Information Systems xxx (xxxx) xxx

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

The GDPR enforcement fines at glance

Jukka Ruohonen *, Kalle Hjerppe

Department of Future Technologies, University of Turku, FI-20014, Turun yliopisto, Finland

ARTICLE INFO

Article history:
Received 8 October 2020
Received in revised form 26 July 2021
Accepted 28 July 2021
Available online xxxx
Recommended by Andrea Tagarelli

Keywords:
Data protection
Privacy
Law enforcement
Public administration
Legal mining
Empirical jurisprudence

ABSTRACT

The General Data Protection Regulation (GDPR) came into force in 2018. After this enforcement, many fines have already been imposed by national data protection authorities in Europe. This paper examines the individual GDPR articles referenced in the enforcement decisions, as well as predicts the amount of enforcement fines with available meta-data and text mining features extracted from the enforcement decision documents. According to the results, three articles related to the general principles, lawfulness, and information security have been the most frequently referenced ones. Although the amount of fines imposed vary across the articles referenced, these three particular articles do not stand out. Furthermore, a better statistical evidence is available with other meta-data features, including information about the particular European countries in which the enforcements were made. Accurate predictions are attainable even with simple machine learning techniques for regression analysis. Basic text mining features outperform the meta-data features in this regard. In addition to these results, the paper reflects the GDPR's enforcement against public administration obstacles in the European Union (EU), as well as discusses the use of automatic decision-making systems in judiciary.

1. Introduction

Data protection has a long history in Europe [1].[1] With respect to the EU, the GDPR repealed the earlier Directive 95/46/EC. Although this directive laid down much of the legal groundwork for EU-wide data protection, its national adaptations, legal interpretations, and enforcement varied both across the member states and different EU institutions [3–5]. In short: it was a paper tiger. Later on, the provisions for both privacy and data protection were strengthened by the inclusion of them in the Charter of Fundamental Rights of the European Union (EU), signed with the Treaty of Lisbon in 2009. The GDPR is the latest manifestation in this path: the goal of the regulation is to protect natural persons with respect to the processing of their personal data, and, therefore, the goal is also to guard their fundamental right to data protection.

The GDPR has been extensively studied in recent years. To put political, economic, and related reasons aside, the reason for the abundance of research originates from the regulation's scope. The fifth Article (A) defines personal data as any information relating to an identified or identifiable natural person. Thus, with few restrictions, as specified in A23 and A89, the GDPR covers all processing activities of personal data, whether manual or automated. This wide scope means that it is difficult to consider the regulation without a context. The protection of personal data is different for information systems than it is for biomedical applications; it differs between scholarly disciplines, from computer science to medicine. The GDPR establishes only a few general principles that are universal. As specified in A5, these include lawfulness, fairness, and transparency, purposefulness, data minimization, accuracy, finite data retention, integrity, confidentiality, and accountability. It is possible to derive design patterns from these principles [6,7], but the patterns are still dependent on a given context. By implication, it is impossible to establish universal guidelines with which sanctions could be avoided. This provides a motivation for the present work to examine the specific articles that have been referenced by data protection authorities (DPAs) when imposing fines according to the conditions specified in A83.

Another motivation stems from the noted administration and governance issues for European data protection practices. Akin to some other public administration domains, such as product safety administration [8], the history of the European data protection has always relied heavily on the ombudsmen-like DPAs instead of enforcement through litigation or criminal law [9,10]. However, a reasonably comprehensive literature search indicates no previous empirical research on the enforcement of this particular regulation, excluding an earlier conference paper [2] upon which the present paper builds. Compared to the conference paper, the present work presents a more thorough examination of the enforcement fines, including the prediction of these by text mining techniques and regression analysis. The predictions

* Corresponding author.
E-mail address: juanruo@utu.fi (J. Ruohonen).

[1] This paper is an extended version of an earlier conference paper presented at COUrT – CAiSE for legal documents workshop [2].

https://doi.org/10.1016/j.is.2021.101876
0306-4379/

are also discussed with respect to a broader debate on automatic decision-making (ADMs) systems used in the public sector. In addition, the work extends the examination toward the GDPR's administrative and political aspects. To these ends, the present paper examines the following three Research Questions (RQs) regarding the enforcement fines:

- RQ_1: *Which GDPR articles have been actively referenced in the recent enforcement cases?*
- RQ_1: *Do the enforcement fines vary across the articles referenced in the enforcement decisions?*
- RQ_1: *How well the recent GDPR fines can be predicted in terms of basic available (i) meta-data and (ii) textual traits derived from the enforcement decisions?*

It is difficult to make prior speculations about potential answers to the questions. Regarding RQ_1, it can be expected that A5 is frequently referenced as it specifies the overall lawfulness condition for processing personal data. But beyond that, the GDPR contains as many as 99 articles, many of which may be used to justify sanctions. As for RQ_2, it could be hypothesized that information security lapses would yield particularly severe penalties; data breaches, in particular, have often been seen as a major deterrent of the GDPR for companies [11]. With respect to RQ_3, there is a more practical motivation: by knowing whether the penalties are predictable by machine learning techniques, a starting point is available for providing further insights in different practical scenarios. These scenarios include the automated archival of enforcement decisions, information retrieval, designation of preventive measures, and last but not least, litigation preparations.

From a data mining perspective, an answer to RQ_3 further paves the way for better understanding whether the manual labor required to construct meta-data from unstructured administrative documents is necessary for predictive tasks—or whether the documents are sufficient themselves. To this end, the paper uses meta-data and text miming features extracted from the decision documents. As such, only black-box predictions are sought; the goal is not to make any legal interpretations whatsoever. The black-box approach also places the paper into a specific branch of existing research dealing with legal and administrative documents. After a brief further motivation for the regulation's enforcement in Section 2, the related branch of work is discussed in Section 3. Thereafter, the paper's structure is straightforward: the dataset and methods are elaborated in Sections 4 and 5, results are presented in Section 6, limitations are discussed in Section 7, and conclusions are summarized in the final Section 8.

2. Background

There are many different viewpoints for approaching the enforcement penalties. One possibility would be to focus on non-compliant products. As noted, however, it is difficult to make generalizations due to the variety of products processing personal data. Another viewpoint is to focus on the regulators instead of the regulation; on the administration of the GDPR by national DPAs and their EU-level coordination institutions. This viewpoint is suitable for the present purposes. In contrast to domain-specific studies on the GDPR and conformance with it, relatively little has also been written from this administrative viewpoint.

In contrast to Directive 95/46/EC, Regulation (EU) 2016/679, the GDPR, is a regulation; it is binding throughout the EU with only a minimal space for national adaptations.[2] In practice, only a few articles in the GDPR provide some but limited room for national maneuvering; these include A6 with respect to relaxation in terms of other legal obligations or public interests, A9 in terms of sensitive data, and A10 regarding criminal matters. Thus, in general, this particular legislation should be interpreted and enforced uniformly through the European Union by national data protection authorities whose *formal* powers are defined in A58. In practice, however, already the resources and thus the *actual* power for enforcement have varied across the member states [12, 13]. Although the budgets of the DPAs have increased after the enactment of the GDPR in 2016 and its later enforcement in 2018, the resources remain scarce according to many critics. The resourcing obstacles were also acknowledged by the European Commission in its 2020 review of the GDPR's implementation. Accordingly, there is still a "need for data protection authorities to be equipped with the necessary human, technical and financial resources to effectively carry out their tasks" [14]. Besides plain budgetary aspects, the lack of human resources is worth emphasizing. As can be concluded from Fig. 1, the amount of personnel employed by national DPAs vary greatly across Europe. There is also an apparent lack of engineers and other technical specialists employed by the national DPAs. Most of the current employees are civil servants specialized to administration, jurisprudence, and related non-technical areas of expertise.

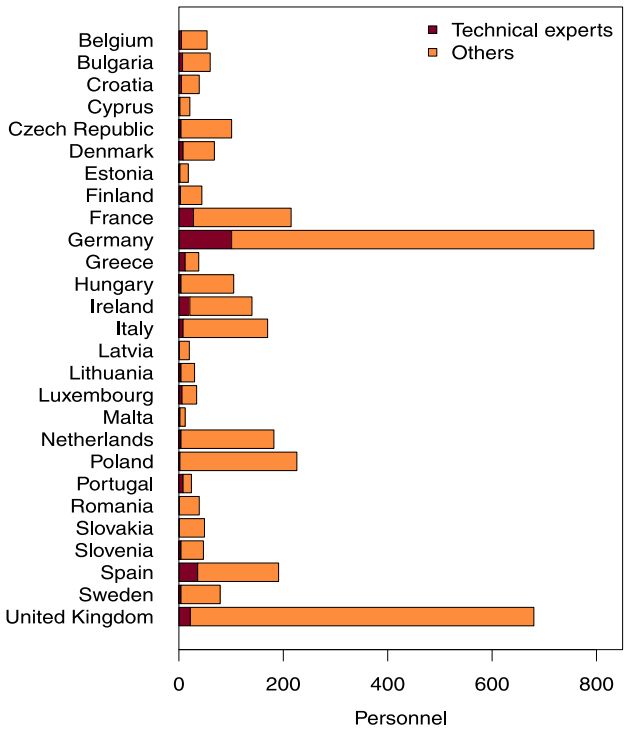

Fig. 1. Personnel employed by DPAs in selected European countries in 2020 (full-time employees addressing technical issues in private sector data processing; based on estimates reported in [15]).

By hypothesis, this evident cross-country variance reflects itself also in terms of the enforcement fines imposed by the national DPAs. There are good reasons to expect that the hypothesis is true. For instance, the enforcement of the GDPR has been continuously criticized by some public authorities and pundits alike. In addition to the lack of resources and the so-called "one-stop-shop" system, there are many other tenets in the criticism, including a lack of transparency and cooperation between DPAs, diverging legal interpretations, cultural conflicts, prioritization inconsistencies, old-fashioned information systems, and general over-tolerance or even reluctance to enforce laws [2,16,17].

[2] Either the GDPR or comparable national laws have been adopted also by countries participating in the EU's internal market via the European Economic Area (EEA) treaty. For brevity, however, this detail is omitted in what follows.

Although already the legacy Directive 95/46/EC established the autonomy of DPAs, data protection issues have also frequently prompted different bureaucratic conflicts and power struggles within national public administration systems [1,18]. Fragmentation in terms of the national adaptations of the 1995 directive was also a well-recognized problem [19]. The interplay between national and EU-level administration has caused additional problems for European data protection [20]. These are hardly unique issues in the European Union in general.

Therefore, these problems and the cross-country incoherence should not be overemphasized. Similar problems exist in many other policy areas in the EU, including closely related ones such as cyber security [21,22] and product safety [8] administration. Given that the GDPR contains information security requirements (as specified particularly in A5 and A32), data protection also aligns with cyber security in Europe. From this viewpoint, the GDPR is best portrayed as a one piece in the EU's broader judicial framework dealing with cyber security, trust, privacy, electronic commerce, and even cyber crime [5,23,24]. The same applies to the enforcement and administration of the corresponding laws. According to recent interviews of some key policy stakeholders, indeed, the role played by DPAs is ranked high also with respect to cyber security [25]. Given this broader viewpoint, perhaps more than anything else, the GDPR's early enforcement problems reflect the general administrative and political problems in the EU. And given these problems in turn, it may be that comprehensive enforcement will be done in court rooms through class actions [16]. To this end, Directive 2019/2161 has already been enacted for allowing collective redress for consumers and their representatives.

3. Related work

Legal mining – for lack of a better term – has emerged in recent years as a promising but at times highly contested interdisciplinary field that uses machine learning techniques to analyze various aspects related to law [26,27]. Although the concrete application domains vary, case law and court cases are the prime examples already because these constitute the traditional kernel of legal scholarship. Within this kernel, existing machine learning applications range from the profiling of judges' personal characteristics [28,29], which may be illegal in some European countries [30], to the prediction of decisions made by the European Court of Human Rights [31,32], the Court of Justice of the European Union [33], and related chief judicial authorities in Europe and elsewhere. These case law examples convey the two traditional functions of applied machine learning; exploratory data mining and forecasting.

Oftentimes, the legal mining domain is further motivated by a traditional rationale for empirical social science research: to better understand trends and patterns in lawmaking and law enforcement; to contrast these with legal philosophies and theories; and so forth. Besides the goal of ensuring consistent rulings [29], the rationale extends to public administration: machine learning may ease the systematic archiving of legal documents and the finding of relevant documents, and, therefore, it may also reduce administrative costs [34]. These administrative aspects reflect the goal of building "systems that assist in decision-making", whereas the predictive legal mining applications seek to build "systems that make decision" [35]. At the risk of a slight overgeneralization, it can be said that the latter systems mostly equate to supervised machine learning models, whereas the assisting systems usually operate with different, law-specific information retrieval techniques. Particularly the information retrieval techniques constitute the backbone in many legal experts systems in practical use. While the present work belongs to the predictive domain, it should be remarked that fully autonomous predictive systems are still rare in law enforcement—and remain highly controversial.

Relying on distinct argumentation styles in legal reasoning [26,36], the information retrieval systems extract and quantify textual data from legal documents into structured collections with a predefined logic and semantics [37–39]. To gain a hint about the extraction, one might consider a legal document to contain some facts, rights, obligations, and prohibitions, statements and modalities about these, and so forth. This illustration helps to understand why a concept of legal linguistics [40] is also sometimes used to describe the information retrieval approaches.

Although applications related jurisprudence are in the mainstream, it is worth noting that similar techniques have also been used to extract requirements for software and systems in order to comply with the laws from which a given extraction is done [39]. Driven by the genuine interest to facilitate collaboration between lawyers and engineers in order to build law-compliant software and systems [41], this rationale has been particularly prevalent in the contexts of data protection and privacy. For instance, previous work has been done to extract requirements from the Health Insurance Portability and Accountability Act in the United States [42]. Against this backdrop, it is no real surprise that data extraction has been applied also for laws enacted in the EU. In particular, there are various existing works on identifying requirements from the GDPR, including those based on manual inspection and user stories [6,43], ontologies and information retrieval [44,45], and formal analysis [46]. By and large, these works have concentrated on providing a better understanding of the GDPR for technical implements and their compliance. Many – but not all [47] – of these previous works also limit themselves to requirements for technical implementations, omitting the organizational requirements, such as the mandate to designate data protection officers specified in A37. As already noted, a different viewpoint is available by focusing on the administration. Thus far, furthermore, the regulation's enforcement has received only minimal attention. Apart from a few short commentaries [17,48,49], no directly comparable previous research seem to exist.

Finally, it should be emphasized that the decision documents released by the national DPAs should not be strictly equated to law-like legal documents. On one hand, the nature of these documents separates the present work from the traditional applications in the legal mining domain; on the other hand, these also enlarge the scope to which the work can be compared. For instance, highly similar machine learning and information retrieval techniques have been used to analyze privacy policies of software and systems [50,51]. Besides aligning closely with the GDPR's requirements [52], these also resemble the decision documents in that neither a universal format nor well-defined semantics exist for representing privacy policies. On that note, the dataset used should be described in more detail.

4. Materials

4.1. Dataset

The EU has not established a common database for archiving and cataloging the GDPR enforcement decisions made by the DPAs. Although the European Data Protection Board (EDPB), which supervises the national DPAs and coordinates pan-European data protection activities, has recently established a specific register for the "one-stop-shop" decisions made under A60 [53], a unified, comprehensive, and robust data source is lacking for the national decisions made and the fines imposed by the DPAs.

To patch this practical but important administrative limitation, several online data collections have recently been established

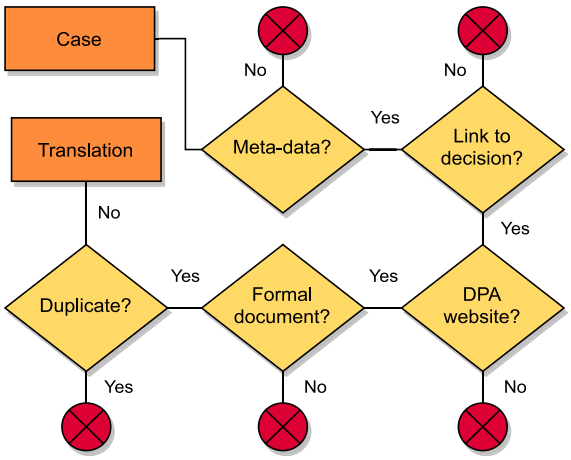

Fig. 2. Sample construction.

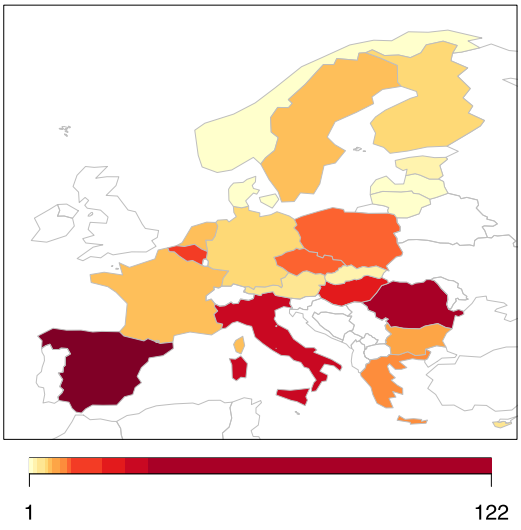

Fig. 3. Countries of origin.

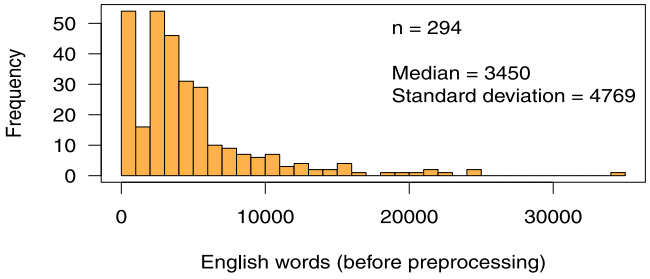

Fig. 4. Decision document lengths.

by non-governmental organizations, companies, and others [54–56]. Also the dataset for the present work is based on an online collection maintained by an international law firm for archiving many of the known GDPR enforcement cases [57]. Given that annotation and labeling are often encountered problems for unstructured collections of law-related documents [51,58], there is a simple but important benefit from using the collection: each archived enforcement case is accompanied by ready-made meta-data supplied by the firm as well as a link to the corresponding decision from a data protection authority.

4.2. Data quality

The dataset is on the small side ($n = 294$), but still sufficient for statistical inference and machine learning computations. Rather than the sample size, the downsides of the dataset collected are elsewhere. In addition to potentially missing cases due to a lack of publicly available information, the archival material is unfortunately incomplete in many respects. The reason originates from the incoherent reporting practices of the national data protection authorities. Therefore, all available cases were obtained from the online collection, but the following steps (see Fig. 2) were followed to construct a sample for the empirical analysis:

1. To maintain coherence between the three research questions, only those cases were included that had both meta-data and links to the decisions available. In terms of the former, some cases lacked meta-data about the fines imposed, the particular articles referenced in the decisions, and even links to the decisions.
2. To increase the quality of the sample, only those cases were included that were accompanied with more or less formal documents supplied on the official websites of the European data protection authorities. By implication, those cases are excluded whose archival material is based online media articles, excerpts collected from annual reports released by the authorities, and related informal or incomplete sources.
3. If two or more cases were referenced with the same decision in the online archive, only one decision document was included but the associated meta-data was unified into a single case by merging the articles references and totaling the enforcement fines imposed.
4. Following recent research [59], all national decisions written in languages other than English were translated to English with Google Translate. In general, such machine translation is necessary due to the EU-wide focus of the forthcoming empirical analysis.

Given these restrictions, the $n = 294$ cases in the sample amount to about 73% of all cases archived to the proprietary online archive at the time of the data collection (24 September, 2020). The coverage is thus good even with the exclusions. However, it should be noted that the quality of the sample is not optimal. Two points warrant a brief discussion in this regard. First, partially due to the data availability issues, the sample is not spatially balanced across Europe. As can be observed from Fig. 3, many of the enforcement fines in the sample were made in Spain, whereas the decisions of German data protection authorities are likely underrepresented in the sample. Germany is also otherwise an exception since there are multiple German DPAs operating at the state level instead of a single data protection authority at the national level.

Second, the authorities in some countries have released highly detailed and rigorous documents about their decisions, while some other authorities have opted for short press releases. Although the length of a document does not necessarily reveal its quality, in the present context, the large variance seen in Fig. 4 illustrates the lack of rigor present in some documents; when imposing fines, which may be substantial under the GDPR, a decision justified with a few thousand words does not seem optimal. It is also worth remarking that most of the documents were supplied in the portable document format (PDF) and informally signed by the authorities (of all documents retrieved, about 77.9% were PDF files; the rest are plain texts appearing on the DPAs' websites). However, scanned PDF documents had to be excluded

due to the automatic data processing. For instance, the scanned PDF documents used in Portugal were omitted (cf. Fig. 3). These data quality issues and their implications are further discussed in Section 7. For the time being, it suffices to again stress that the quality issues are related to the general administrative and political shortcomings in the EU.

4.3. Preprocessing

The textual aspects for RQ_3 are derived from the translated decisions. A conventional "bag-of-words" approach is used for extracting the features. This choice is justifiable due to the nature of the dataset. The machine-translation, which is necessary for a EU-wide analysis, largely prevents robust use of semantic approaches based on word embeddings, part-of-speech (PoS) tagging, and related techniques. Furthermore, the decision documents vary greatly across Europe in terms of style, conventions, format, and other linguistic elements. In essence, each European DPA tends to use a distinct style and convention for documenting its decisions. This variance further implies that the information retrieval techniques developed in the legal mining domain cannot be readily applied.

Nevertheless, some preprocessing is still necessary. Nine steps were used for the task. To begin with, (1) all translated decision documents were lower-cased and (2) tokenized according to white space and punctuation characters; (3) only alphabetical tokens recognized as English words were included; (4) common and custom stopwords were excluded; (5) tokens with lengths less than three characters or more than twenty characters were excluded; and (6) all tokens were lemmatized into their common English dictionary forms. A common natural language processing library [60] was used for this processing together with a common English dictionary [61]. In addition to the common stopwords supplied in the library, the twelve most frequent tokens were used as custom excluded stopwords: *data*, *article*, *personal*, *protection*, *processing*, *company*, *authority*, *regulation*, *information*, *case*, *art*, and *page*.

After these initial steps, (7) five separate corpora were constructed by using k-grams with $n = 1, \ldots, 5$. These contain sequences of adjacent lemmatized tokens; for instance, the phrase *condicio sine qua non* yields three 2-grams: *condicio sine*, *sine qua*, and *qua non*. In general, k-gram models are commonly used in text mining as these often improve predictions and ease interpretation. The legal mining domain is not an exception in this regard [28,32]. After the construction of these five corpora, (8) each one was pruned by excluding those k-grams that occurred in a given corpus only once. Finally, (9) term frequency inverse document frequency (TF–IDF) scores were calculated for the k-grams in each corpus (for the exact formula used see [62]). In general, TF–IDF is often preferred as it penalizes frequently occurring terms. It is also worth remarking that other common weighting schemes (see, e.g., [63,64]) did not notably change the empirical predictions reported.

5. Methods

Descriptive statistics are used to answer to RQ_1, and ordinary least squares (OLS) to RQ_2. Regarding the latter question, two OLS models are estimated: a restricted one in which only the articles referenced in the decisions are present, and an unrestricted one that includes rest of the meta-data. A logarithm of the enforcement fines is used as the dependent variable in both OLS regression models.

The restricted regression model equates to the conventional analysis-of-variance (ANOVA). For the unrestricted OLS model, the additional meta-data aspects include dummy variables for the following features: (i) the *year* of a given enforcement case; (ii) the *country* in which the given fine was imposed; and (iii) the *sector* of the violating organization. The last feature was constructed manually by using five categories: individuals, public sector (including associations, political parties, universities, etc.), telecommunications, private sector (excluding telecommunications), and unknown sector due to a lack of meta-data supplied in the online archive. Together with an intercept, the unrestricted model contains 55 independent variables.

The question RQ_3 requires a different strategy. The reason is sparsity: there are only 294 enforcement decisions, while each k-gram corpus contains thousands of k-grams. In fact: even after the eight preprocessing steps noted in Section 4.3, there are over 35 thousand features in each $k > 1$ corpus (see Fig. 5). Fortunately, the problem is not uncommon, and dimension reduction is the generic solution for addressing it. To this end, each corpus was further pruned with the `nearZeroVar` function available from the package [65] used for computation. It drops those features that have only one unique value, as well as those features that have a very few unique values whose frequency is large with respect to the second most common value.

Then, three common dimension reduction methods for regression analysis are used: principal component regression (PCR), partial least squares (PLS), and ridge regression. In essence, PCR uses uncorrelated linear combinations as the independent variables; PLS is otherwise similar but also the dependent variable is used for constructing the combinations. Ridge regression is based on a different principle: the dimensionality is reduced by shrinking some of the regression coefficients toward zero. All three are classical and well-documented regression methods (for summaries of the statistical background see [66] and [67]). All are also widely used in applied research [68]. In general, all three methods are further known to yield relatively similar results in applied work. Given these points, it is more relevant to proceed by elaborating the practical computation than to describe the methods themselves.

Thus, in terms of practical computation, the number of components for the PCR and PLS models, and the shrinkage parameter for the ridge regression, is optimized during the training while the results are reported with respect to a randomly selected test set containing 20% of the enforcement cases. Centering (but not scaling) is used prior to the training with a 5-fold cross-validation. Computation is carried out with the *caret* package [65] in conjunction with the *pls* [69] and *foba* [70] packages. Although root-mean-square errors (RMSEs) are used for optimization, the results are summarized with mean absolute errors (MAEs) due to their straightforward interpretability. These are defined as the arithmetic means of the absolute differences between the observed and predicted fines in the test set.

As for answering to RQ_3 in general, each of the three regression estimators is used to estimate six models (see Fig. 5). The first contains the meta-data features; the second and third models the pruned 2-gram and 3-gram features; and so on. If the meta-data model outperforms the textual feature models, at least one of the estimators should show smaller MAEs compared to the MAEs from any of the fifteen models using the k-gram features.

6. Results

6.1. Fines

The GDPR enforcement fines imposed vary greatly in the dataset. As can be seen from Fig. 6, a range from about e^6 euros to e^{12} euros capture the majority of the enforcement fines observed. This range amounts roughly from about four hundred to 163 thousand euros.

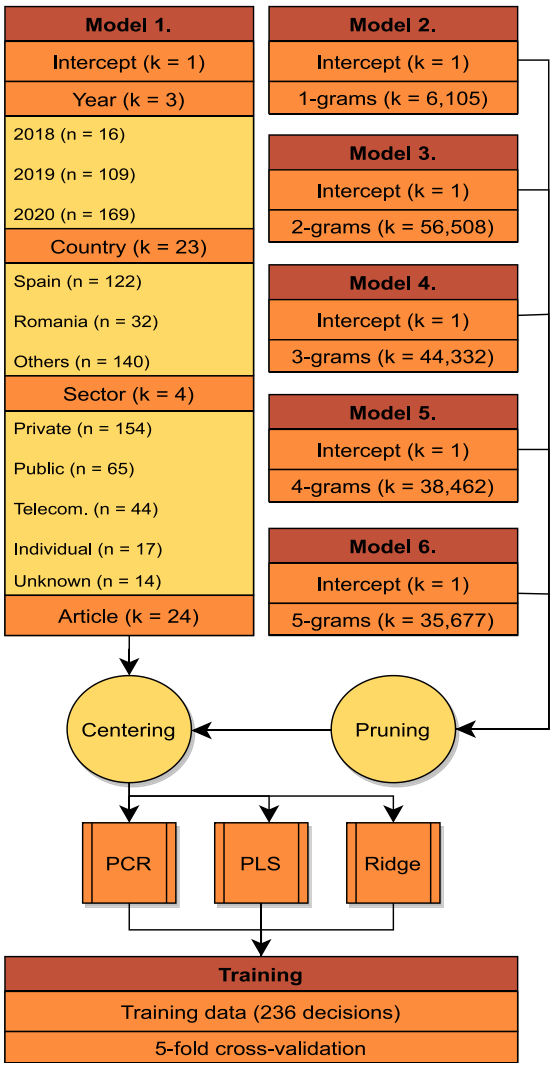

Fig. 5. Model construction.

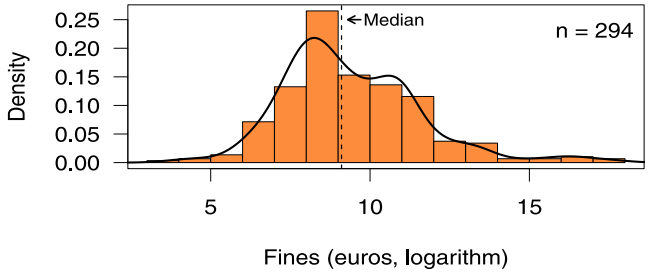

Fig. 6. Enforcement fines in the sample.

That said, the distribution has a fairly long tail; also a few large, multi-million euro fines are present in the sample. Therefore, the sample cannot be considered biased even though the restrictions discussed in Section 4.2 exclude some of the largest enforcement cases, including the announcements about the intention to fine the British Airways and Marriott International by the Information Commissioner's Office in the United Kingdom. Although these two excluded cases are (at least at the time of writing) preliminary announcements, they are still illuminating in the sense that both were about large-scale data breaches of consumer data. Given that data breaches have been estimated to

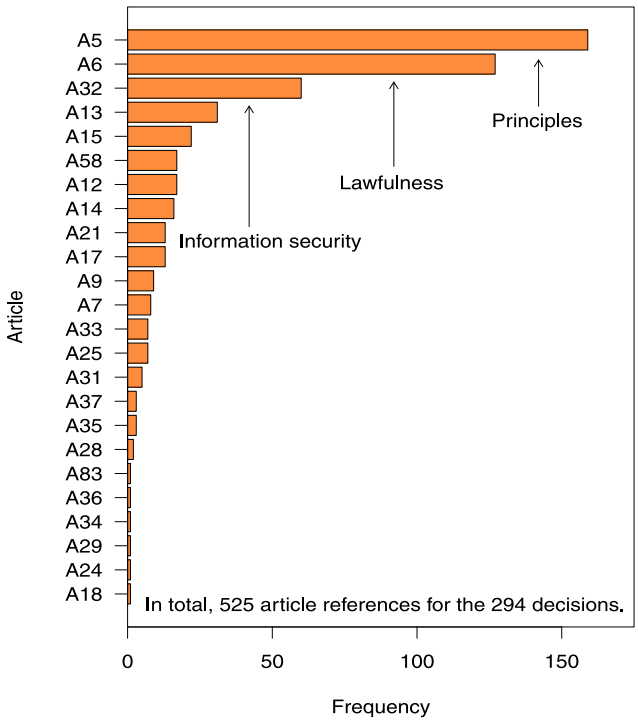

Fig. 7. Articles referenced in the enforcement cases.

cause hundreds of millions (or more) of economic and societal losses [71,72], even the few large fines in the sample are clearly on the small side. This point reinforces the earlier remarks about the enforcement problems.

6.2. Articles

Articles A5 and A6 have been the most frequently referenced ones in the enforcement decisions (see Fig. 7). This observation is not surprising; these two articles are perhaps the most fundamental ones among the ninety-nine articles laid down in the GDPR. Article A5 specify the accountability criterion and the mandate to be able to demonstrate compliance. These are fundamental practically for all software products processing with personal data [6]. Article A6, in turn, specifies the six conditions under which the lawfulness of processing personal data can be established in the EU under the GDPR. Thus, it is no real wonder that as many as 67% of the enforcement decisions have referenced either A5, A6, or both of these.

Article A32, which addresses the security of processing personal data explicitly, has been the third most frequently referenced article in the decision documents. Given that particularly the recital (f) in A5 align with A32 [5], it can be concluded that many of the decisions have dealt with data breaches and other security lapses. When taking a look at the twenty-five 2-grams with the highest TF–IDF scores, different security issues are indeed apparent; *black list*, *technical organizational* and *appropriate technical* (which both refer to A5), *security breach*, *unauthorized access*, *security measure*, *security policy*, and so forth.

In addition, many references have also been made to numerous other articles in the GDPR. As many as 31 references have been made to A13 and 22 references to A15. The former specifies the informing obligations to data subjects, whereas the latter defines the conditions under which they can access their personal data. In general, these references reflect the criticism about the non-compliance of many organizations with respect to their respect of the new rights granted to individuals [73]. Furthermore,

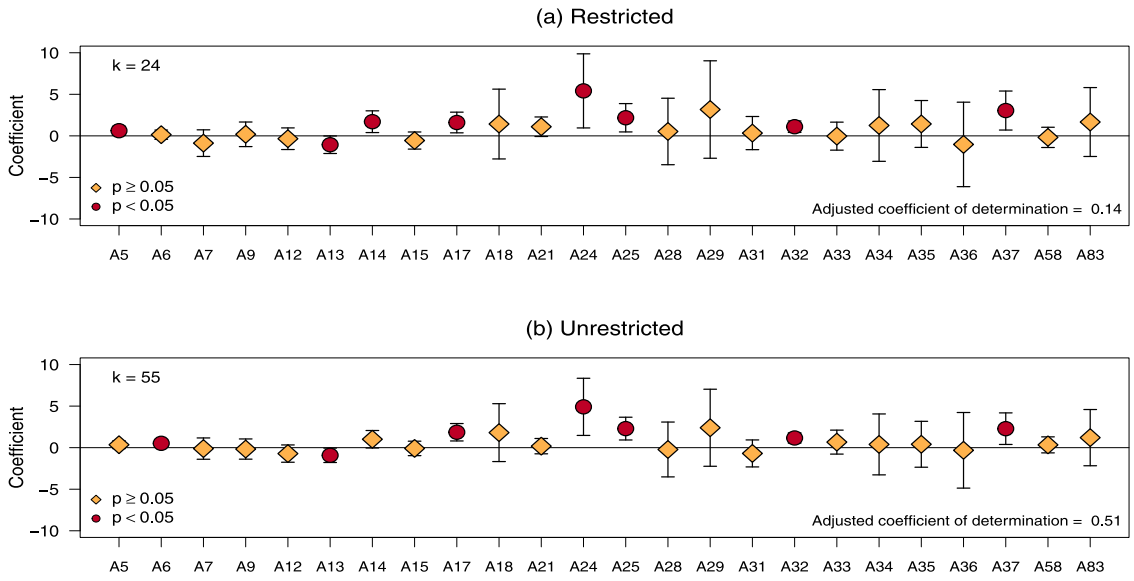

Fig. 8. Enforcement fines across articles (logarithm, OLS, 95% CIs).

more than seven references have been made to A58 (the powers granted to DPAs), A12 and A14 (transparency requirements), A21 (the right to object), A17 (the right to erasure), A9 (sensitive personal data), and A7 (the conditions for consent). Particularly the seventeen references to A58 are worth emphasizing; it seems that some organizations have been also unwilling to cooperate with public authorities. In this regard, it is further worth remarking the references made to the obligations to designate data protection officers (A37), conduct impact assessments (A35), and consult supervisory authorities (A36), to name three examples. Furthermore, less frequent references have been made in the decisions to numerous other articles. Interestingly, though, no references have been made to A22 (the right to object automatic decision-making that have legal consequences for data subjects). This observation again pinpoints toward the diverging legal interpretations in Europe [74]. But all in all, as a whole, the GDPR articles referenced hint that the regulation's full scope is slowly being enforced by the data protection authorities. Indirectly, the references to articles such as A35, A36, and A37 further hint that DPAs are also using their soft power for improving data protection. In addition to the enforcement decisions as a deterrent against poor practices, such soft power includes public relations, promotion of instructions and guidelines, raising of awareness, and other things.

Turning to the regression analysis, the OSL estimates are summarized in Fig. 8. There are three points worth making about the estimates. First, the regression coefficients are highly similar between the restricted model including only the articles and the unrestricted model containing all available meta-data. In addition to the similarity in terms of magnitude, only three coefficients differ between the two models with respect to statistical significance at the conventional level. Second, the overall performance is even surprisingly good for the unrestricted model; the adjusted R^2 is as high as 0.51. Given that the unrestricted model yields a value of 0.14, much of the performance is attributable to the other three meta-data features. Of these features, none of the dummy variables are statistically significant for the sector of an infringing party. Hence, the year of enforcement and the country of origin are particularly relevant for explaining the overall variation in the enforcement fines. This supports the earlier discussion about cross-country variation in the GDPR's enforcement and the administration of data protection in general. Third, none of the coefficients forcefully stand out in terms of their magnitudes. When looking at the coefficients with relatively tight confidence intervals (CIs), it is evident that variation is present but the magnitude of this variation is not substantial. Most of the coefficients remain in the range $[-5, 5]$. It is particularly noteworthy that in both models the coefficients for A32 (the information security requirements) are statistically significant and have a positive sign, but with only modest magnitudes. The small magnitudes apply also to A5. The observation is generally surprising, given that data breaches could be expected to yield particularly severe penalties. But according to the dataset, this expectation does not hold ground.

6.3. Predictions

The results from the cross-validated predictions are summarized in Fig. 9. It shows the mean absolute errors across the 48 models trained. These errors are small. Given that all MAEs are below e^2 euros, the predictions are generally decent enough. Another point worth remarking is that Ridge regression with 1-grams outperforms all other models. Therefore, the answer to RQ_3 is twofold: while meta-data gives decent predictions, the mechanical black-box textual features yield slightly better ones. The estimates seem acceptable also upon a close visual examination. For instance, in Fig. 10, even the outlying large fine is estimated with a reasonable error. Adding a dummy variable for it and re-estimating the models indicates only small improvements in the MAEs. But as will be soon noted in Section 7, outliers still remain a potential concern for the prediction of future enforcement fines.

The same concern can be raised also from the illustration in Fig. 11, which indicates potential problems in the training process, including the possibility of over-fitting (the MAEs for the training refer to the best cross-validated models). In other words, there are fairly large gaps in the performance between the training and test sets. Though, these gaps apply only to the PLS and Ridge regression estimators. Given that 1-grams yield the best performance also with the PCR estimator (see Fig. 9), which does not exhibit notable train–test gaps, the overall conclusion regarding RQ_3 is not threatened—for this particular dataset.

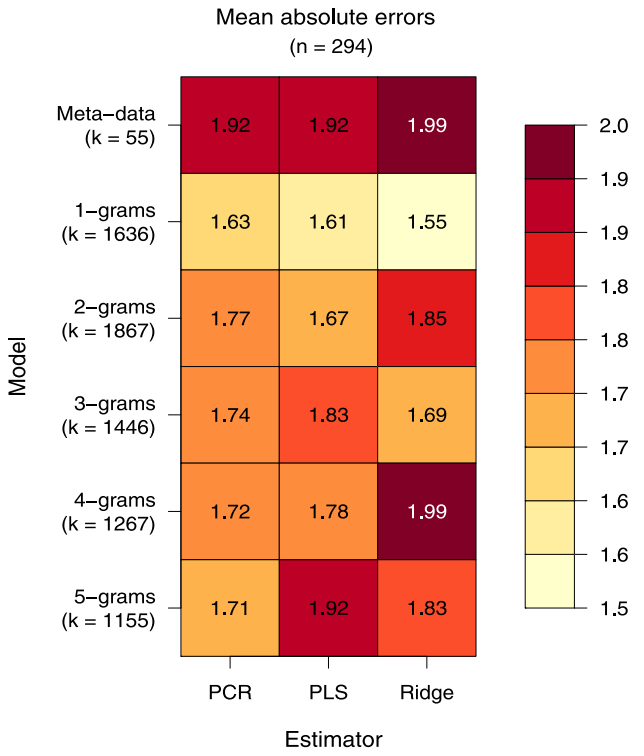

Fig. 9. Prediction performance (MAEs).

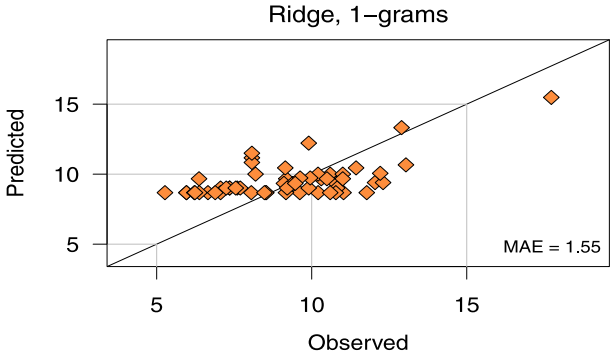

Fig. 10. Observed and predicted values in the test set.

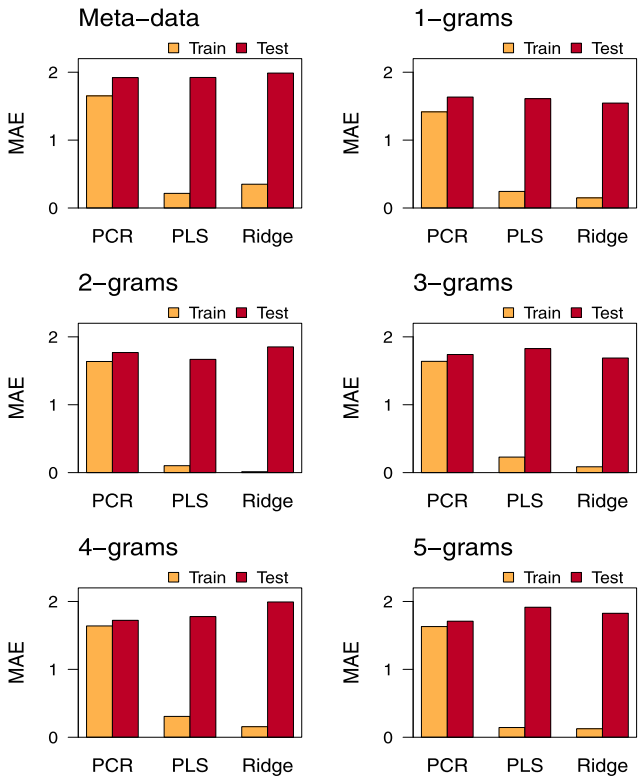

Fig. 11. Training and testing (MAEs).

7. Limitations

Some limitations should be acknowledged. To raise the generality, the discussion that follows addresses these together with points about automatic decision-making systems for judiciary. Given the discussion in Section 3, the paper aligns with the idea of systems making decisions; the results presented can therefore be seen as an output from a prototype-like automatic decision-making system.

The limitations can be further framed with the difficult concepts of reproducibility and replicability (or repeatability). These concepts are often used interchangeably. Sometimes, these are even defined in conflicting ways (see for instance [75] versus [76]) even though the intention remains the same. For the present purposes, repeatability can be defined as a "property of an experiment: the ability to repeat – or not – the experiment described in a study", whereas reproducibility is understood as "a property of the outcomes of an experiment: arriving – or not – at the same conclusions, findings, or values" [77]. With some caveats, basic repeatability should be possible by carefully following the steps described in Sections 4 and 5 . But the text mining context almost necessarily adds some caveats.[3]

Among the caveats is preprocessing. Even with the guidelines, achieving *perfect* repeatability may be difficult. But this difficulty comes from a necessity: there are tens of thousands of *k*-grams in the corpora used to answer to RQ_3, and even after preprocessing, the amounts are large and require dimension reduction methods for regression analysis. This comes at a cost for repeatability. A related caveat is the machine-translation of the decision documents to English. As a proprietary translation engine was used, it is impossible to guarantee that exactly the same results would be obtained in the future. With regard to the translations themselves, there is existing discussion about the use of Google Translate in scientific and scholarly applications [79,80]. For the present purposes, it suffices to briefly continue the discussion by noting that even small translation mistakes may have exceptionally dire consequences in judiciary [81]. However: as only conventional TF–IDF weights were used, mistakes in translation semantics are a lesser concern for the present paper. That said, furthermore, a proprietary online database was used to obtain the meta-data as well as to collect the decisions from the primary sources. Even if a repeatable code would be supplied for the data retrieval, it is impossible to guarantee that exactly same data would be retrieved due to the third-party source.

But what do these repeatability concerns imply for ADM systems used in judiciary? Clearly, in this context, nearly perfect repeatability should be guaranteed as otherwise it becomes difficult, if not impossible, to challenge and justify a system's decision. In this regard, there is an interesting recent discussion about preprocessing and dimension reduction methods in automatic decision-making systems used in judiciary: as preprocessing is

[3] In order to facilitate potential replications, the dataset used for the statistical computation is available online [78].

absurd in the legal mining domain because it may change cases extra-judicially, regularization and related methods should be used instead [82,83]. Further problems arise from the proprietary, closed source nature of most ADMs. In general, it is notoriously difficult to audit such systems [84]. And once again, the problem is not merely about auditing; it is about the use of private sector systems for public sector services [74,85,86]. Thus, the repeatability concerns are graver on the side of ADMs due to the consequences to individuals subjected to the decisions made by the systems.

Analogous concerns apply to the reproducibility of values, such as the distribution of the fines in Fig. 6. Here, the biggest concern is generalizability. Although the about 73% of decisions from the online archive could be reasonably assumed to generalize toward all decisions in this particular archive, these may not generalize toward the whole population of GDPR enforcement decisions during the period studied. Again, the problem is unavoidable because neither the national DPAs together nor the EU institutions have provided a rigorous archive for all decisions made in Europe. In short: because the statistical population remains unknown, generalizability cannot be guaranteed.

These concerns about reproducibility of values translate into potential issues in the reproducibility of findings, such as those in Fig. 7 (RQ_1) and the ANOVA results in Fig. 8 (RQ_2). The regression predictions (RQ_3) are further threatened by other issues. For instance, the dataset is not balanced across Europe (see Fig. 3). There are potential issues also with the sectoral breakdown (see Section 5) because some countries (such as Finland) have excluded public sector from the scope of A83. But all things considered, do these problems threaten the reproducibility of conclusions, the answers to the three research questions?

By argument, the answer is negative: regardless of the repeatability and reproducibility threats, a future study should find frequent references to A5 and A6 in particular (RQ_1), variance of the enforcement fines across the articles (RQ_2), and decent predictions by conventional regression methods (RQ_3). Of these conclusions, the one given to RQ_3 is the most contestable. The ADM context illustrates the issue better than the analysis presented.

In general, ADMs used in judiciary have severe problems in reorienting themselves according to changes in law and court practice [74]. Thus: if future enforcement pushes the magnitudes toward billion-euro fines, say, the predictions would be inaccurate at best and haphazard at worst. Yet the real issue is not about potential inaccuracies. Predicting the GDPR enforcement fines is as a sensible research question as any for scholarly work, but, throughout this paper, an implicit question has lingered along: should automated systems for determining fines be deployed in a society? If the answer is no, or even maybe, there should be a thorough political discussion in the society.

8. Conclusion

The following points summarize the answers reached:

1. Based on a dataset constructed via a third-party collection – which is necessary because the public administrations involved have not been able or willing to provide adequate public data, thus casting the accountability and transparency of the enforcement into a somewhat dismal light – the articles related to the general principles (A5), lawfulness (A6), and information security (A32) have been the most frequently referenced ones in the recent enforcement decisions done by the public administrations (RQ_1). The observation is not surprising. Article A5 is a go-to article due to its explicit responsibility dictate, and each one who computes with personal data must satisfy one of the legal basis in Article A6 (with few exceptions).
2. However, the enforcement fines are not forcefully larger or smaller for the decisions referencing A5, A6, and A32. Particularly A32 is surprising in this regard considering the harms caused by data breaches. But, in general, the enforcement fines do vary across the articles referenced (RQ_2). That said, a slightly stronger statistical explanation is available by knowing the year of enforcement and the country in which a public authority making a decision is located. This statistical result reinforces the discussion in Section 2 on the administrative problems in enforcing the GDPR.
3. Both the meta-data available from the third-party and the textual features extracted from the enforcement documents released by the responsible public authorities provide enough material for decent predictions of the enforcement fines (RQ_3). The textual features seem to outperform the meta-data, suggesting the plausibility of using a black-box predictive system for foresight. The average error is less than ten euros.

There are a couple of prolific paths for further research. Besides potential reproduction of the conclusions, first, it seems reasonable to argue that future research should focus on providing a more nuanced analysis of the enforcement decisions. A better understanding on the logic and arguments used in the decision documents is necessary for moving forward with the domain of legal linguistics described in Section 3. Eventually, it may be possible also in the GDPR context to build systems that assist in decision-making—even if actual fine-imposing ADMs are seen as unachievable or undesirable. The second path follows. By a fine-grained analysis, it should be also possible to establish implicit compliance frameworks for implementations. Fundamentally—like with all laws, the GDPR's enforcement should not depend on sanctions but on acquiescence.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: There is a conflict of interest with respect to any researcher funded by the grant no. 327391 of the Strategic Research Council at the Academy of Finland.

Acknowledgments

This research was funded by the Strategic Research Council at the Academy of Finland (grant no. 327391).

References

[1] D.H. Flaherty, Governmental surveillance and bureaucratic accountability: Data protection agencies in western societies, Sci. Technol. Hum. Values 11 (1) (1986) 7–18.
[2] J. Ruohonen, K. Hjerppe, Predicting the amount of GDPR fines, in: Proceedings of the First International Workshop CAiSE for Legal Documents (COUrT 2020), CEUR-WS, Grenoble (online), 2020, pp. 3–14.
[3] D. Erdos, Statutory regulation of professional journalism under European data protection: Down but not out? J. Media Law 8 (2) (2016) 229–265.
[4] G.G. Fuster, The Emergence of Personal Data Protection As a Fundamental Right of the EU, Springer, Cham, 2014.
[5] E. Ventrella, The symbiotic relationship between privacy and security in the contet of the general data protection regulation, ERA Forum 20 (2020) 455–469.
[6] K. Hjerppe, J. Ruohonen, V. Leppänen, The general data protection regulation: Requirements, architectures, and constraints, in: Proceedings of the 27th IEEE International Requirements Engineering Conference (RE 2019), IEEE, Jeju Island, 2019, pp. 265–275.
[7] S. Shastri, M. Wasserman, V. Chidambaram, GDPR anti-patterns, Commun. ACM 64 (2) (2021) 59–65.

[8] J. Ruohonen, A review of product safety regulations in the European Union, 2021, Archived manuscript, available online: https://arxiv.org/abs/2102.03679.
[9] T. Dalenius, Data protection legislation in Sweden: A statistician's perspective, J. R. Stat. Soc. A (General) 142 (3) (1979) 285–298.
[10] P. Hustinx, The role of data protection authorities, in: S. Gutwirth, Y. Poullet, P. De Hert, C. de Terwangne, S. Nouwt (Eds.), Reinventing Data Protection?, Springer, Cham, 2002, pp. 131–137.
[11] N.N. Neto, S.E. Madnick, Developing a global data breach database and the challenges encountered, J. Data Inf. Qual. 13 (1) (2021) 1–33.
[12] C.J. Bennett, C.D. Raab, Revisiting the governance of privacy: Contemporary policy instruments in global perspective, Regulation & Governance (2018) (Published online in September).
[13] B. Custers, F. Dechesne, A.M. Sears, T. Tani, S. van der Hof, A comparison of data protection legislation and policies across the EU, Comput. Law Secur. Rev. 34 (2) (2018) 234–243.
[14] European Commission, Communication from the Commission to the European Parliament and the Council: Data Protection as a Pillar of Citizens' Empowerment and the EU's Approach to the Digital Transition – Two Years of Application of the General Data Protection Regulation. COM(2020) 264 Final, 2020, available online in 2020: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0264.
[15] J. Ryan, Europe's Governments are Failing the GDPR: Brave's 2020 Report on the Enforcement Capacity of Data Protection Authorities, 2020, Available online in 2020: https://brave.com/wp-content/uploads/2020/04/Brave-2020-DPA-Report.pdf.
[16] F. Casarosa, Transnational collective actions for cross-border data protection violations, Internet Policy Rev. 9 (3) (2020).
[17] R.L.P. Mahieu, J. Ausloos, Harnessing the collective potential of GDPR access rights: Towards an ecology of transparency, Internet Policy Rev. (Opinion) (2020) available online in August: https://policyreview.info/articles/news/harnessing-collective-potential-gdpr-access-rights-towards-ecology-transparency/1487.
[18] K. Yesilkagit, Institutional compliance, European networks of regulation and the bureaucratic autonomy of national regulatory authorities, J. Eur. Publ. Policy 18 (7) (2011) 962–979.
[19] G. Pearce, N. Platten, Achieving personal data protection in the European union, J. Common Market Stud. 36 (4) (1998).
[20] A.-S. Lind, J. Reichel, Administrating data protection – or the fort knox of the European composite administration, Kritische Vierteljahresschrift Für Gesetzgebung Und Rechtswissenschaft 97 (1) (2014) 44–57.
[21] J. Ruohonen, An acid test for Europeanization: Public cyber security procurement in the European Union, Eur. J. Secur. Res. 5 (2) (2020) 349–377.
[22] J. Ruohonen, S. Hyrynsalmi, V. Leppänen, An outlook on the institutional evolution of the European union cyber security apparatus, Gov. Inf. Q. 33 (4) (2016) 746–756.
[23] A. Mantelero, G. Vaciago, M.S. Esposito, N. Monte, The common EU approach to personal data and cybersecurity regulation, Int. J. Law Inf. Technol. (2021) 1–33, (Published online in January).
[24] D. Wicki-Birchler, The budapest convention and the general data protection regulation: Acting in concert to curb cybercrime? Int. Cybersecur. Law Rev. (2020) 1–10, (Published online in September).
[25] P. Sterlini, F. Massacci, N. Kadenko, T. Fiebig, M. van Eeten, Governance challenges for European cybersecurity policies: Stakeholder views, IEEE Secur. Privacy 18 (1) (2020) 46–54.
[26] A. Dyevre, W. Wijtvliet, N. Lampach, The future of European legal scholarship: Empirical jurisprudence, Maastricht J. Eur. Comp. Law 26 (3) (2019) 348–371.
[27] P. Leith, The rise and fall of the legal expert system. International review of law, Comput. Technol. 30 (3) (2016) 94–106.
[28] C.I. Hausladen, M.H. Schubert, E. Ash, Text classification of ideological direction in judicial opinions, Int. Rev. Law Econ. 62 (2020) 105903.
[29] R. Wang, Legal technology in contemporary USA and China, Comput. Law Secur. Rev. 39 (2020) 105459.
[30] C. Calomme, Why Open Legal Data and Analytics are Not Without Risks, Centre for IT & IP Law (CiTiP) Blog, KU Leuven, 2020, available online in April https://www.law.kuleuven.be/citip/blog/why-open-legal-data-and-analytics-are-not-without-risks/.
[31] Z. Liu, H. Chen, A predictive performance comparison of machine learning models for judicial cases, in: Proceedings of the IEEE Symposium Series on Computational Intelligence (SSCI 2017), IEEE, Honolulu, 2017, pp. 1–6.
[32] M. Medvedeva, M. Vols, M. Wieling, Using machine learning to predict decisions of the European court of human rights, Artif. Intell. Law 28 (2019) 237–266.
[33] K. Moodley, P.V.H. Serrano, G. van Dijck, M. Dumontier, Similarity and relevance of court decisions: A computational study on CJEU cases, in: M. Araszkiewicz, V. Rodríguez-Doncel (Eds.), Legal Knowledge and Information Systems – JURIX 2019: The Thirty-Second Annual Conference, IOS Press, Madrid, 2019, pp. 63–72.
[34] R. Chhatwal, N. Huber-Fliflet, R. Keeling, J. Zhang, H. Zhao, Empirical evaluations of active learning strategies in legal document review, in: Proceedings of the IEEE International Conference on Big Data (Big Data 2017), IEEE, Boston, 2017, pp. 1428–1437.
[35] E. Nissan, Computer tools and techniques for lawyers and the judiciary, Cybern. Syst. 49 (4) (2018) 201–233.
[36] K. Atkinson, T. Bench-Capon, D. Bollegala, Explanation in AI and law: Past, present and future, Artificial Intelligence 289 (2020) 103387.
[37] H. Bhuiyan, F. Olivieri, G. Governatori, M.B. Islam, A. Bond, A. Rakotonirainy, A methodology for encoding regulatory rules, in: Proceedings of the MINing and REAsoning with Legal Texts (MIREL 2019), CEUR-WS, Madrid, 2019, pp. 1–13.
[38] N. Holzenberger, A. Blair-Stanek, B. van Durme, A dataset for statutory reasoning in tax law entailment and question answering, in: Proceedings of the Natural Legal Language Processing Workshop (NLLP 2020), CEUR-WS, 2020, pp. 31–38.
[39] A. Sleimi, M. Ceci, N. Sannier, M. Sabetzadeh, L. Briand, J. Dann, A query system for extracting requirements-related information from legal texts, in: Proceedings of the IEEE 27th International Requirements Engineering Conference (RE 2019), IEEE, Jeju Island, 2019, pp. 319–329.
[40] F. Vogel, H. Hamann, I. Gauer, Legal linguistics: Corpus analysis as a new tool for legal studies, Law Soc. Inquiry 43 (4) (2018) 1340–1363.
[41] N. van Dijk, A. Tanas, K. Rommetveit, C. Raab, Right engineering? The redesign of privacy and personal data protection, Int. Rev. Law Comput. Technol. 32 (2–3) (2018) 230–256.
[42] T.D. Breaux, M.W. Vail, A.I. Anton, Towards regulatory compliance: Extracting rights and obligations to align requirements with regulations, in: Proceedings of the 14th IEEE International Requirements Engineering Conference (RE 2006), IEEE, Minneapolis, 2006, pp. 49–58.
[43] C. Bartolini, S. Daoudagh, G. Lenzini, E. Marchetti, GDPR-based user stories in the access control perspective, in: M. Piattini, P.R. da Cunha, I.G.R. de Guzmán, R. Pérez-Castillo (Eds.), Proceedings of the 12th International Conference on the Quality of Information and Communications Technology (QUATIC 2019), Springer, Ciudad Real, 2019, pp. 3–17.
[44] M. Palmirani, M. Martoni, A. Rossi, C. Bartolini, L. Robaldo, Legal ontology for modelling GDPR concepts and norms, in: M. Palmirani (Ed.), Legal Knowledge and Information Systems – JURIX 2018: The Thirty-First Annual Conference, IOS Press, Groningen, 2018, pp. 91–100.
[45] D.A. Tamburri, Design principles for the general data protection regulation (GDPR): A formal concept analysis and its evaluation, Inf. Syst. 91 (2020) 101469.
[46] E. Arfelt, D. Basin, S. Debois, Monitoring the GDPR, in: K. Sako, S. Schneider, P.Y.A. Ryan (Eds.), Proceedings of the 24th European Symposium on Research in Computer Security (ESORICS 2019), in: Lecture Notes in Computer Science, vol. 11735, Springer, Luxembourg, 2019, pp. 681–699.
[47] C. Meurisch, M. Mühlhäuser, Data protection in AI services: A survey, ACM Comput. Surv. 54 (2) (2021) 40:1–40:38.
[48] C. Barrett, Emerging trends from the first year of EU GDPR enforcement, Scitech Lawyer 16 (3) (2020) 22–25.
[49] A. Erickson, Comparative analysis of the EU's GDPR and Brazil's LGPD: Enforcement challenges with the LGPD, Brooklyn J. Int. Law 44 (2) (2019) 859–888.
[50] H. Harkous, K. Fawaz, R. Lebret, F. Schaub, K.G. Shin, K. Aberer, Polisis: Automated analysis and presentation of privacy policies using deep learning, in: Proceedings of the 27th USENIX Security Symposium (USENIX Security 2018), USENIX, Baltimore, 2018, pp. 531–548.
[51] M. Lippi, P. Pał ka, G. Contissa, F. Lagioia, H.-W. Micklitz, G. Sartor, P. Torroni, CLAUDETTE: An automated detector of potentially unfair clauses in online terms of service, Artif. Intell. Law 27 (2019) 117–139.
[52] K. Hjerppe, J. Ruohonen, V. Leppänen, Extracting layered privacy policy purposes from web services, in: Proceedings of the International Workshop on Privacy Engineering (IWPE 2020), IEEE, Genova (online), 2020.
[53] European Data Protection Board, Register of art. 60 final decisions, 2020, Available online in August: https://edpb.europa.eu/our-work-tools/consistency-findings/register-for-article-60-final-decisions_en.
[54] International network of privacy law professionals (NPLP), 2020, GDPR Fines Database. Available online in August: https://gdpr-fines.inplp.com/.
[55] noyb, et al., GDPRhub, 2020, Available online in August: https://gdprhub.eu/index.php.
[56] PrivacyAffairs, GDPR fines tracker & statistics, 2020, Available online in August: https://www.privacyaffairs.com/gdpr-fines/.
[57] CMS Law.Tax, GDPR Enforcement Tracker, 2020, Data obtained in 24 February from: https://enforcementtracker.com/.
[58] S. Sharafat, Z. Nasar, S.W. Jaffry, Data mining for smart legal systems, Comput. Electr. Eng. 78 (2019) 328–342.
[59] J. Ruohonen, A dip into a deep well: Online political advertisements, valence, and European electoral campaigning, in: The Proceedings of the 2nd Multidisciplinary International Symposium on Disinformation in Open Online Media (MISDOOM 2020), Springer, Leiden (online), 2020.

[60] The Natural Language Toolkit (NLTK), The natural language toolkit (NLTK). Version 3.5, 2020, Available online in September: http://www.nltk.org.
[61] L. Németh, K. Hendricks, C. McNamara, et al., Hunspell, version 1.7.0, 2020, available online in February https://github.com/hunspell/hunspell.
[62] J. Ruohonen, V. Leppänen, Toward validation of textual information retrieval techniques for software weaknesses, in: M. Elloumi, M. Granitzer, A. Hameurlain, C. Seifert, B. Stein, A.M. Tjoa, R. Wagner (Eds.), Proceedings of the 29th International Conference on Database and Expert Systems Applications (DEXA 2018), in: Communications in Computer and Information Science, vol. 903, Springer, Regensburg, 2018, pp. 265–277.
[63] H. Fang, T.T.C. Zhai, A formal study of information retrieval heuristics, in: Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2004), ACM, Sheffield, 2004, pp. 49–56.
[64] R. Jin, J.Y. Chai, L. Si, Learn to weight terms in information retrieval using category information, in: Proceedings of the 22nd International Conference on Machine Learning (ICML 2005), ACM, Bonn, 2005, pp. 353–360.
[65] M. Kuhn, et al., Caret: Classification and Regression Training. R package version 6.0-85, 2020, available online in February: https://cran.r-project.org/web/packages/caret/.
[66] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer, New York, 2011.
[67] H.A.L. Kiers, A.K. Smilde, A comparison of various methods for multivariate regression with highly collinear variables, Stat. Methods Appl. 16 (2007) 193–228.
[68] B. Hemmateenejad, M. Akhond, F. Samari, A comparative study between PCR and PLS in simultaneous pectrophotometric determination of diphenylamine, aniline, and phenol: Effect of wavelength selection, Spectrochim. Acta A 67 (2007) 958–965.
[69] B.-H. Mevik, R. Wehrens, The pls package: Principal component and partial least squares regression in R, J. Stat. Softw. 18 (2) (2007) 1–23.
[70] T. Zhang, foba: Greedy Variable Selection. R package version 0.1, 2008, available online in February: https://cran.r-project.org/web/packages/foba/.
[71] B. Edwards, S. Hofmeyr, S. Forrest, Hype and heavy tails: A closer look at data breaches, J. Cybersecurity 2 (1) (2016) 3–14.
[72] O.I. Poyraz, M. Canan, M. McShane, C.A. Pinto, T.S. Cotter, Cyber assets at risk: Monetary impact of U.S. personally identifiable information mega data breaches, The Geneva Papers on Risk and Insurance – Issues and Practice 45 (2020) 616–638.
[73] R.L.P. Mahieu, H. Asghari, M. van Eeten, Collectively exercising the right of access: Individual effort, societal effect, Internet Policy Rev. 7 (3) (2018).
[74] M. Suksi, Administrative due process when using automated decision-making in public administration: Some notes from a finnish perspective, Artif. Intell. Law (2020) (Published online in May).
[75] A. Repar, M. Martinc, S. Pollak, Reproduction, replication, analysis and adaptation of a term alignment approach, Lang. Resour. Eval. 54 (2020) 767–800.
[76] J. Ruohonen, S. Hyrynsalmi, V. Leppänen, The sigmoidal growth of operating system security vulnerabilities: An empirical revisit, Comput. Secur. 55 (2015) 1–20.
[77] K.B. Cohen, J. Xia, P. Zweigenbaum, T. Callahan, O. Hargraves, F. Goss, N. Ide, A. Névéol, C. Grouin, L.E. Hunter, Three dimensions of reproducibility in natural language processing, in: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Miyazaki, 2018, pp. 156–165.
[78] J. Ruohonen, K. Hjerppe, The GDPR enforcement fines at glance: Dataset, 2021, Available online in July: https://doi.org/10.7910/DVN/DQ38VC.
[79] F. Daniele, Performance of an automatic translator in translating medical abstracts, Heliyon 5 (10) (2019) e02687.
[80] M. Groves, K. Mundt, Friend or foe? Google translate in language for academic purposes, Engl. Spec. Purposes 37 (2015) 112–121.
[81] J. Scott, J. O'Shea, How legal documents translated outside institutions affect lives, businesses and the economy, Int. J. Semiotics Law (2021) 1–43, Published online in February.
[82] A. Bibal, M. Lognoul, A. de Streel, B. Frénay, Legal requirements on explainability in machine learning, Artif. Intell. Law (2020) 1–21, (Published online in July).
[83] L. Boswell, A. Prakash, On the fairness of 'fake' data in legal AI. Archived manuscript, 2020, available online in September: https://arxiv.org/abs/2009.04640.
[84] B. Waltl, R. Vogl, Increasing transparency in algorithmic-decision-making with explainable AI, Datenschutz Und Datensicherheit – DuD 42 (2018) 613–617.
[85] T. Kerikmäe, E. Pärn-Lee, Legal Dilemmas of Estonian Artificial Intelligence Strategy: In Between of E-SOciety and Global Race, AI & SOCIETY, 2020, (Published online in July).
[86] M. Kuziemskia, G. Misuraca, AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings, Telecommun. Policy 44 (6) (2020) 101976.