the learning rate. The model was ne-tuned each time for 25,000 steps. Interestingly, the best-performing learning rate is the same (4e-3) for all the models and tasks.

4.5 Evaluation measures

Evaluation of multi-label classi cation. We use precision, recall, and F1-score metric to evaluate the extraction of both coarse-grained and ne-grained data practices from privacy policies segments. Since we are in a multi-label classi cation setting, we adapt the traditional single-label metrics to this setting by using the label-based metrics [52]: precision, recall, and F1-score for the j-th class label $\sim_9$ are de ned as follows:

[3] We used the pretrained model available at Hugging Face models hub.

Table 1: Results of categories prediction by the local classi-ers approach

Categories	CNN			XLNet		
	P	R	F1	P	R	F1
Introductory/Generic	74	40	52	76	54	63
Policy Change	85	60	71	73	65	69
Speci c audiences	90	77	83	85	80	82
Privacy Contact Info	87	52	65	84	75	79
1[BC] Party Collection	67	87	76	84	81	83
Data Retention	52	39	45	58	36	44
3[A3] party sharing/collection	71	85	78	76	87	81
User Choice/Control	45	79	58	66	69	67
Practice Not Covered	39	39	38	40	44	42
Data Security	79	48	60	77	68	72
Access, Edit, Deletion	87	35	50	75	72	74
Do Not Track	100	29	45	93	100	96
Macro-Average	72	55	62	76	71	73

$$\%_9 = \frac{)\%_9}{)\%_9 \, \%_9} \qquad '_9 = \frac{)\%_9}{)\%_9 \, \#_9} \qquad 1_9 = \frac{2 \ \%_9 \ '_9}{\%_9 \, '_9}$$

$)\%_9$, $\%_9$, $)\#_9$ and $\#_9$ are the number of true positive, false positive, true negative and false negative test examples with respect to class label $\sim_9$. To measure the global performance over a set of labels we compute the macro-average of each metric by averaging on the set of labels S.[4]

For the task of values prediction, we only report the result of attributes and values that we use to automate formal compliance checking of privacy policies, such that the reported metrics are the macro-average over the values necessary for compliance checking.

Evaluation of span extraction. We use the F1-score, as in the SQuaD dataset [34], to evaluate the extraction of spans associated with the values by comparing the ground truth target to the generated target.

4.6 Results and Discussion

Local classi ers approach. In Table 1, we report the results of the evaluation of the CNN and XLNet experiments on categories prediction. We present the results of CNN and the XLNet for the task of values prediction in Table 2. XLNet has superior performance comparing to CNN for the task of categories predictions. However, it has signi cantly lower performance than CNN for the second task, because there are enough examples to ne-tune XLNet for categories prediction but not enough for values prediction.

Text-to-text approach. We report the results of evaluating the two tasks on the test dataset in Table 3, Table 4, and Table 5. We observe that the individual ne-tuning and multi-task ne-tuning have a close recall for both tasks, but they di er signi cantly in precision. By ne-tuning each task separately, we obtain a 2.6% precision improvement for the task of categories prediction and

[4] It is worth noting that we don't use the same precision, recall and F1-score as in [14] where they use the macro-average of each metric predicting the presence and absence of the label.

Figure 4: The hierarchical multi-label classi cation of OPP-115 is converted into two text-to-text tasks. The rst task is to predict categories of data-practice from an input segment. We then retrieve the attributes of each predicted category to feed them and their category as input to the second task. The second task is to predict the values of the input attributes and to generate the corresponding span of text (highlighted).

Figure 5: Input example to train T5 for categories prediction

Figure 6: Example of an input to train T5 for prediction of attributes' values and the corresponding spans of text

Table 2: Results of values prediction by CNN and XLNet for attributes used in compliance checking. We report the macro-average over the values of each attribute.

Attribute	CNN			XLNet		
	P	R	F1	P	R	F1
action rst-party	44	47	45	16	33	21
does/does not	93	77	84	84	82	83
personal information type	73	58	64	56	49	52
purpose	74	56	64	74	64	69
retention period	79	62	69	50	11	18
access type	61	51	55	30	26	27
Macro-average	72	60	65	51	44	47

4.5% precision improvement for values prediction. This behavior

Table 3: Results of categories prediction by T5

Category	Multi-task Fine-tuning			Task 1 Fine-tuning		
	P	R	F1	P	R	F1
Introductory/Generic	68	61	64	69	58	63
Policy Change	82	67	74	82	69	75
Speci c audiences	94	84	88	91	85	87
Privacy Contact Info	81	82	82	79	69	74
1^{BC}Party Collection	88	87	87	90	85	87
Data Retention	65	47	55	77	52	62
3^{A3}party sharing/collection	83	86	85	85	85	85
User Choice/Control	65	60	62	66	64	65
Practice Not Covered	46	45	46	49	47	48
Data Security	73	66	69	78	70	74
Access, Edit, Deletion	82	84	83	78	76	77
Do Not Track	100	75	85	100	62	76
Macro-Average	77	70	73	79	69	73

Table 4: Results of values prediction for attributes used in compliance checking by T5. We report the macro-average over the values of each attribute.

Attribute	Multi-task Fine-tuning			Task 2 Fine-tuning		
	P	R	F1	P	R	F1
action rst-party	55	56	55	62	58	60
does/does not	90	78	83	91	83	87
personal information type	72	61	66	73	63	68
purpose	72	65	69	74	67	70
retention period	53	47	50	50	25	33
access type	62	58	60	71	70	70
Macro-average	67	61	63	70	61	65

where separate models trained on each task outperforms the multi-task model is coherent with previous ndings [4, 26, 33].

The performance of span extraction, presented in Table 5 is low in comparison with the performance of transformers models

on similar tasks such as reading comprehension or named entity recognition, which might be due to the relatively small number of training examples given to T5.

Table 5: Results of evaluation of span extraction by T5.

Multi-task Fine-tuning			Task 2 Fine-tuning		
P	R	F1	P	R	F1
64	57	52	65	59	54

Local classi ers approach vs. text-to-text approach. We present in Table 6 the macro-average of the di erent experiments for both tasks categories and values prediction. Fine-tuning separate models of T5 for each task achieves the highest F1 score. Both XLNet and T5 of the transformers family signi cantly outperform CNN on the rst task. However, the performance of transformers on the task of values prediction is at its best close to CNN's performance.

The local classi ers approach requires ne-tuning separate models for each attribute. Consequently, it decreases the number of training examples seen by each model, explaining the signi cant performance gap between task 1 and task 2. This performance gap is more important for XLNet than CNN, which could result from the high number of parameters of XLNet we need to ne-tune in comparison to the CNN architecture. T5 has even more parameters than XLNet, but its performance does not drop as signi cantly as XLNet. We can explain this di erence by the nature of the text-to-text approach where we ne-tune one model of T5 on the prediction of the values of all the attributes instead of individual ne-tuning for each attribute. Hence, T5 sees much more examples than XLNet.

Table 6: Macro-Average of precision, recall and F1-score for categories prediction (Task 1) and values prediction (Task 2) by the four models.

Approach		Task 1			Task 2		
		P	R	F1	P	R	F1
Local classi ers	CNN	72	55	62	72	60	65
	XLNet	76	71	73	51	44	47
Text-to-Text: T5	Multi-task FT	77	70	73	67	61	63
	Individual FT	79	69	73	70	61	65

5 RULE BASED COMPLIANCE CHECKING

5.1 From rules to code

This section describes our rule-based approach to automate the formal compliance checking of privacy policies. Privacy policies must comply with Articles 12, 13, and 14 of the GDPR. We limit our experiments to Articles 13 and 14, listing the mandatory information that privacy policies must contain, for which we can use information extraction algorithms, described in the previous section. To verify compliance with Article 12, we will need to develop other algorithms to assess how the mandatory information is communicated to data subjects (language complexity, length of sentences, etc.). Legal experts manually converted rules from Articles 13 and 14 into code using the OPP-115 taxonomy. As the OPP-115 taxonomy does not cover all the concepts of the GDPR, we only encoded the articles presented in the second column of Table 7.

Table 7: List of mandatory information from articles 13 and 14 of the GDPR encoded by us using the OPP-115 taxonomy.

Mandatory information	Article Reference
Identity of the controller	13 1.a; 14 1.a
Contact details of the controller	13 1.a; 14 1.a
Purpose of the processing of personal data	13 1.c; 14 1.c
Right to data portability	13 2.b; 14 2.c
Right to erasure	13 2.b; 14 2.c
Right to recti cation	13 2.b; 14 2.c
Right to access	13 2.b; 14 2.c
Data retention period or the criteria of retention period	13 2.a; 14 2.a
The recipients or categories of recipients of the personal data	13 1.e; 14 1.e
Categories of personal data	14 1.d

We plan to build a GUI for legal experts to convert compliance rules into code. To do so, we choose JsonLogic to serialize obtained rules as a JSON le. JsonLogic provides a simple mechanism to share rules between the front-end and back-end of a GUI. It comes with a parser that we use to build a rst-order logic inference engine with Python. In Figure 7 we present an example of a rule encoded with JsonLogic and OPP-115 taxonomy. We consider that the purpose of processing personal data is mentioned in the privacy policy if there is at least one data practice whose category is First Party Collection and the attribute "Purpose" has a value di erent from Unspeci ed .

```
{"some":[{ "var" :"data_practices"},
          {"and":[{ "==" :[{ "var" :"category"},"first_party"]},
                  {"!=" :[{ "var" :"attributes.purpose"},"unspecified"
                   ]}]}]}
```

Figure 7: Example of encoding a GDPR rule with JsonLogic. The rule states the obligation of mentioning the purpose of the processing of personal data.

5.2 Evaluation

The OPP-115 taxonomy was created before the entry into force of the GDPR. To evaluate its capacity to capture GDPR concepts, we create a dataset of 30 privacy policies where legal experts indicate each mandatory information's presence. We use this dataset as ground truth of mandatory information listed in Table 7.

The ground truth dataset contains two types of privacy policies: 15 privacy policies are from the OPP-115 dataset and 15 post-GDPR privacy policies are from the corpus released by [23]. We extract data practices from privacy policies and feed them to the inference engine to check for the presence of each mandatory information. For the rst type of privacy policy, we use the ground truth data practices extracted manually by legal experts from [43]. For the second type, we use data practices predicted by T5. Therefore, any error on the rst type of privacy policies will not be due to machine learning errors but due to the OPP-115 taxonomy used to encode mandatory information.

We report metrics of both the absence and presence of information in Table 8. Our objective is to detect non-compliance and send the documents for review by experts, so we need to maximize the number of absent mandatory information we can detect.

Table 8: Results of mandatory information detection from both OPP-115 and post-GDPR privacy policies.

Dataset	P	R	F1
OPP-115: absence	93	88	90
OPP-115: presence	93	97	95
Post-GDPR: absence	78	78	78
Post-GDPR: presence	91	91	91

Most of the errors on the OPP-115 dataset (see Figure 8) are caused by the difficulty of aligning OPP-115 concepts with GDPR concepts. For example, to encode "Data retention period" we used the "Retention period" attribute. However, even when the annotators select a value for the "Retention period" it [48] it does not always concern all the collected personal data. In contrast, the GDPR requires that the retention period is stated for all of the data.

The majority of errors on the post-GDPR policies are when the algorithm does not detect the right to data portability. This type of error is expected: because data portability has not been widely adopted before the GDPR (Art. 20), and no privacy policy from the OPP-115 dataset mentions the right to data portability.

Although our rules did not capture the right to data portability, T5 correctly predicted the category "data portability" from this sentence: "Data portability, that is to say the possibility of receiving these data in a structured format that is readable by an automatic device and of sending them to another processing owner without any impediments." T5 could predict new categories of data practices due to its language understanding capabilities that enable few and zero-shot inference [22, 50, 53]. Other new categories from the GDPR articles, predicted for post-GDPR policies are: "data minimisation" (Art. 5.1(c)), "data accuracy" (Art. 5.1(d)), "legitimate interests" (Art. 6.1(f)), "lawful basis" (Art. 6), and "right to object" (Art. 21).

Figure 8: The distribution of errors over types of mandatory information for OPP-115 and post-GDPR privacy policies

6 CONCLUSION

This study designed a theoretical framework to implement and monitor GDPR compliance in the data supply chain through a document-based approach, for which we defined three key tasks. We proposed a formal and substantive approach to verify GDPR-compliance of privacy policies. It is worth highlighting that, as a potential next step, our framework could be adapted to other compliance documents like DPIAs and/or ROPAs. More broadly, research is also needed to implement this framework in a multi-document setting, where data processing activities are described in multiple documents.

Our second significant contribution is the experimentation on the automation of formal compliance checking of privacy policies, which could be generalized to other documents in the data supply chain as a means to improve end-to-end compliance. We build a system combining machine learning and rules to detect the presence of information required by the GDPR. We fine-tuned the T5 model in a multi-task setting and achieved good performance predicting both coarse-grained and fine-grained data practices with only one model. The T5 model also extracts the spans of text corresponding to the fine-grained data practices. These spans of text could be used to explain the predicted values.

We used the OPP-115 taxonomy to encode 10 GDPR rules from Articles 13 and 14 concerning the information a privacy policy should contain. We evaluated the system on a corpus of 30 privacy policies, where legal experts indicated the presence of mandatory information. Although OPP-115 taxonomy is pre-GDPR, it proved capable of capturing some mandatory information in both pre-GDPR and post-GDPR policies. Currently, it is one of the most valuable resources in our research community. Still, it is not enough to encode both GDPR rules and data protection activities defined in compliance documents. Thus, there is a need for a new corpus of data protection documents from the data supply chain, to automate compliance checking tasks, which we leave for the future work.

Additionally, we pointed that T5 was able to predict new categories such as data portability. This capacity of zero-shot prediction can be leveraged to assist law and privacy scholars in creating a GDPR taxonomy compatible with the variety of compliance documents in the data supply chain.

REFERENCES

[1] 2017. The True Cost of Compliance with Data Protection Regulations. Technical Report. Ponemon Institute LLC.

[2] 2019. ICO Guidance: Update report into adtech and real time bidding. Technical Report. Information Commissioner's Office. 19 21 pages.

[3] David Restrepo Amariles, Aurore Clément Troussel, and Rajaa El Hamdani. 2020. Compliance Generation for Privacy Documents under GDPR: A Roadmap for Implementing Automation and Machine Learning. arXiv preprint arXiv:2012.12718 (2020).

[4] Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. arXiv preprint arXiv:1907.05019 (2019).

[5] Jaspreet Bhatia, Travis D Breaux, Joel R Reidenberg, and Thomas B Norton. 2016. A theory of vagueness and privacy risk perception. In 2016 IEEE 24th International Requirements Engineering Conference (RE). IEEE, 26 35.

[6] Giuseppe Contissa, Koen Docter, Francesca Lagioia, Marco Lippi, Hans-W Micklitz, Przemysaw Paka, Giovanni Sartor, and Paolo Torroni. 2018. Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence. Available at SSRN 3208596 (2018).

[7] Elisa Costante, Yuanhao Sun, Milan Petkovi¢, and Jerry Den Hartog. 2012. A machine learning solution to assess privacy policy completeness: (short paper).