Available online at www.sciencedirect.com

ScienceDirect

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

Computer Law & Security Review

Pairing EU directives and their national implementing measures: A dataset for semantic search

Roger Ferrod[a], *Denys Amore Bondarenko*[a], *Davide Audrito*[b,*], *Giovanni Siragusa*[a]

[a] *Department of Computer Science, University of Torino, Corso Svizzera 185 - 10149 Torino, Italy*
[b] *Legal Studies Department, University of Bologna, Via Zamboni 27 - 40126 Bologna, Italy*

ARTICLE INFO

Keywords:
Legal harmonization
EU legislation
Domestic implementation
Search engine
Dataset

ABSTRACT

European Directives (EUDs) are binding upon Member States as to the results to be achieved, but leave to national authorities the choice of form and methods. Therefore, Member States adopt *ad hoc* National Implementing Measures (NIMs) that mostly reproduce the contents of EUDs and transpose them into domestic legislation. This well-known process is defined as "legal harmonization" and consists of the gradual, although ambiguous, approximation of national legal orders as a result of the adoption of European legislation. In order to contribute to the analysis of this phenomenon, we collect a large and unique dataset composed of European and domestic legislative sources, which is an essential requirement to automatically pair EUDs and the corresponding NIMs, in light of their semantic similarity. The first results show the feasibility of the proposed task to discern NIMs from national legislation that does not contribute to implementing EUDs, thus constituting the foundation for a semantic search engine. We believe that our effort can promote future applications and research directions, with the ultimate aim to support traditional legal methodology, facilitate citizens' access to rights, support public administrations, and, more in general, promote democracy and the rule of law in the European Union. Data and source code are available at https://doi.org/10.17632/mkx5sb3mnw.

1. Introduction

The aim of this work is to introduce a novel semantic search task designed to automatically identify National Implementing Measures (NIMs) of European Directives (EUDs), which – by nature – require the adoption of *ad hoc* domestic legislation to implement their provisions. The first step of this research effort is the release of a new multilingual dataset that collects EUDs and NIMs of five different Member States, accompanied by some baselines that introduce the ranking task, i.e. the ability to automatically pair EU Directives provisions and their corresponding National Implementing Measures.

This research is inspired by the most advanced technologies for semantic search and in particular by the use of deep learning for dense text retrieval. The state of the art in this field is represented by transformer-based word embeddings and contrastive learning (Xiong et al., 2021; Zhan et al., 2021), which aim to bring semantically similar examples closer and, at the same time, dissimilar examples far from each other. We

* Corresponding author at: Legal Studies Department, University of Bologna, Via Zamboni 27 - 40126 Bologna, Italy.
E-mail address: davide.audrito2@unibo.it (D. Audrito).

https://doi.org/10.1016/j.clsr.2023.105862
0267-3649/

argue that this approach could be applied to the legal domain, as there is a higher similarity between EU norms and their corresponding national legislation. This hypothesis, which lays the foundation of Sulis et al. (2022), was proved by our baselines as well.

We intend to support this line of research by releasing a wide collection of texts specifically designed for deep learning and contrastive learning. These resources include European norms and their implementations (i.e. "positive examples"), as well as domestic laws that do not contribute to transposing EUDs and can be consequently seen as "negative examples".

The dataset for this study includes all the European Directives that are currently in force, which were downloaded in their English, French, German, Spanish and Italian versions. Extracting national legislation was far more complex for three main reasons: 1) it required a detailed comparison of legislative sources to be included in the dataset, well beyond their mere presence in the transposition tables; 2) multilingual versions of laws required additional linguistic and semantic analysis; 3) the availability of national Official Journals in different formats and the large number of legislative sources made data extraction challenging.

National laws were then linked to the directives that they implement in an unsupervised approach, as indicated in the transposition tables on the EUR-Lex website.[1] To avoid the negative impact of irrelevant information on the performance of the model, we developed a filter that uses an IDF-inspired and similarity-based function to score the value of each article, as described more in details in Section 4.2. Indeed, both European and national laws commonly include articles that are semantically independent of the subject of the document. These provisions typically concern the date of entry into force, recipients, maximum penalties for infringements, and other similar details that are mostly common to different legislative sources with only minor variations.

Moreover, to provide a benchmark for comparison, we included two baselines in our study. The first is based on the BM25 ranking function to match EUDs and NIMs; the second employs Sentence Transformers models to encode articles, followed by dense similarity search.

Our contribution is manifold for two main reasons. 1) This methodology could assist competent authorities in reporting applicable national legislative measures that were adopted to fulfill obligations arising from EU directives. Similarly, legal practitioners would be helped in their comparative analysis of national frameworks and in their research on legislative harmonization through the adoption of EU legislation. 2) This research is relevant to boost domestic implementation of EUDs, considering that citizens have the right to lodge complaints before national courts and the Court of Justice of the European Union (CJEU) in case of failure to fulfill obligations. In other words, this search engine proposal could support citizens to check the effective transposition of EU norms into domestic legal orders. For the same reason, such a tool would have a supportive role in judicial interpretation and decision-making in the proceedings in which the CJEU reviews the consistency of domestic legislation with EU law.

Similar objectives have been pursued by previous research in the area of legal informatics and – manually – by legal experts, in the framework of EU-funded projects. This notwithstanding, the latter method is extremely time-consuming although able to deliver excellent results.

The dataset and the baselines made available in this article stand as a starting point for further research in the legal area, as they combine the most advanced and successful technologies for semantic search with multilingual legislation.

2. Background

Historically, there are two main reasons for the foundation of the European Union. First, after the end of World War II, it became necessary to establish an economic and political cooperation framework in order to prevent States' nationalism and imperialism resulting in their inability to expand their trade flows, which was at the origin of World War I (Formigoni, 2018). After 1945, economic nationalism and self-sufficient ambitions were indeed mostly set aside in the EU. Second, the establishment of a common internal market was functional to the economic interests of largest banks, investment funds, big companies and to the recovery of the EU economy, which was in a deplorable condition.

Essentially, the EU was grounded upon a regional legal order of sovereign States that jointly abide by common legislation and fundamental principles that have a constitutional nature, i.e. the rule of law, solidarity, freedom, democracy and equality. Over the decades, the founding treaties, the EU legislation and the case law of the European Court of Justice fostered the creation of a common legal framework across Member States' borders and legal orders.

The establishment of the EU internal market was the milestone of legal harmonization and uniformity processes and one of the most significant achievements of the EU, which allows citizens and businesses to pursue their fundamental freedoms, including the free movement of goods and services (Baaij, 2012). Furthermore, Articles 114 and 115 of the Treaty on the Functioning of the European Union (TFEU), which enables the EU to adopt measures on the functioning of the internal market, were invoked to pursue harmonization in many other European legal and policy areas, including a common legal framework for national contract laws (Baaij, 2012). This trend towards harmonization and uniformity concerns the legal areas in which the EU has been conferred legislative competence, e.g. internal market, consumer rights, environment, labor, mobility and migration.

2.1. *Legal harmonization in the European union*

The effectiveness of harmonization is linked to a variety of drivers, including the nature of legal instruments, the competence conferred upon the EU, multilingualism and the vagueness of language. In this regard, the Court of Justice of the European Union (CJEU) has enshrined the so-called "doctrine of

[1] EUR-Lex is the main repository of EU official sources, which aims to ensure citizens' right to be informed about EU issues. The platform enforces the principles codified in the Charter of Fundamental Rights of the European Union (CFREU), namely the right of access to the documents of the EU institutions according to Article 42, but also Articles 11, 41 and 44 thereof.

consistent interpretation", which was defined as "the obligation of national courts and administrative authorities to interpret the applicable national law as much as possible in a way that ensures the fulfillment of obligations derived from European law"(Prechal, 2007). This notwithstanding, common legislation does not always result in effective harmonization. For this reason, some scholars argue that achieving full linguistic and legal harmonization in the EU is a myth (Šarčević, 2016).

Among other parameters, multilingualism has a relevant role in the CJEU judicial decision-making and represents a research challenge also for Natural Language Processing (NLP) and Machine Learning (ML). Multilingual versions of EU legal acts directly affect legal interpretation at the Court, because a relatively new "hermeneutic paradigm" replaces traditional interpretation methods, including literal and systematic interpretation (Łachacz and Mańko, 2013). More generally, "the EU legal culture emerges through translation as a hybrid supranational pan-European construct with mutual dependencies on national legal cultures" (Sosoni and Biel, 2018). This influence of national legal traditions hinders the authoritativeness of the Court of Justice of the European Union. The Court is increasingly accused of going beyond its powers, including in the use of the comparative law method of interpretation (Łachacz and Mańko, 2013). The comparative law method has been rising over the most recent years and justifies the need to compare definitions and legal concepts under different national legal orders, possibly through computational approaches (Audrito et al., 2022). The balance between consensus and pursued objectives is the first reason to engage with the comparative law method, which results in the ability of the CJEU to promote a judicial dialogue with national courts by taking into account their decisions.

First, once the Court embraces the interpretation of domestic tribunals, the effectiveness of EU law will increase, because national judicial and administrative authorities will be more likely to recognize the decision of the CJEU. Second, the comparative approach can give rise to a constructive interaction between the EU and the domestic legal orders, considering that engagement among judiciaries can result in the adoption of a certain interpretation rather than another. Third, the comparative law method, by promoting "value diversity" fosters the enforcement of Article 4(2) TEU, pursuant to which the EU is to respect the national identities of the Member States. In conclusion, the comparative law method may be relied upon to clarify EU law provisions within the so-called "federal common law-making" (Lenaerts and Gutman, 2006).

2.2. *Directives and regulations as sources of harmonization*

As mentioned in the introduction, EUDs represent the official legislative source to pursue legal harmonization in the European Union. According to Article 288(3) (TFEU), directives are binding upon each Member State to which it is addressed as to the result to be achieved, but leave to Member States the choice of form and methods. In other words, EU legislators approximate the national legislative framework by adopting EUDs, which obligate MSs as to the results to be achieved. However, the form and methods used to achieve such objectives are left to the discretion of Member States' legislatures and governments.

Over the last decades, the European Commission (EC), the European Parliament (EP) and the Council have increasingly preferred the adoption of Regulations (EURs) instead of EUDs, since they are directly applicable throughout the EU, do not require further transposing legislation and uniform domestic legal orders. Nevertheless, regulations are not always able to prevent the occurrence of domestic peculiar regulatory frameworks in harmonized and unified legal domains, such as in the case of the General Data Protection Regulation (GDPR).

The GDPR has marked a meaningful shift in the creation of a European common legislative standard on the protection of personal data. However, Member States have followed different approaches to comply with the GDPR and this resulted in the occurrence of peculiarities in the national legal frameworks on the protection of personal data. In light of the foregoing, both EUDs and EURs should be considered to analyze the harmonization of laws in the EU, although in point of law EU Treaties merely rely on EUDs to enhance approximation of domestic legislative frameworks.

3. Related works

Several research works have proposed the application of Natural Language Processing (NLP) methods to legal text in order to accomplish different goals, such as identifying legislative documents (Boella et al., 2012), extracting named entities from the text (Nanda et al., 2017a) or predicting judicial decisions (Aletras et al., 2016; Medvedeva et al., 2021). For instance, Humphreys et al. (2015) developed a system to map recitals to legal provisions via a cosine similarity score over TF-IDF vectors; results indicate that their system achieves high accuracy as a result of the large number of true negatives present in their unbalanced dataset. Mandal et al. (2017) implemented several models for document similarity in order to identify similar court cases from the Indian Supreme Court. Siragusa et al. (2021) proposed machine learning models for Legal Textual Entailment, focusing on standard procedures that a company has to implement to protect their data. Such procedures are generally defined by ISO (the International Organization for Standardization) and their more specific counterpart NIST (National Institute of Standards and Technology).

In the context of legal harmonization analysis, our work pursues the same objective of Nanda et al. (2020, 2019, 2017b), i.e. identifying domestic transpositions of European Directives (EUDs). In Nanda et al. (2019), the authors defined a gold dataset of 43 EUDs and their corresponding National Implementing Measures (NIMs) in three different languages (Italian, English and French). In Nanda et al. (2020), the authors decided to explore the hidden links, i.e. links that are not explicitly referred to within the text, present in the EUDs and NIMs. To accomplish their goal, they hired legal experts to annotate a subset composed of 5 EUDs and their corresponding NIMs from the original corpus of Nanda et al. (2019).

Differently from them – and given the scope of our dataset – we excluded the hidden links as in Nanda et al. (2020) because of their challenging identification and dissimilar nature. Furthermore, since our objective is to create a scalable and un-

supervised dataset, we avoided the paragraph-by-paragraph alignment in contrast to Nanda et al. (2019), as it would require a manual annotation effort.

Other projects coherent with our research purposes are CrossJustice[2] and Facilex.[3] CrossJustice is a project developed under the "European Justice Programme" between 2019 and 2021. It aims to examine whether EU citizens can exercise their rights under the six European Directives on procedural safeguards for persons accused or suspected of a crime, providing an external monitoring of the implementation of EU legislation. In the project, the level and the adequacy of the implementation of EUDs is evaluated manually by legal experts.

Facilex follows the CrossJustice project, allowing judicial cooperation scenarios and automatic advice to legal practitioners, through a free-access online platform. This platform provides three functions to the users: a legal database, a customized single test advisory module and a harmonization test advisory module.

Although our efforts are concentrated on the creation and distribution of a new dataset, we also propose and test a simple text ranking task, i.e. a ranked list of articles given a specific query. Traditionally, search has been carried out with sparse methods and exact term matching, with ranking functions such as TF-IDF or BM25. These methods are based on statistical features such as term and document frequency or document length, but they are powerless in case of terms mismatch, i.e. the so-called *vocabulary mismatch problem* (Furnas et al., 1987). Apart from enriching query and document representation, models have evolved beyond the limits of exact term matching with techniques such as Latent Semantic Analysis (Deerwester et al., 1990) or Latent Dirichlet Allocation (Wei and Croft, 2006), though they have never achieved large adoption. On the contrary, dense representation models have become widely popular. They are based on word embeddings, like word2vec (Le and Mikolov, 2014), on which the relevance score can be easily calculated through cosine similarity between dense vectors. Upon its arrival, BERT (Devlin et al., 2019) changed the landscape, shortly becoming the state of the art for many NLP tasks. Nogueira and Cho (2019) were the first to use BERT for text ranking. Since then, many other works have followed and defined the state of the art. Among these research efforts we highlight ANCE (Xiong et al., 2021) and STAR/ADORE (Zhan et al., 2021) which have directly inspired our work.

The techniques mentioned in the previous paragraph were applied, in the legal domain, by Kim et al. (2015), Shao et al. (2020). In particular, Kim et al. (2015) proposed two unsupervised models – based on TF-IDF and LDA – for legal information retrieval, with the aim of finding relevant Japanese civil law articles given a specific query. Shao et al. (2020) developed a model for case law retrieval founded on both BM25 and BERT according to a re-ranking scheme, in which the first model (based on BM25) finds all potentially relevant cases, and the second one (based on BERT) refines the search.

Another dataset, conceptually similar to our work, is proposed by Bhattacharya et al. (2019) who collected 50 queries from Section Acts of Indian law and 10,685 documents from the Indian Supreme Court. However, in the literature and in the specific case of European legislation, there are no attempts at automatic text ranking, nor datasets available. Consequently – to the best of our knowledge – our work represents a unique and innovative tool for European law.

4. Dataset

The main objective of our dataset is to collect – in a machine-readable format – all the measures adopted by Member States to implement EU directives and align them with the directives themselves. By cross-referencing the transpositions reported on the EUR-Lex website and national official journals, we thus retrieved national laws associated with EU acts.

However, due to the different web platforms, the non-harmonized formats, as well as missing or incomplete references, the documents listed on EUR-Lex could only be partially retrieved. More details can be found in the following paragraphs.

4.1. *Sources*

The texts of European Directives (EUDs) are freely available[4] on the EUR-Lex website, either in HTML or PDF format. We retrieved all the directives implemented in at least one of the countries of our interest, excluding "corrigenda", consolidated versions and legal sources no longer in force at the time being.

The retrieval of national laws was instead more challenging, since Member States collect laws in varying formats and store them on highly different web platforms. To standardize the process, we considered transpositions reported in national official journals and excluded other legal sources, e.g. judgments and administrative acts. This choice is coherent with best practices adopted in the area of computational law and legal informatics, in which the accuracy of traditional legal methodology often needs to be balanced with the management of huge amounts of data and the specific requirements of computational efforts. We selected legal sources following a careful comparative approach on main domestic legislative instruments that are commonly adopted to follow up on obligations enshrined in EUDs, taking into consideration the occurrences published on the EUR-Lex website in the section "national implementations" and the results of the CrossJustice project, in which legal experts produced reports on implementing measures of six EUDs following a State-by-State method.

In conclusion, we approached data selection pursuant to our future research goal, which is setting-up a search engine aimed at reducing the burden of legal data extraction and

[2] crossjustice.eu
[3] site.unibo.it/facilex

[4] Researchers are allowed to elaborate on, and reuse data published on EUR-Lex according to the copyright notice published thereon, which clarifies that "unless otherwise specified, you can re-use the legal documents published in EUR-Lex for commercial or non-commercial purposes".

Table 1 – Selected legislative sources for irrelevant articles sampling; the historical period runs from the first legislature of the last constitutional order (reported in table) until 25/11/2022.

State	Legislative Sources	Starting Period	Web Sources
Italy	Decreto del presidente, Decreto legge	08/05/1948	http://www.gazzettaufficiale.it http://www.normattiva.it
France	Loi, Décret, Ordonnance	09/12/1958	http://www.legifrance.gouv.fr
Spain	Ley, Real decreto	23/03/1979	http://www.boe.es
Ireland	Act of the Oireachtas, Statutory Instrument	01/01/1938*	http://www.irishstatutebook.ie
Austria	*all sources in:* Staats- Bundesgesetzblatt Landesgesetzblatt	19/12/1945	http://www.ris.bka.gv.at

*The first Government of Ireland dates back to 29 December 1937, but the website does not offer the possibility of such a precise selection.

pairing for legal experts and public administration when reporting on directives implementing measures. Unfortunately, the legislative sources cited on EUR-Lex are not always easily machine-readable. In fact, the lack of complete references, inaccessible web pages, as well as poorly formatted texts, reduced textual sources from a nominal amount of 11,726 transpositions to 9,313. Similarly, the number of directives decreased from 1117 to 906.

To simulate a real-world scenario of text retrieval, we sampled – from the same national sources – further acts that are legally independent of the directives. These acts therefore represent the collection of irrelevant documents that a good model must be able to identify and discriminate. For a mere computational reason, and given the extent of the archives, we restricted the search to a few law sources, namely ordinary laws, decrees and their equivalents in other jurisdictions, all within a limited historical period (Fig. 3b). The selected timeframe begins with the first legislature of the latest constitutional order in a given country (e.g. French Fifth Republic, Italian Republic or the 1937 Constitution of Ireland). More details about the selected legislative sources are reported in Table 1.

4.2. *Text transformation*

Legal texts are structured around chapters and sections depending on their length, which include, in turn, articles, paragraphs and subparagraphs. Laws may also contain attached sources like annexes and appendices. Therefore, we tried to conform to this structure by subdividing the text into articles and, possibly, into paragraphs. We excluded preambles, annexes and appendices, as well as possible notes placed after signatures. Unfortunately, although homogeneous within a country, article and paragraph delimiters vary between Member States. Case in point is the distinction between the Italian "articolo" and "comma", and the German "Artikel" and "Absatz". Furthermore, the format and automatic parsing of the text do not always allow for an unambiguous subdivision. For this reason, we relied on the XML/HTML structure whenever possible, even though it is usually available for the most recent documents only. In the occurrence of particularly ambiguous documents, we suspended the fine-grained subdivision of the text.

As previously outlined, not all articles contain relevant information for our research, though. Transposition deadlines, financial clauses, the dates of entry into force and other details do not add useful information to our semantic search task. We consequently decided to remove such cases. To achieve this, we developed a filter inspired by Inverse Document Frequency (IDF), which assumes that common articles appearing in multiple documents are less informative than rare articles. The scoring function for an article a, is the following:

$$\text{article\_idf}(a, C) = \frac{|C|}{|\{d \in C : \text{sim}(d, a) > \lambda\}|} \tag{1}$$

where C is the article corpus, $\text{sim}(\cdot)$ is a similarity function, and λ is the similarity threshold above which two articles are considered semantically identical.

Irrelevant articles are often short and entirely contained inside longer and more elaborate articles. This is the case with articles stating fines for law violations. Therefore, we consider asymmetric similarities to be the best choice for evaluating article similarities. For the similarity function of the Eq. (1), we use Asymmetric Jaccard Similarity (AJS) defined as:

$$AJS(A, B) = \frac{|A \cap B|}{|A|} \tag{2}$$

AJS allows us to penalize short, recurring articles without impacting rare, longer ones. The λ parameter was set to 0.80 based on empirical results.

Once the scores are computed, we remove provisions with the lowest scores. Since the scores distribution looks almost like a binomial distribution, we decided to apply Otsu's thresholding method (Otsu, 1979) to automatically calculate the cutoff of the IDF filter. Otsu's threshold minimizes the intra-class variance while maximizing the inter-class variance of two classes, which in our case represent relevant and irrelevant articles respectively. After getting the cutoffs approved by the legal domain expert, we proceeded to remove articles that are below the cutoff. An example of the distribution with the respective cutoff can be seen in Fig. 1.

To prepare the dataset for machine learning applications, we divided the collection of articles into training and testing sets. The corpus of national laws remained unchanged, while the directive articles were split into training and testing sets in a 90–10 ratio. Throughout this process, we have also ensured that no testing directive, even in a different language, appears in the training set.

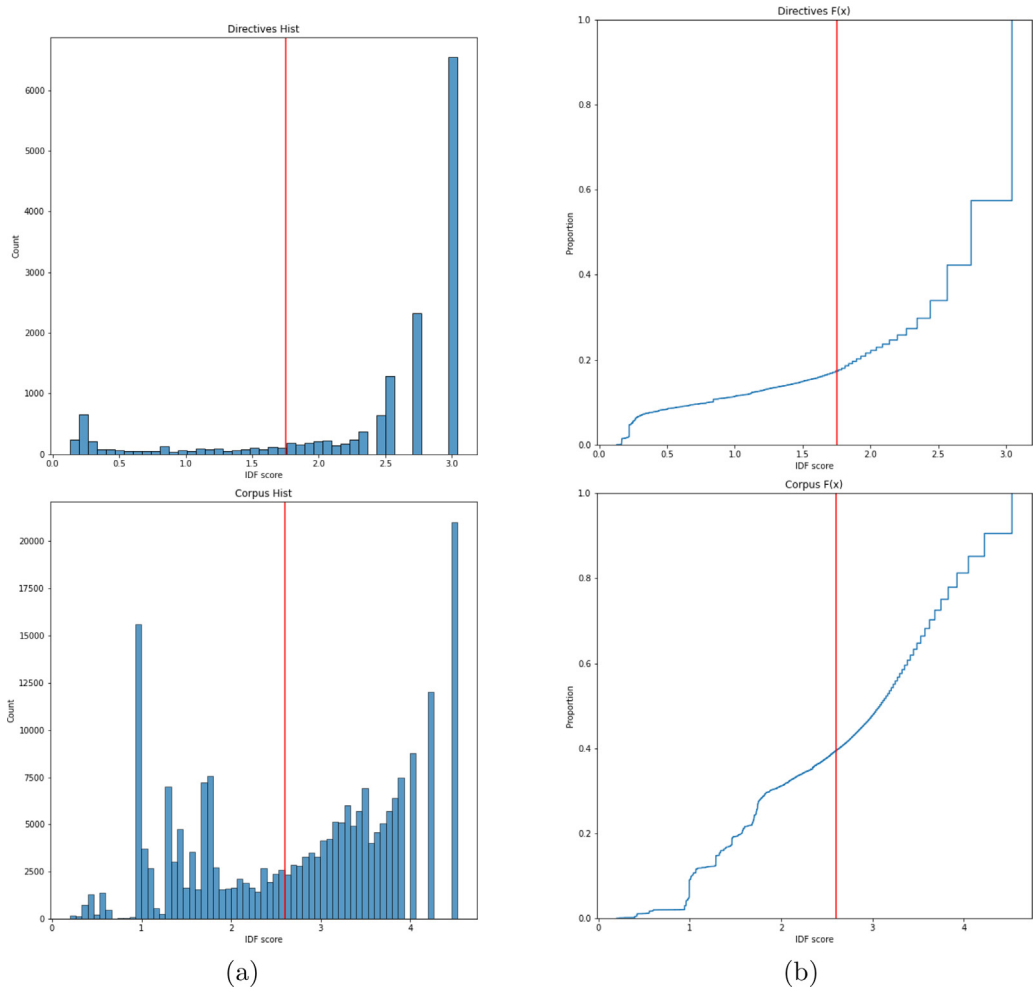

Fig. 1 – Example of histogram (a) and cumulative distribution (b) of article-IDF scores in the Italian subset, with directive articles at the top and national articles at the bottom; the red line indicates the cutoff established by the Otsu's method. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.3. *Data*

The final dataset consists of 57,330 directive articles and 939,160 articles from the national archives, representing 906 unique directives and 9016 documents respectively. Among the provisions from national archives, 827,093 (88%) are unrelated from any directive and are therefore considered irrelevant. The remaining 112,067 articles (12%) implement at least one directive, with 1338 articles associated with multiple directives.

At the basis of our project lies the idea of overcoming the supervised learning approach, which characterized works such as CrossJustice, to prioritize – instead – the creation of a dataset entirely free of expensive manual annotation. We believe that exclusively the Self-Supervised Learning (SSL) approach is able to effectively achieve training scalability. In fact, a careful textual analysis by domain experts can hardly reach a sufficient number of examples to train a modern Deep Learning model. On the contrary, the reports on measures implemented to abide by directives, which are compulsorily submitted by national public authorities to the EU Commission, offer the required training signal, despite being entirely unsupervised data.

We envision that training could be performed on the multilingual dataset just described by exploiting the text augmentation naturally provided by the different languages, while testing could take place on a state-by-state basis.

Each directive is linked to one or multiple national laws and their corresponding articles. On average, each directive has 136 implementing articles (which we will call "positive examples"), with a maximum amount of 2497 articles per directive and a minimum amount of 1. The distribution is shown in Fig. 2. The ranking task is described more in details in Section 5, while the distribution of articles in the different Member States is reported in Table 2 and Fig. 3a. With an average of 200 words, the length of the articles that compose the dataset varies greatly, though. Indeed, except for EUDs which are translations of the same document, there is variety

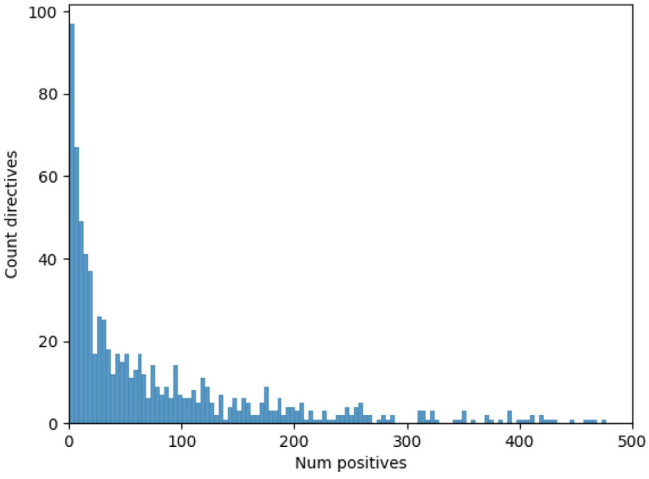

Fig. 2 – Number of positive articles (NIMs) for each EU Directive (EUD) they implement.

Table 2 – Distribution of articles and their origin within the dataset; EUDs provisions are considered queries and divided into train and test sets, while NIMs articles belong to the collection of documents on which to apply the search.

	TOT	ITA	FRA	ESP	IRL	AUT
Queries	57,330	11,514	11,386	11,249	11,344	11,837
train	51,588	10,362	10,248	10,109	10,214	10,665
test	5742	1152	1138	1140	1130	1182
Corpus	939,160	135,221	236,762	209,795	157,601	199,781

between Member States' legislation, because of the different document structures and legal traditions. More details can be found in Table 3.

We publish the dataset in three versions: 1) the parsed collection of directives and national laws, 2) the parsed and filtered collection of articles and 3) the ML-ready dataset split into train and test sets. The transposition tables for the five Member States are included in the dataset, which is available

Table 3 – Mean and Standard Deviation of word count for EUDs articles (queries) and NIMs (corpus).

Country	Words	
	Queries	Corpus
ITA	243±436.36	213±523.51
FRA	258±462.68	200±864.60
ESP	279±502.35	226±439.60
IRL	247±436.26	159±277.32
AUT	212±380.65	342±1823.54
TOT	248±445.35	231±997.57

at https://doi.org/10.17632/mkx5sb3mnw[5], linking the directives (denoted by the CELEX number) to their national implementations (marked by a unique hash string). The ML-ready dataset is a shuffled collection of articles from different countries, with the CELEX number as a label, the name of the country implementing the measure and the identification hash of the national transposition. Compared to the first two versions, this dataset was further filtered to remove directives that – after text transformation – are left without corresponding national counterparts.

5. Ranking task

To overcome the issues related to the usual length that characterizes legal texts, as well as the heterogeneity in contents covered by the same text, we have structured the ranking task on an article-by-article basis. In this setup, an article from an EU directive acts as the query to which one or more articles from the domestic law corpus must correspond. The aim is to retrieve all articles related to a given directive within the whole legislation.

In this paper, we do not deal with the development of such a search engine. On the contrary, we focus on presenting the

[5] For the purpose of the double-blind review, we removed the link to the data and source code. We will make them available in case of acceptance.

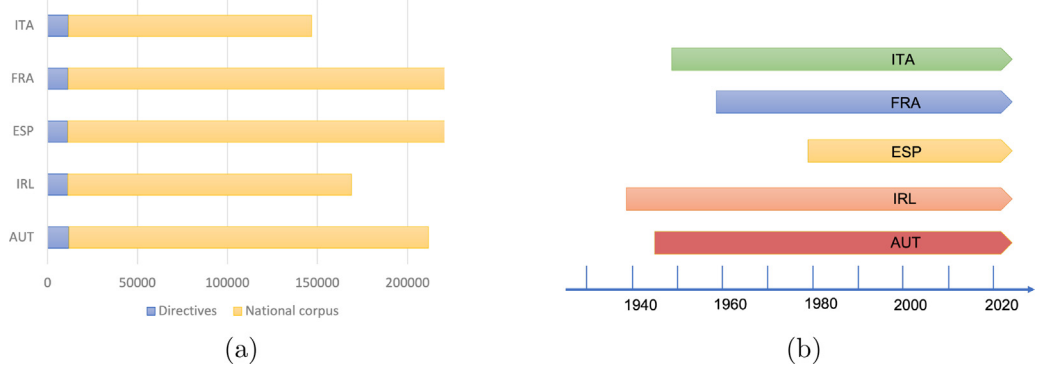

Fig. 3 – (a) Dataset size (number of articles) and its composition, with EUD articles shown in blue and NIM articles in yellow, b) Historical period used for irrelevant articles sampling. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

dataset and evaluating the feasibility of the ranking task using two baseline methods.[6] The first one is based on the Okapi BM25 ranking function, in which retrieval is formulated in terms of inner products on sparse bag-of-words vectors with exact term matching. The second method, instead, adopts an Approximate Nearest Neighbor (ANN) search over dense vectors provided by the multi-language version of Sentence-BERT (Reimers and Gurevych, 2019).

We relied on the Rank-BM25 library (Brown, 2020) and the tokenization provided by spaCy (Honnibal et al., 2020) (it_core_news_lg, fr_core_news_lg, en_core_web_lg, es_core_news_lg, de_core_news_lg models). We also removed stopwords and applied the Snowball stemmer.

For dense retrieval, instead, we used the DistilBERT-based Sentence Transformers model provided by *distiluse-base-multilingual-cased-v1*. Since the model is trained with sequences of 128 tokens, we only used the first 128 tokens of an article, keeping uppercase and lowercase. Through a mean pooling function, we then obtained a single embedding vector for each article and, after L2 normalization, the Flat Index (IndexFlatIP) of FAISS (Johnson et al., 2019) was used for similarity search over the embedding space.

5.1. Results

We evaluated the results with standard ranking metrics such as Hits, Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR). More formally, given a query q from the test set Q, the corpus of documents C and a ranked list $R(q) = [(d_i, s_i)]^k$ truncated at length k, where d_i represents the document at rank i and s_i its ranking score, the metrics are defined as follows:

$$Hits@k(Q) = \frac{1}{|Q|} \sum_{q \in Q} \sum_{d \in R(q)} rel(q, d)$$

$$P@k(R, q) = \frac{\sum_{d \in R} rel(q, d)}{|R|}$$

$$AP@k(R, q) = \frac{\sum_{(i,d) \in R} P@i \cdot rel(q, d)}{\sum_{d \in C} rel(q, d)}$$

$$MAP@k(Q) = \frac{1}{|Q|} \sum_{q \in Q} AP@k(R, q)$$

$$RR@k(R, q) = \frac{1}{lower\_rank_i}$$

$$MRR@k(Q) = \frac{1}{|Q|} \sum_{q \in Q} RR@k(R, q)$$

where $rel(q, d)$ is an indicator function that designates whether the document d is relevant to the query q, and *lower_rank* is the smallest rank number of a relevant document.

The results of BM25 and Sentence-Transformers are reported, respectively, in Tables 4 and 5. Unsurprisingly, word embeddings alone underperform BM25, as already shown in the literature (Gao et al., 2021; Lee et al., 2019; Luan et al., 2021). Dense retrieval models therefore require appropriate training, according to the state-of-the-art techniques currently available: the collected dataset has been specifically designed for this purpose.

Table 4 – Ranking results obtained with the BM25 baseline.

Country	Hits				MAP			MRR		
	@1	@3	@5	@10	@3	@5	@10	@3	@5	@10
ITA	0.35	0.76	1.08	1.73	0.39	0.38	0.36	0.39	0.40	0.40
FRA	0.24	0.58	0.83	1.30	0.30	0.30	0.29	0.30	0.31	0.32
ESP	0.23	0.53	0.77	1.26	0.26	0.25	0.24	0.26	0.27	0.27
IRL	0.16	0.38	0.55	0.86	0.20	0.20	0.20	0.20	0.21	0.21
AUT	0.16	0.40	0.57	0.90	0.20	0.21	0.20	0.20	0.21	0.22
avg	0.23	0.53	0.76	1.21	0.27	0.27	0.26	0.27	0.28	0.28

Table 5 – Ranking results obtained with the Sentence-Transformers baseline.

Country	Hits				MAP			MRR		
	@1	@3	@5	@10	@3	@5	@10	@3	@5	@10
ITA	0.19	0.39	0.54	0.79	0.21	0.21	0.19	0.21	0.22	0.22
FRA	0.07	0.15	0.21	0.32	0.08	0.09	0.08	0.08	0.09	0.09
ESP	0.13	0.29	0.38	0.56	0.15	0.15	0.15	0.15	0.16	0.16
IRL	0.11	0.22	0.31	0.43	0.13	0.13	0.13	0.13	0.13	0.14
AUT	0.07	0.14	0.20	0.31	0.08	0.09	0.08	0.09	0.09	0.09
avg	0.11	0.24	0.33	0.48	0.13	0.13	0.13	0.13	0.14	0.14

6. Conclusions & future work

In this work, we propose an innovative dataset that collects European Directives (EUDs) and their National Implementing Measures (NIMs) for text retrieval applications. We also proposed a ranking task and evaluated its feasibility with two baselines: BM25 and pre-trained Sentence Transformers, which can be used as benchmarks for future developments.

Since BM25 is not trainable, and the possibility to improve its performance is limited, we believe that promising results could be expected from trained dense retrieval methods, following state-of-the-art models such as ANCE (Xiong et al., 2021) and STAR/ADORE (Zhan et al., 2021). With respect to these latest works, we believe that our dataset can be used in a Contrastive Learning scenario, exploiting EUDs and NIMs as "positive examples" and irrelevant laws as "negative examples". We are certainly interested in exploring this possibility, which we will leave as future work.

This research effort, which lays the foundation to automatize the task of pairing EUDs and NIMs, is of high interest for both EU and national public administration, reporting authorities and for legal research more in general. In the future, the dataset might be enriched by including legislation from other jurisdictions, broadening the multilingual spectrum and increasing legal sources, especially EU Regulations and judicial

[6] The source code of both baselines is published, together with the dataset, at https://doi.org/10.17632/mkx5sb3mnw.

decisions. This upgrade might be useful to assess the contribution of national courts in the harmonization of domestic legal orders within the EU, taking into account the duty of consistent interpretation, which was first codified by the CJEU in *Von Colson* and *Kamann*.

In our view, this research work could ultimately pave the way for novel research directions, in order to pursue citizens' access to rights, to support public administrations and – more in general – to promote democracy and the rule of law in the European Union.

Declaration of Competing Interest

Potential conflict of interest exists: We wish to draw the attention of the Editor to the following facts, which may be considered as potential conflicts of interest, and to significant financial contributions to this work: The nature of potential conflict of interest is described below: No conflict of interest exists. We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Data availability

Data will be made available on request.

REFERENCES

Aletras N, Tsarapatsanis D, Preoţiuc-Pietro D, Lampos V. Predicting judicial decisions of the european court of human rights: a natural language processing perspective. PeerJ Comput Sci 2016;2:e93.

Audrito D, Sulis E, Humphreys L, Di Caro L. Analogical lightweight ontology of eu criminal procedural rights in judicial cooperation. Artif. Intell. Law 2022:1–24.

Baaij CJ. The role of legal translation in legal harmonization. Kluwer Law International BV; 2012.

Bhattacharya P, Ghosh K, Ghosh S, Pal A, Mehta P, Bhattacharya A, et al. Fire 2019 AILA track: Artificial intelligence for legal assistance. Proceedings of the 11th annual meeting of the forum for information retrieval evaluation; 2019. p. 4–6.

Boella G, Di Caro L, Humphreys L, Robaldo L, van der Torre L. NLP challenges for eunomos, a tool to build and manage legal knowledge. Language resources and evaluation (LREC); 2012. p. 3672–8.

Brown D.. Rank-BM25: A Collection of BM25 Algorithms in Python. 2020.

Deerwester SC, Dumais ST, Landauer TK, Furnas GW, Harshman RA. Indexing by latent semantic analysis. J Am Soc Inf Sci 1990;41:391–407.

Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, editors. Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2–7, 2019, Volume 1 (Long and Short Papers). Association for Computational Linguistics; 2019. p. 4171–86.

Formigoni G. Storia della politica internazionale nell'età contemporanea (1815–1992). Il mulino; 2018.

Furnas GW, Landauer TK, Gomez LM, Dumais ST. The vocabulary problem in human-system communication. Commun ACM 1987;30(11):964–71.

Gao L, Dai Z, Chen T, Fan Z, Durme BV, Callan J. Complement lexical retrieval model with semantic residual embeddings. In: Hiemstra D, Moens M, Mothe J, Perego R, Potthast M, Sebastiani F, editors. Advances in information retrieval - 43rd European conference on IR research, ECIR 2021, Virtual Event, March 28, - April 1, 2021, Proceedings, Part I. Springer; 2021. p. 146–60.

Honnibal M., Montani I., Van Landeghem S., Boyd A.. spaCy: Industrial-strength Natural Language Processing in Python2020;.

Humphreys L, Santos C, Di Caro L, Boella G, Van Der Torre L, Robaldo L, et al. Mapping recitals to normative provisions in eu legislation to assist legal interpretation. JURIX; 2015. p. 41–9.

Johnson J, Douze M, Jégou H. Billion-scale similarity search with GPUs. IEEE Trans. Big Data 2019;7(3):535–47.

Kim MY, Xu Y, Goebel R. Legal question answering using ranking SVM and syntactic/semantic similarity. New frontiers in artificial intelligence: JSAI-isAI 2014 workshops, LENLS, JURISIN, and GABA, Kanagawa, Japan, October 27–28, 2014, Revised Selected Papers. Springer; 2015. p. 244–58.

Łachacz O, Mańko R. Multilingualism at the court of justice of the european union: theoretical and practical aspects. Stud Logic Grammar Rhetoric 2013;34(1):75–92.

Le Q, Mikolov T. Distributed representations of sentences and documents. ICML'14: Proceedings of the 31st international conference on international conference on machine learning. JMLR.org, 2014.

Lee K, Chang MW, Toutanova K. Latent retrieval for weakly supervised open domain question answering; 2019. p. 6086–96.

Lenaerts K., Gutman K.. "Federal common law" in the European union: A comparative perspective from the United States. 2006.

Luan Y, Eisenstein J, Toutanova K, Collins M. Sparse, dense, and attentional representations for text retrieval. Trans. Assoc. Comput. Linguist. 2021;9:329–45.

Mandal A, Chaki R, Saha S, Ghosh K, Pal A, Ghosh S. Measuring similarity among legal court case documents. Proceedings of the 10th annual ACM India compute conference; 2017. p. 1–9.

Medvedeva M, Üstün A, Xu X, Vols M, Wieling M. Automatic judgement forecasting for pending applications of the european court of human rights. ASAIL/LegalAIIA@ ICAIL; 2021. p. 12–23.

Nanda R, Di Caro L, Boella G, Konstantinov H, Tyankov T, Traykov D, et al. A unifying similarity measure for automated identification of national implementations of european union directives. Proceedings of the 16th edition of the international conference on articial intelligence and law; 2017. p. 149–58.

Nanda R, Humphreys L, Grossio L, John AK. Multilingual legal information retrieval system for mapping recitals and normative provisions. Legal knowledge and information systems: JURIX 2020: The thirty-third annual conference, Brno, Czech Republic, December 9–11, 2020. IOS Press; 2020. p. 123.

Nanda R, Siragusa G, Di Caro L, Boella G, Grossio L, Gerbaudo M. Unsupervised and supervised text similarity systems for automated identification of national implementing measures of european directives. Artif. Intell. Law 2019;27:199–225.

Nanda R, Siragusa G, Di Caro L, Theobald M, Boella G, Robaldo L, et al. Concept recognition in european and national law. JURIX; 2017b. p. 193–8.

Nogueira R., Cho K.. Passage re-ranking with bert. 2019; arXiv:1901.04085.

Otsu N. A threshold selection method from gray-level histograms. IEEE Trans. Syst. Man Cybern. 1979;9(1):62–6.

Prechal S. Direct effect, indirect effect, supremacy and the evolving constitution of the european union. The fundamentals of EU law revisited: Assessing the impact of the constitutional debate; 2007. p. 35–71.
Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP); 2019. p. 3982–92.
Šarčević S. The myth of EU terminology harmonization on national and EU level. Language and culture in EU law. Routledge; 2016. p. 225–36.
Shao Y, Mao J, Liu Y, Ma W, Satoh K, Zhang M, et al. BERT-PLI: Modeling paragraph-level interactions for legal case retrieval. IJCAI; 2020. p. 3501–7.
Siragusa G, Robaldo L, Di Caro L, Violato A. Textual entailment for cybersecurity: An applicative case. J. Appl. Logics 2021;8(4):975.
Sosoni V, Biel L. Eu legal culture and translation. JLL 2018;7:1–7.
Sulis E, Humphreys LB, Audrito D, Di Caro L. Exploiting textual similarity techniques in harmonization of laws. In: Bandini S, Gasparini F, Mascardi V, Palmonari M, Vizzari G, editors. AIxIA 2021 – advances in artificial intelligence. Cham: Springer International Publishing; 2022. p. 185–97.
Wei X, Croft WB. LDA-based document models for ad-hoc retrieval. SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: Association for Computing Machinery; 2006. p. 178–85.
Xiong L, Xiong C, Li Y, Tang K, Liu J, Bennett PN, et al. Approximate nearest neighbor negative contrastive learning for dense text retrieval. 9th International conference on learning representations, ICLR 2021, Virtual Event, Austria, May 3–7, 2021. OpenReview.net, 2021.
Zhan J, Mao J, Liu Y, Guo J, Zhang M, Ma S. Optimizing dense retrieval model training with hard negatives. SIGIR '21: Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval. New York, NY, USA: Association for Computing Machinery; 2021. p. 1503–12.