GDPR Article Retrieval based on Domain-adaptive and Task-adaptive Legal Pre-trained Language Models

Andrea Simeri[1], Andrea Tagarelli[1,*]

[1]Dept. Computer Engineering, Modeling, Electronics, and Systems Engineering (DIMES), University of Calabria, 87036 Rende (CS), Italy

Abstract

The General Data Protection Regulation (GDPR) is an European regulation on data protection and privacy for all individuals within the European Union (EU) and the European Economic Area (EEA), and for all foreign subjects dealing with European citizens data. Therefore, the GDPR has important legislation implications that hold beyond EU member states. In this paper, we address the problem of GDPR article retrieval through the use of pre-trained language models (PLMs). Our approach features several key aspects, which include both domain-general and domain-specific pre-trained BERT models, further powered by self-supervised task-adaptive pre-training stages, with or without data enrichment based on recitals. Our study endeavors to demonstrate the potential of PLMs in addressing the challenges posed by the GDPR's intricate legal framework, thus ultimately facilitating efficient access to GDPR provisions for government agencies, law firms, legal professionals, and citizens alike.

Keywords

law article retrieval, domain adaptation, legal language models, artificial intelligence and law

1. Introduction

The General Data Protection Regulation (GDPR) stands as one of the most significant legal frameworks for data protection and privacy in recent years. Enforced by the European Union (EU) since May 2018, the GDPR has garnered global attention due to its wide-reaching impact on businesses, organizations, and individuals, transcending geographical boundaries. While initially conceived to safeguard the data rights of EU citizens, its influence extends far beyond EU member states, making it a pivotal legislation worldwide.

The GDPR is comprised of two components, namely the *articles* and *recitals*. The GDPR articles constitute the legal requirements that must be followed by organizations to demonstrate compliance. The GDPR currently in force consists of 99 articles, which are organized into 11 chapters: 'general provisions', 'principles', 'rights of the data subject', 'controller and processor', 'transfers of personal data to third countries or international organisations', 'independent supervisory authorities', 'cooperation and consistency', 'remedies, liability and penalties', 'provisions relating to specific processing situations', 'delegated acts and implementing acts', 'final provisions'. In addition to the articles, introductory statements and explanations, called *recitals*, provide context and guidance for the interpretation of the provisions of the regulation, thus

LIRAI'23: 1st Legal Information Retrieval meets Artificial Intelligence Workshop co-located with the 34th ACM Hypertext Conference, September 4, 2023, Rome, Italy

*Corresponding author.

andrea.simeri@dimes.unical.it (A. Simeri); andrea.tagarelli@unical.it (A. Tagarelli)

representing a valuable resource for determining the meaning and scope of the GDPR articles. The recitals are 173 in total, each associated to one or more articles; in turn, each article is associated with zero, one or more recitals.

The GDPR hence encompasses a comprehensive set of norms and principles that regulate the collection, processing, and transfer of personal data. Its provisions, such as the right to be forgotten, consent requirements, and data subject rights, have brought about significant changes in the digital landscape. Indeed, the complexity and scope of the GDPR pose challenges for government agencies, law firms, legal professionals, and citizens seeking to navigate its intricacies and access relevant regulations. Automating the search for GDPR information is demanding to address the need for efficient access to the GDPR contents. By employing advanced NLP technologies, the process of accessing and understanding GDPR provisions can be greatly facilitated.

In this regard, *pre-trained language models* (PLMs) such as BERT and GPT like models are the most helpful and attractive tools, given their widely known remarkable capabilities in various NLP tasks, including text classification, question answering, and document retrieval. In particular, we notice that opting for BERT-like models over GPT-like models for GDPR article retrieval offers several key advantages. BERT-like models prioritize precision, accuracy, and contextual understanding, mitigating the risks associated with hallucinations and ensuring reliable interpretations of the GDPR contents. Their emphasis on pre-training and fine-tuning, coupled with higher transparency and explainability than GPTs, make BERT-like models well-suited for addressing the complex task of article retrieval in the GDPR context, empowering users to access and comprehend data protection regulations with confidence and accuracy.

However, despite the potential benefits of leveraging PLMs for GDPR article retrieval, we are not aware of PLMs that have been specifically trained on GDPR texts to date. In this work, we aim to fill this gap by training BERT-based models for the task of GDPR article retrieval. More specifically, we train and fine-tune a pool of BERT models for a sequence classification task on the GDPR articles, with or without data enrichment based on the GDPR recitals. Our selected BERT models include not only the general domain (i.e., base) BERT but also the *legal BERT* models in [1], particularly the from-scratch pre-trained and further pre-trained versions. Furthermore, we originally propose two self-supervised task-adaptive pre-training strategies, namely *Related Sentence Prediction* and *Multiple Choice Answering*, which show key advantages on different query sets, which vary in terms of source, length, and lexical characteristics.

By harnessing the power of PLMs' contextual language understanding, we aim to provide an efficient and effective means for government agencies, law firms, legal professionals, and citizens to access and retrieve GDPR regulations. The outcomes of our research might contribute to enhancing accessibility, comprehension, and application of the GDPR, benefiting a wide range of stakeholders in their efforts to comply with data protection regulations and uphold individuals' rights.

2. Related work

Most existing approaches have focused on checking the GDPR *compliance* of privacy policies.

[2] proposes a conceptual model for characterizing the content of privacy policies in terms of information elements that one can expect to find in them (e.g., controller's identity and

contact). Based on named entity recognition and Glove word embeddings, such information elements are extracted and used to train a SVM model for a task of multi-label classification. The approach in [3] distinguishes between coarse-grained and fine-grained practices based on the OPP-15 taxonomy, and model them as a directed acyclic graph with a three level structure (i.e.; categories, attributes and values) so that the extraction of data practices is treated as a hierarchical multi-label classification task converted into two text-to-text tasks, one for each level of the label hierarchy, based on a T5 model. The extracted information are fed into a rule-based system that encodes the GDPR articles 13 and 14 under the supervision of legal experts in accord with the OPP-115 taxonomy.

[4] identifies four types of regulatory entities within policies of web services, which are used to define 16 classes. A BiLSTM is trained for a multi-class classification task, and a BERT summarizer is applied prior to the evaluation of context similarity for adhering vs. non-adhering policies. [5] exploits a supervised variant of Latent Dirichlet Allocation (i.e., Labeled LDA) to model topics associated with various types of violations of GDPR articles within real privacy incidents. The most relevant words associated with the topics induced by Labeled LDA are used to augment the training instances from the CMS.Law GDPR Enforcement Tracker database[1] to learn an LSTM classifier at GDPR article level. [6] leverages FastText word embeddings and a CNN model to predict privacy disclosure requirements according to GDPR articles 13 and 14. [7] is concerned with compliance of data processing agreements (DPAs), i.e., legally binding agreements that regulate the data processing activities according to GDPR. In relation to a predefined set of 45 requirements extracted from the GDPR provisions relevant to DPA, the proposed approach aims to assess whether a DPA is GDPR compliant, by comparing semantic-role-based representations of DPAs against predefined representations of the requirements. Also, the approach further provides recommendations about missing information in the DPA. In the context of GDPR compliance concerning the Italian Public Administration, [8] proposes a framework to detect security breaches related to unlawful disclosure of health information in public documents. Personally identifiable information as named entities are extracted and used to feed a machine learning classifier (e.g., SVM, XGBoost) for a binary classification task (i.e., compliant or non-compliant).

Unlike our work, the above studies have made limited use of PLMs and only focused on completeness or compliance/violation w.r.t. the GDPR; moreover, with regard to the latter aspect, the requirements checking is often carried out only within few GDPR articles (e.g., 13 and 14), although other articles contain fundamental GDPR requirements as well. By contrast, our work is the first to embrace the more general task of GDPR article retrieval by leveraging PLMs, and also powering them through self-supervised task-adaptive pre-training schemes. While our models can serve as a basis for further tasks like compliance checking — in fact, we recognize it as ongoing work (cf. Conclusions) — they offer a more general and versatile solution to automate and ease access to GDPR.

[1] https://www.enforcementtracker.com/

3. Self-supervised task-adaptive pre-training strategies

Task-adaptive fine-tuning is known to be the commonly used approach to enable a direct application of a pre-trained model to a downstream task. On the other hand, *domain-adaptive pre-training* allows for a more extensive customization of a pre-existing out-of-the-box model to a specialized language domain. In the legal domain, two approaches to domain-adaptive pre-training have been adopted: one is to continue pre-training the model using a legal corpus, while the other is to start the pre-training process from scratch using a legal corpus. An exemplary study adopting both approaches is provided in [1] for the development of the well-known Legal-BERT models.

However, the availability of legal data for a particular task may be limited, which can hinder the effective training of the model. Consequently, the model may struggle to fully grasp the meaning of legal texts and generalize the acquired knowledge in enough detail to handle unknown inputs successfully. It has been demonstrated that pre-training a model on a legal corpus does not always guarantee significant improvements over fine-tuning a corresponding domain-general pre-trained model on the target task (e.g., [9]). According to [10, 11], the advantages of domain-specific pre-training are particularly evident when dealing with low-resource downstream tasks.

Nonetheless, there exists another form of pre-training, which is to train a (pre-trained) model on a smaller corpus of the specialized domain such that the corpus is directly related to the target task but its documents are not annotated with the target class labels. This form of unsupervised pre-training, called *task-adaptive pre-training*, has shown competitiveness in comparison to domain-adaptive pre-training and can also enhance performance when combined with it for the downstream task [11]. However, these findings have not been proven specifically for the legal domain, presenting opportunities for further research in this area.

In this work, we aim to fill the above gap by developing the first approaches to task-adaptive pre-training of BERT models tailored to the GDPR article retrieval task. We shall describe our two proposed approaches in the following sections.

3.1. Related Sentence Prediction

Besides Masked Language Modeling, BERT was unsupervisedly trained on another pre-training objective, called Next Sentence Prediction (NSP), to cover a variety of downstream tasks involving sentence pairs. Given two word sequences as input, the NSP task is to determine if the second sequence is subsequent to the first in a document.

Inspired by the idea underlying NSP, we propose the *Related Sentence Prediction* (RSP) approach to task-adaptive pre-training. Basically, RSP is to predict if two given sentences in input are related to each other or not; the *relatedness* concept can be defined in more ways, and in this work we shall consider the location of two sentences within the same GDPR article.

Let us denote with $\mathcal{A}$ and $\mathcal{R}$ the sets of GDPR articles and recitals, respectively. Each article A_i can be modeled as a sequence $A_i = (p_{i,1}, \ldots, p_{i,n_i})$, where ps denote textual units belonging to A_i, i.e., sentences – following an analogy with NSP – or, more generally, paragraphs constituting A_i. For any $A_i \in \mathcal{A}$, we define $\mathsf{RSP}_D(A_i)$, with $D \in \{\mathcal{A}, \mathcal{R}\}$, as a meta-function expressing *relatedness* between A_i and different portions of the GDPR, thus producing a set of training

instances as associations between textual units of A_i and textual units from the document collection D. This means that, depending on the choice of D, sentences/paragraphs of A_i are coupled with either sentences/paragraphs from other article(s) than A_i, or with recitals. We refer to the first case (i.e., $D = \mathcal{A}$) as *article-level RSP*, and to the second case (i.e., $D = \mathcal{R}$) as *recital-level RSP*. In both cases, two types of training instances are built, which are *contrastive* to each other: the ones referring to "positive" relatedness (denoted with superscript $+$) and the other ones referring to "negative" relatedness (denoted with superscript $-$). We shall provide our definitions next.

Article-level RSP. For any article $A_i \in \mathcal{A}$, we define

$$\mathsf{RSP}_{\mathcal{A}}(A_i) = \mathsf{RSP}^{+}(A_i, A_i) \cup \mathsf{RSP}^{-}(A_i, \mathcal{A} \setminus A_i), \tag{1}$$

where $\mathsf{RSP}^{+}(A_i, A_i)$ is a function producing a set of intra-article pairings over the textual units of A_i as positive relatedness associations, and $\mathsf{RSP}^{-}(A_i, \mathcal{A} \setminus A_i)$ is a function producing a set of inter-article pairings between textual units of A_i and textual units from articles different from A_i, as negative relatedness associations. The above functions are specified so as to satisfy the following minimum requirements. First, to ensure balance between positive and negative associations, the number of training instances as pairs derived from A_i, which is bounded by $|A_i|(|A_i| - 1)/2$, is used to constrain the number of training instances derived by pairing units from A_i and units from any other article A_j, with $j \neq i$. Second, the choice of A_j is, by default, made uniformly at random, although constraints could be added to get A_j more or less "topically distant" from A_i (e.g., selecting A_j from the same chapter of A_i or from a different one). Third, multiple choices of A_j are made if $|A_j| < |A_i|$ (i.e., multiple articles need to be involved to form a number of negative associations to equal the number of positive ones); in general, using multiple articles against A_i can be useful to diversify the negative associations with A_i, thus obtaining a mix of negative training instances at different hardness levels.

Recital-level RSP. Leveraging recitals for training a model on an RSP task is strongly justified since they are originally conceived in the GDPR as essential complements for the articles. Based upon this, we provide a function definition analogous to the article-level one, which is as follows:

$$\mathsf{RSP}_{\mathcal{R}}(A_i) = \mathsf{RSP}^{+}(A_i, R^{(i)}) \cup \mathsf{RSP}^{-}(A_i, \mathcal{R} \setminus R^{(i)}), \tag{2}$$

where $R^{(i)}$ denotes the set of recitals associated with article A_i, the positive relatedness function $\mathsf{RSP}^{+}(A_i, R^{(i)})$ yields a set of training instances as pairs obtained by the textual units of A_i and the recitals in $R^{(i)}$, and the negative relatedness function $\mathsf{RSP}^{-}(A_i, \mathcal{R} \setminus R^{(i)})$ yields a set of training instances as pairs obtained by textual units of A_i and recitals not in $R^{(i)}$. It should be noted that we specify associations between portions of articles and the entire recitals, as we want to provide a maximal context based on recitals for each of the articles.

3.2. Multiple Choice Answering

Our second proposed task-adaptive pre-training approach is *Multiple Choice Answering* (MCA), which is defined as choosing the correct answer from a set of possible answers relating to an input query. In a sense, this task can be seen as a blend between the RSP task and the Masked Language Modeling task, which is at the core of BERT-like pre-training. In fact, the latter

requires to choose the correct word to fill in a mask from a set of possible options based on the context of an input sentence. Analogously, MCA requires to choose from a set of sentences that are presented as either positively or negatively related to an input one, in a similar fashion to the RSP task. Also, we again consider the opportunity of enriching the training through the recitals, therefore we shall distinguish between *article-level MCA* and *recital-level MCA*.

Article-level MCA. Given an integer $k > 1$ as the number of choices, for any article $A_i \in \mathcal{A}$, we define

$$\mathsf{MCA}_{\mathcal{A},k}(A_i) = \bigcup_{j=1}^{|A_i|} \langle \mathsf{MCA}^{+}(p_{i,j}, A_i), \mathsf{MCA}^{-}(p_{i,j}, \mathcal{A} \setminus A_i) \rangle, \tag{3}$$

where $\mathsf{MCA}^{+}(p_{i,j}, A_i)$ yields a pairing between $p_{i,j}$ and another unit randomly chosen from A_i (i.e., the positive or correct choice for $p_{i,j}$), and $\mathsf{MCA}^{-}(p_{i,j}, \mathcal{A} \setminus A_i)$ yields a $(k-1)$-sized set of pairings between $p_{i,j}$ and units each selected from randomly chosen articles A_j, with $j \neq i$. Overall, for each A_i, a set of training instances is computed, where each training instance is a tuple of size $k+1$ (i.e., a sentence/paragraph and its relating k choices). Note that the positions of the k choices are actually randomly shuffled so that the position of the correct choice is variable through all training instances.

Recital-level MCA. By changing the answering choice context from articles to recitals, we have the following definition:

$$\mathsf{MCA}_{\mathcal{R},k}(R^{(i)}) = \bigcup_{j=1}^{|A_i|} \langle \mathsf{MCA}^{+}(p_{i,j}, R^{(i)}), \mathsf{MCA}^{-}(p_{i,j}, \mathcal{R} \setminus R^{(i)}) \rangle, \tag{4}$$

where $R^{(i)}$ denotes the set of recitals associated with article A_i, and the positive, resp. negative, MCA functions have analogous definitions to the corresponding article-level ones. Note however that, consistently with the recital-level RSP definition, recitals are considered as atomic units.

4. Training and evaluation methodologies

4.1. Model selection and settings

Our study is versatile w.r.t. the choice of PLM to deal with the GDPR search and retrieval. As happened in the past for other novel applications of PLMs, demonstration through BERT (`bert-base-uncased`) [12] represents a primary choice. Moreover, this allows us to naturally couple evaluation based on a domain-general BERT model with evaluation based on legal specialized counterparts, which are the well-known family of models in [1]. Specifically, we use (i) the main model, named LegalBERT (`legal-bert-base-uncased`), which was pre-trained from scratch on large corpora including EU legislation, US contracts and cases, and UK legislation; (ii) a EU specific model, named EULegalBERT (`bert-base-uncased-eurlex`), which was further pre-trained starting from BERT base using EU legislation only.[2]

Table 1 provides a summary of our developed models. Suffix -r is used to denote the *recital-enriched* models, i.e., the recital-level RSP or MCA based models, as well as the base BERT,

[2]LEGAL-BERT models are available at https://huggingface.co/nlpaueb/legal-bert-base-uncased

Table 1
Summary of our developed models fine-tuned on the GDPR article retrieval task

Model name	task-adaptive pre-training	data-enrichment	Model name	task-adaptive pre-training	data-enrichment
BERT	✗	✗	BERT-r	✗	✓
LegalBERT	✗	✗	LegalBERT-r	✗	✓
EULegalBERT	✗	✗	EULegalBERT-r	✗	✓
BERT-RSP	RSP	✗	BERT-RSP-r	RSP	✓
BERT-MCA	MCA	✗	BERT-MCA-r	MCA	✓
LegalBERT-RSP	RSP	✗	LegalBERT-RSP-r	RSP	✓
LegalBERT-MCA	MCA	✗	LegalBERT-MCA-r	MCA	✓
EULegalBERT-RSP	RSP	✗	EULegalBERT-RSP-r	RSP	✓
EULegalBERT-MCA	MCA	✗	EULegalBERT-MCA-r	MCA	✓

LegalBERT, and EULegalBERT which were fine-tuned based on articles and recitals as the training data. Note also that, in the fine-tuning stage, each of the models was trained for 10 epochs, using cross-entropy as loss function, AdamW optimizer and initial learning rate selected within [1e-5, 5e-5] on batches of 256 examples.

4.2. Training data preparation

Data enhancement. GDPR articles exhibit three distinctive traits in their logical structure. First, the text of an article is commonly organized as a numbered sequence of paragraphs (commas), each sometimes formatted as an enumerated list of points. Second, an article often contains references to specific paragraphs or points of the same article as well as of other articles. Third, one or more recitals can be associated with specific paragraphs, subparagraphs, or points within the same article.

Such features of the GDPR articles prompted us to carry out some preprocessing aimed to enhance them for feeding a language model. In particular, we pursued a threefold goal: (i) to enrich the article contents by expanding the references to (portions of) other articles or chapters; (ii) to refactor structured parts of an article in order to resolve anaphoric passages; and (iii) to produce a recital-based labeling of the articles at the finest level, by leveraging associations of recitals to the individual paragraphs of an article, when available. While the latter required no particular effort, the first two objectives were accomplished through a semi-automatic process with manual supervision to produce a reliable outcome. Specifically, with regard to objective (i), we resolved each reference by replacing it either with the original text of the referred part or with an inferred short description, in the form of a citation (i.e., enclosed by quotation marks). Concerning objective (ii), any enumerated list of points in an article was either replaced by as many paragraphs as the number of points, each equally preceded by the common premise of the point list, or just flattened by keeping the premise once followed by the enumeration, in case of relatively short texts as points in the list.

Fine-tuning article labeling schemes. Creating a training dataset for the downstream task, i.e., GDPR article retrieval, requires that the entire corpus must be used to embed its knowledge fully, and each class label must correspond to a specific GDPR article since we want to learn how to classify at the article level. To this purpose, in order to build the training set for the fine-tuning task, we resort to an unsupervised article-labeling scheme proposed in [13]

which is designed to select and combine portions of each article, while ensuring balance of the contributions of each article, which clearly have different lengths. This scheme applies a round-robin method to iterate over replicas of the same group of training instances per article until a minimum number $minK$ of instances (to be produced for each article) is reached. In this work, we used the *unigram with parameterized emphasis on the title* scheme [13] creates a set of training instances for each article which is comprised of round-robin selected sentences from the article, along with replicas of the article's title; also, $minK$ was set to 64 instances.

4.3. Query sets

We are not aware of any publicly available benchmark for evaluating retrieval models on GDPR data. Therefore, we built our own query data as test sets, by varying them in terms of source, length, and lexical characteristics. We define the following query sets: **QA**, which contains sentences randomly extracted from the GDPR articles; **QpA**, where each sentence in QA is *paraphrased* through an English-Spanish-Italian-English machine translation of the queries (via Google Translate); **QR** and **QpR**, which are analogous of QA and QpA, respectively, but replacing articles with recitals; **QC**, which contains *expert commentary* texts related to the GDPR articles, i.e., a set of opinions and comments provided by experts in data protection and privacy;[3] **QCs**, where each comment in QC is broken down into its constituting sentences.

Each of the **QA** and **QpA**, resp. **QR** and **QpR**, sets contains 661, resp. 138, queries, with an average of 60 ($\pm$ 35), resp. 112 ($\pm$ 61), words per query. Also, **QC**, resp. **QCs**, contains 45, resp. 272, queries, with an average of 169 ($\pm$ 52), resp. 28 ($\pm$ 13), words per query.

4.4. Assessment criteria

Each query is associated with one article (ground-truth). For each article A_i, we first computed the following statistics: the recall for A_i (Rec_i), i.e., the number of queries s.t. A_i was correctly predicted out of all queries actually pertinent to A_i, the precision for A_i ($Prec_i$), i.e., the number of queries s.t. A_i was correctly predicted out of all predictions of A_i, and the F-measure for A_i, i.e., $F_i = 2Prec_i Rec_i/(Prec_i + Rec_i)$. Then, we averaged over all articles to obtain the per-article average *precision* ($Prec$), *recall* (Rec), *micro-averaged F-measure* (F^μ) as the average over all F_is, and *macro-averaged F-measure* (F^M) as the harmonic mean of $Prec$ and Rec.

In addition, we accounted for the top-3 predictions and the position (rank) of the correct article in predictions: the former is the fraction of correct article labels that are found in the top-3 predictions (i.e., top-3-probability results in response to each query), and averaging over all queries, which is the *recall@3* ($Rec@3$); the latter is the *mean reciprocal rank* (MRR) considering for each query the rank of the correct prediction over the classification probability distribution, and averaging over all queries.

5. Results

We organize our presentation of the results into two parts: the first reporting the performance of our models trained on articles only, and the second considering the recital-enriched models.

[3] https://gdpr-text.com

Table 2
Performance results of base BERT, LegalBERT, and EULegalBERT (*Bold values correspond to the best model, for each evaluation criterion and query set*)

Query type	Model	Rec	$Prec$	F^{μ}	F^{M}	$Rec@3$	MRR
QA	BERT	0.992	0.982	0.985	0.987	0.992	0.988
	LegalBERT	**0.996**	**0.987**	**0.990**	**0.991**	**0.997**	**0.993**
	EULegalBERT	0.984	0.934	0.946	0.958	0.965	0.962
QpA	BERT	0.949	0.930	0.929	0.939	0.952	0.933
	LegalBERT	**0.975**	**0.944**	**0.951**	**0.959**	**0.971**	**0.962**
	EULegalBERT	0.878	0.793	0.812	0.833	0.900	0.870
QR	BERT	0.667	0.620	0.596	0.643	0.739	0.683
	LegalBERT	**0.703**	**0.674**	**0.658**	**0.688**	**0.804**	**0.740**
	EULegalBERT	0.519	0.483	0.451	0.501	0.609	0.557
QpR	BERT	0.624	0.574	0.550	0.598	0.717	0.657
	LegalBERT	**0.707**	**0.685**	**0.662**	**0.696**	**0.775**	**0.726**
	EULegalBERT	0.371	0.325	0.291	0.347	0.558	0.478
QCs	BERT	0.340	0.590	0.390	0.431	**0.570**	0.519
	LegalBERT	**0.357**	**0.668**	**0.419**	**0.465**	0.548	**0.482**
	EULegalBERT	0.215	0.425	0.262	0.285	0.349	0.327
QC	BERT	0.556	0.669	0.590	0.607	0.778	0.752
	LegalBERT	**0.794**	**0.862**	**0.806**	**0.826**	**0.889**	**0.848**
	EULegalBERT	0.385	0.474	0.399	0.425	0.689	0.558

Table 3
Performance results of task-adaptive pre-trained BERT models and comparison with base BERT (*Bold values correspond to the best model, for each evaluation criterion and query set*)

Query type	Model	Rec	$Prec$	F^{μ}	F^{M}	$Rec@3$	MRR
QA	BERT	0.992	0.982	0.985	0.987	0.992	0.988
	BERT-RSP	0.990	0.980	0.981	0.985	0.994	0.989
	BERT-MCA	**0.998**	**0.990**	**0.993**	**0.994**	**0.997**	**0.996**
QpA	BERT	**0.949**	**0.930**	**0.929**	**0.939**	0.952	0.933
	BERT-RSP	0.930	0.895	0.895	0.913	0.926	0.906
	BERT-MCA	0.943	0.920	0.917	0.931	**0.961**	**0.945**
QR	BERT	0.667	0.620	0.596	0.643	0.739	0.683
	BERT-RSP	**0.684**	0.629	0.606	0.655	0.710	0.665
	BERT-MCA	0.675	**0.644**	**0.615**	**0.659**	**0.768**	**0.701**
QpR	BERT	0.624	0.574	0.550	0.598	0.717	0.657
	BERT-RSP	**0.671**	**0.594**	**0.579**	**0.630**	0.703	0.649
	BERT-MCA	0.588	0.530	0.512	0.557	**0.746**	**0.682**
QCs	BERT	0.340	0.590	0.390	0.431	0.570	0.519
	BERT-RSP	0.334	0.687	0.420	0.450	0.526	0.493
	BERT-MCA	**0.428**	**0.702**	**0.503**	**0.532**	**0.596**	**0.546**
QC	BERT	0.556	0.669	0.590	0.607	0.778	0.752
	BERT-RSP	0.613	0.702	0.641	0.655	0.733	0.720
	BERT-MCA	**0.754**	**0.821**	**0.758**	**0.786**	**0.867**	**0.811**

5.1. Training on articles only

Comparison of BERT models. Table 2 shows results obtained by the base BERT, LegalBERT, and EULegalBERT on the various query sets.

First, all models achieve high scores across all metrics over QA and QpA queries, indicating their ability to accurately retrieve relevant information; clearly, this is not surprising since QA and QpA queries contain information seen during the models' training. More challenging are the QR/QpR and QCs/QC queries which are based on contents from recitals and commentaries,

Table 4
Performance results of task-adaptive pre-trained LegalBERT models and comparison with base LegalBERT (*Bold values correspond to the best model, for each evaluation criterion and query set*)

Query type	Model	Rec	$Prec$	F^{μ}	F^{M}	$Rec@3$	MRR
QA	LegalBERT	0.996	0.987	0.990	0.991	0.997	0.993
	LegalBERT-RSP	**0.998**	**0.995**	**0.996**	**0.996**	0.995	0.995
	LegalBERT-MCA	**0.998**	0.991	0.994	0.995	**0.998**	**0.996**
QpA	LegalBERT	0.975	0.944	0.951	0.959	0.971	0.962
	LegalBERT-RSP	0.957	0.924	0.929	0.941	0.961	0.949
	LegalBERT-MCA	**0.980**	**0.952**	**0.959**	**0.966**	**0.982**	**0.969**
QR	LegalBERT	**0.703**	**0.674**	**0.658**	**0.688**	**0.804**	**0.740**
	LegalBERT-RSP	0.665	0.627	0.603	0.646	0.746	0.686
	LegalBERT-MCA	0.663	0.625	0.597	0.643	0.768	0.697
QpR	LegalBERT	**0.707**	**0.685**	**0.662**	**0.696**	**0.775**	**0.726**
	LegalBERT-RSP	0.697	0.675	0.645	0.686	0.754	0.694
	LegalBERT-MCA	0.654	0.613	0.586	0.633	0.754	0.692
QCs	LegalBERT	0.357	0.668	0.419	0.465	0.548	0.482
	LegalBERT-RSP	0.312	0.678	0.402	0.427	0.489	0.455
	LegalBERT-MCA	**0.426**	**0.692**	**0.493**	**0.527**	**0.629**	**0.560**
QC	LegalBERT	**0.794**	**0.862**	**0.806**	**0.826**	0.889	0.848
	LegalBERT-RSP	0.624	0.702	0.641	0.661	0.822	0.732
	LegalBERT-MCA	0.754	0.810	0.764	0.781	**0.911**	**0.860**

Table 5
Performance results of task-adaptive pre-trained EULegalBERT models and comparison with base EULegalBERT (*Bold values correspond to the best model, for each evaluation criterion and query set*)

Query type	Model	Rec	$Prec$	F^{μ}	F^{M}	$Rec@3$	MRR
QA	EULegalBERT	0.984	0.934	0.946	0.958	0.965	0.962
	EULegalBERT-RSP	**0.996**	**0.990**	**0.992**	**0.993**	**0.998**	**0.995**
	EULegalBERT-MCA	0.986	0.965	0.973	0.975	0.986	0.983
QpA	EULegalBERT	0.878	0.793	0.812	0.833	0.900	0.870
	EULegalBERT-RSP	**0.955**	**0.942**	**0.939**	**0.948**	**0.970**	**0.957**
	EULegalBERT-MCA	0.921	0.877	0.881	0.898	0.955	0.927
QR	EULegalBERT	0.519	0.483	0.451	0.501	0.609	0.557
	EULegalBERT-RSP	**0.689**	**0.667**	**0.633**	**0.678**	**0.754**	**0.691**
	EULegalBERT-MCA	0.665	0.624	0.605	0.644	0.674	0.651
QpR	EULegalBERT	0.371	0.325	0.291	0.347	0.558	0.478
	EULegalBERT-RSP	**0.648**	**0.608**	**0.585**	**0.627**	**0.754**	**0.682**
	EULegalBERT-MCA	0.638	0.571	0.566	0.603	0.681	0.657
QCs	EULegalBERT	0.215	0.425	0.262	0.285	0.349	0.327
	EULegalBERT-RSP	0.344	0.670	0.425	0.455	0.533	0.491
	EULegalBERT-MCA	**0.391**	**0.730**	**0.453**	**0.509**	**0.544**	**0.506**
QC	EULegalBERT	0.385	0.474	0.399	0.425	0.689	0.558
	EULegalBERT-RSP	0.603	0.724	0.644	0.658	**0.778**	0.750
	EULegalBERT-MCA	**0.763**	**0.882**	**0.801**	**0.818**	**0.778**	**0.781**

respectively. Generally, LegalBERT consistently outperforms BERT, which would indicate the benefits of adapting the models to the legal domain prior to the GDPR article retrieval task. However, EULegalBERT shows significantly lower performance compared to the other two models. This performance gap should be ascribed to the fact that EULegalBERT results from a further pre-training of BERT over EU specific resources, which limited knowledge expansion w.r.t. that gained by LegalBERT over a much larger set of legal corpora in relation to BERT.

Impact of task-adaptive pre-training. Let us now focus on the impact of task-adaptive

pre-training on the GDPR article retrieval task, whose results are summarized in Tables 3–5; on each of those tables, we also report the scores achieved by the corresponding model without task-adaptive pre-training (cf. Table 2).

Considering first BERT models (Table 3), we find that BERT-MCA achieves the highest scores for most query types, and the advantage over the other two models are particularly evident for the most difficult query sets. BERT-RSP generally performs better than the base BERT in terms of precision, recall, and f-measures, with the exception of QA/QpA query sets, which would indicate that task-adaptation based on the RSP pre-training task can even worsen performance on article retrieval when the query content does not deviate much from the training data. Moreover, BERT-RSP consistently falls behind BERT-MCA, which highlights the superiority of the latter form of task-adaptive pre-training when applied to BERT.

As reported in Table 4, task-adaptive pre-training of LegalBERT showcases quite different trends from the BERT counterpart. While MCA maintains a significant advantage over RSP especially for the commentary-based queries (i.e., QCs and QC), the same does not hold for the other queries, especially the recital-based queries (i.e., QR and QpR). Also, on the latter query types and on QC, both variants of task-adaptive pre-training are not able to improve performance over the base LegalBERT. This would suggest MCA (and RSP) might not necessarily take advantage when, as it is the case for LegalBERT, the base model has a pre-training knowledge on a legal domain that would enclose the targeted one (i.e., GDPR in our setting).

Table 5 shows how EULegalBERT appears to take advantage when task-adapted via RSP, on query sets QA- QpR, or via MCA, on commentary-based queries. However, compared to the previous results, EULegalBERT models are still outperformed by LegalBERT and BERT models (apart from very few exceptions, such as precision on QCs and QC queries).

Overall, our findings suggest that task-adaptive pre-training, especially based on MCA, can yield better results when applied to base BERT and its legal pre-trained models, and this particularly holds in terms of $Rec@3$ and MRR criteria.

5.2. Training with recital data enrichment

Table 6 shows results corresponding to the recital-enriched models. For each query set and criterion, the table reports the percentage variation (i.e., increase/decrease) achieved by a recital-enriched model w.r.t. its corresponding non-recital-enriched model; moreover, the last row of each query-set subtable shows the absolute best-performing model, considering both the recital-enriched models and the previously analyzed models.

First, we notice that while the improvements are negligible for the QA/QpA query types, some significant increase in performance holds for the QCs/QC query types, particularly for the BERT and EULegalBERT models. Also, it clearly does not come to our surprise that leveraging recitals is beneficial for all domain-adaptive and task-adaptive pre-trained models when evaluated on recital-based queries (i.e., QR and QpR). In such cases, the absolute best model is LegalBERT-r, which also indicates that task-adaptive pre-training is not needed for the target task as its lack can be well compensated by the recital data enrichment.

More importantly, task-adaptive pre-training, especially based on MCA, reveals to be essential to maximize performance in relation to the non-recital-based query sets. Moreover, a combination of MCA and recital-enrichment leads to the absolute best model for the QCs type. This is

Table 6
Percentage variations of the recital-enriched BERT-r, LegalBERT-r, and EULegalBERT-r w.r.t. their respective base models (i.e., BERT, LegalBERT, and EULegalBERT). The absolute best-performing model is reported in the last row of each query-set subtable.

Query type	Model	Rec	$Prec$	F^{μ}	F^{M}	$Rec@3$	MRR
QA	BERT-r	-0.51%	+0.28%	-0.28%	-0.11%	-0.15%	-0.49%
	BERT-RSP-r	+0.39%	+0.79%	+0.69%	+0.59%	+0.46%	+0.69%
	BERT-MCA-r	+0.27%	+0.13%	+0.28%	+0.20%	+0.31%	+0.31%
	LegalBERT-r	-0.67%	+0.03%	-0.58%	-0.32%	0%	-0.39%
	LegalBERT-RSP-r	-0.34%	-0.16%	-0.28%	-0.25%	-0.15%	-0.22%
	LegalBERT-MCA-r	-0.27%	-0.14%	-0.20%	-0.21%	-0.30%	-0.26%
	EULegalBERT-r	+0.14%	+3.47%	+2.69%	+1.82%	+2.67%	+1.88%
	EULegalBERT-RSP-r	+1.32%	+6.17%	+5.04%	+3.75%	+3.14%	+3.27%
	EULegalBERT-MCA-r	+1.13%	+5.84%	+4.72%	+3.49%	+3.46%	+3.20%
	LegalBERT-RSP (-MCA)*	0.998	0.995	0.996	0.996	0.998*	0.9965*
QpA	BERT-r	-4.03%	-6.46%	-5.91%	-5.27%	-0.16%	-1.42%
	BERT-RSP-r	-0.48%	-1.27%	-0.72%	-0.88%	+0.64%	+1.20%
	BERT-MCA-r	-1.19%	-2.96%	-3.08%	-2.09%	+0.16%	-1.14%
	LegalBERT-r	-3.46%	-4.39%	-4.29%	-3.93%	+0.31%	-0.79%
	LegalBERT-RSP-r	-1.30%	+0.10%	-0.96%	-0.59%	+0.16%	-0.68%
	LegalBERT-MCA-r	-1.89%	-1.62%	-1.59%	-1.75%	-0.16%	-0.63%
	EULegalBERT-r	-10.23%	-2.18%	-7.21%	-6.17%	-2.86%	-3.78%
	EULegalBERT-RSP-r	+8.46%	+16.85%	+14.61%	+12.71%	+7.23%	+9.25%
	EULegalBERT-MCA-r	+8.96%	+16.98%	+14.79%	+13.03%	+7.40%	+8.66%
	LegalBERT-MCA	0.980	0.952	0.959	0.966	0.982	0.969
QR	BERT-r	+41.83%	+51.99%	+57.80%	+46.92%	+34.31%	43.83%
	BERT-RSP-r	+5.89%	+7.92%	+8.57%	+6.94%	+5.88%	+4.08%
	BERT-MCA-r	+4.64%	+2.39%	+4.30%	+3.46%	+10.78%	+5.22%
	LegalBERT-r	+34.55%	+39.80%	+43.14%	+37.18%	+23.42%	+32.69%
	LegalBERT-RSP-r	+4.10%	+4.46%	+3.95%	+4.28%	-2.70%	-0.19%
	LegalBERT-MCA-r	+12.09%	+12.51%	+12.29%	+12.31%	+2.70%	+5.13%
	EULegalBERT-r	+82.15%	+95.01%	+108.48%	+88.59%	+63.10%	+76.29%
	EULegalBERT-RSP-r	+29.58%	+31.65%	+35.17%	+30.64%	+22.62%	+22.64%
	EULegalBERT-MCA-r	+44.10%	+47.45%	+51.29%	+45.81%	+26.19%	+29.00%
	LegalBERT-r	**0.946**	**0.942**	**0.941**	**0.944**	**0.993**	**0.982**
QpR	BERT-r	+48.05%	+60.76%	+66.33%	+54.41%	+36.36%	+46.15%
	BERT-RSP-r	+11.29%	+9.67%	+11.90%	+10.44%	+6.06%	+7.05%
	BERT-MCA-r	+8.93%	+12.44%	+11.81%	+10.73%	+11.11%	+7.53%
	LegalBERT-r	+35.72%	+38.70%	+43.24%	+37.22%	+28.04%	+35.30%
	LegalBERT-RSP-r	+4.08%	+4.33%	+4.48%	+4.21%	+1.87%	+2.88%
	LegalBERT-MCA-r	+1.39%	-1.83%	-2.12%	-0.27%	+6.54%	+4.51%
	EULegalBERT-r	+150.46%	+180.98%	+212.85%	+165.86%	+74.03%	+100.73%
	EULegalBERT-RSP-r	+80.38%	+93.20%	+107.180	+87.00%	+31.17%	+42.34%
	EULegalBERT-MCA-r	+95.110%	+110.51%	+129.21%	+103.03%	+32.47%	+48.59%
	LegalBERT-r	**0.960**	**0.950**	**0.949**	**0.955**	**0.993**	**0.982**
QCs	BERT-r	-1.20%	+9.66%	+1.84%	+2.51%	-6.45%	-9.89%
	BERT-RSP-r	-0.21%	+14.13%	+9.03%	+4.59%	-1.94%	-0.21%
	BERT-MCA-r	+17.88%	+16.34%	+23.19%	+17.31%	+8.39%	+4.63%
	LegalBERT-r	+11.43%	+6.32%	+10.03%	+9.60%	0%	+2.87%
	LegalBERT-RSP-r	-0.70%	-2.51%	+5.24%	-1.34%	+10.74%	+8.78%
	LegalBERT-MCA-r	+26.91%	+15.20%	+25.25%	+22.57%	+12.08%	+17.86%
	EULegalBERT-r	-30.57%	-7.88%	-28.59%	-24.314	-16.842	-18.575
	EULegalBERT-RSP-r	+105.58%	+80.23%	+97.89%	+96.32%	+64.21%	+69.46%
	EULegalBERT-MCA-r	+67.82%	+60.49%	+70.08%	+65.29%	+65.26%	+61.23%
	LegalBERT-MCA-r	**0.452**	**0.770**	**0.524**	**0.570**	**0.614**	**0.569**
QC	BERT-r	+13.88%	+5.64%	+10.04%	+9.99%	+8.57%	-2.21%
	BERT-RSP-r	+13.06%	+19.29%	+15.70%	+15.80%	+2.86%	-1.00%
	BERT-MCA-r	+15.71%	+4.06%	+11.09%	+10.12%	+8.57%	+5.19%
	LegalBERT-r	-6.57%	-6.54%	-7.28%	-6.56%	+2.50%	-2.56%
	LegalBERT-RSP-r	-2.71%	-3.22%	-2.82%	-2.96%	+2.50%	-1.29%
	LegalBERT-MCA-r	-6.86%	-5.16%	-5.09%	-6.05%	+2.50%	-2.38%
	EULegalBERT-r	-7.94%	+11.56%	-0.61%	-0.11%	-22.58%	-15.89%
	EULegalBERT-RSP-r	+101.77%	+75.89%	+94.54%	+89.27%	+22.58%	+45.16%
	EULegalBERT-MCA-r	+61.18%	+54.10%	+64.33%	+57.93%	+12.90%	+36.89%
	LegalBERT (-MCA)*	0.794	0.862	0.806	0.826	0.911*	0.860*

another remarkable aspect since it supports our intuition of the beneficial effect of integrating the complementary role of recitals for the GDPR articles into task-adaptive pre-training of models to be fine-tuned for the article retrieval task.

6. Conclusions

Summary. We addressed the problem of GDPR article retrieval through the use of PLMs, which include both domain-general and domain-specific pre-trained BERT models, further powered by self-supervised task-adaptive pre-training stages, with or without data enrichment based on recitals. To the best of our knowledge, this is the first study that explores a number of aspects concerning both domain-adaptive and task-adaptive legal pre-trained language models for the task of GDPR article retrieval.

Ongoing work. We are currently working on an evaluation of our proposed models on the CMS.Law GDPR Enforcement Tracker database. Preliminary experimental results have shown the effectiveness of our models, both in absolute terms and in relation to the article classification approach proposed in [5] relying on labeled LDA and an LSTM model (cf. Related Work).

As previously discussed, our models for GDPR article retrieval can serve as a basis for a variety of similarity based tasks, including question-answering. Indeed, we are working on further developments of our models to deal with GDPR compliance and violation checking tasks. In particular, we are developing a hybrid framework based on a combination of BERT-like models and ChatGPT-like models in order to take advantage of similarity search and classification capabilities of the former and conversational functionality of the latter.

References

[1] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 2898––2904. doi:10.18653/v1/2020.findings-emnlp.261.

[2] D. Torre, S. Abualhaija, M. Sabetzadeh, L. C. Briand, K. Baetens, P. Goes, S. Forastier, An AI-assisted Approach for Checking the Completeness of Privacy Policies Against GDPR, in: Proc. of the 28th IEEE International Requirements Engineering Conference (RE 2020), IEEE, 2020, pp. 136–146. doi:10.1109/RE48521.2020.00025.

[3] R. E. Hamdani, M. Mustapha, D. R. Amariles, A. C. Troussel, S. Meeùs, K. Krasnashchok, A combined rule-based and machine learning approach for automated GDPR compliance checking, in: Proc. of the Eighteenth International Conference for Artificial Intelligence and Law (ICAIL 2021), ACM, 2021, pp. 40–49. doi:10.1145/3462757.3466081.

[4] L. Elluri, S. S. L. Chukkapalli, K. P. Joshi, T. Finin, A. Joshi, A BERT based approach to measure web services policies compliance with GDPR, IEEE Access 9 (2021) 148004–148016. doi:10.1109/ACCESS.2021.3123950.

[5] A. Aleroud, F. Masalha, A. A. Saifan, Identifying GDPR privacy violations using an augmented LSTM: toward an ai-based violation alert systems, in: Proc. of the IEEE International Conference on Parallel & Distributed Processing with Applications, Big Data

& Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom), IEEE, 2021, pp. 1617–1624. doi:10.1109/ISPA-BDCloud-SocialCom-SustainCom52081.2021.00216.

[6] T. A. Rahat, M. Long, Y. Tian, Is your policy compliant?: A deep learning-based empirical study of privacy policies' compliance with GDPR, in: Proc. of the 21st Workshop on Privacy in the Electronic Society (WPES 2022), ACM, 2022, pp. 89–102. doi:10.1145/3559613.3563195.

[7] O. Amaral, M. I. Azeem, S. Abualhaija, L. C. Briand, NLP-based Automated Compliance Checking of Data Processing Agreements against GDPR, CoRR abs/2209.09722 (2022). doi:10.48550/arXiv.2209.09722.

[8] F. Lorè, P. Basile, A. Appice, M. de Gemmis, D. Malerba, G. Semeraro, An AI framework to support decisions on GDPR compliance, Journal of Intelligent Information Systems (2023) 1–28.

[9] A. Simeri, A. Tagarelli, Exploring domain and task adaptation of LamBERTa models for article retrieval on the Italian Civil Code, in: Proc. of the 19th Conference on Information and Research science Connecting to Digital and Library science (IRCDL 2023), volume 3365 of *CEUR Workshop Proceedings*, 2023, pp. 130–143. URL: https://ceur-ws.org/Vol-3365/paper4.pdf.

[10] S. Wang, M. Khabsa, H. Ma, To pretrain or not to pretrain: Examining the benefits of pretraining on resource rich tasks, in: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), ACL, 2020, pp. 2209–2213.

[11] S. Gururangan, A. Marasovic, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL), ACL, 2020, pp. 8342–8360.

[12] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT 2019), Association for Computational Linguistics, 2019, pp. 4171–4186.

[13] A. Tagarelli, A. Simeri, Unsupervised law article mining based on deep pre-trained language representation models with application to the Italian civil code, Artif. Intell. Law 30(3) (2022) 417–473. Published: 15 September 2021. doi:10.1007/s10506-021-09301-8.