

# A practical application of Artificial Intelligence techniques for legal context analysis

Ilaria Angela Amantea<sup>1</sup>[0000-0003-1329-1858], Guido Boella<sup>1</sup>[0000-0001-8804-3379], Chiara Bonfanti<sup>1</sup> \* [0009-0007-8015-7786], Michele Colombino<sup>1</sup> \*\* [0009-0007-3248-1661], Luigi Di Caro<sup>1</sup> [0000-0002-7570-637X],  
Giorgia Iacobellis<sup>1</sup> \*\* [0009-0003-1730-7711], Susanna Marta<sup>2</sup> [0009-0003-4014-261X], Rachele Mignone<sup>1</sup> \* \* \* [0009-0009-2699-8730],  
Marianna Molinari<sup>2</sup> [0009-0003-1832-8135], Ivan Spada<sup>1</sup> \* \* \* [0009-0002-0459-1189],  
Emilio Sulis<sup>1</sup> [0000-0003-1746-3733], and Laurentiu Jr Marius Zaharia<sup>1</sup>  
\* \* [0009-0002-3559-8367]

<sup>1</sup> Computer Science Department - University of Turin, Italy  
{rachele.mignone, ivan.spada, michele.colombino, ilariaangela.amantea,  
emilio.sulis, luigi.dicaro, guido.boella}@unito.it  
{chiara.bonfanti, giorgia.iacobellis, laurentiu.zaharia}@edu.unito.it  
<sup>2</sup> Law Department - University of Turin, Italy  
{marianna.molinari, susanna.marta}@unito.it

**Abstract.** Legal management systems typically focus on specific tasks, pursuing organizational improvement, documentation management, and decision-making. This contribution explores automatic classification, taxonomic alignment, information extraction, and legal context analysis in a real case study. We propose a practical application that does not consider the different tasks separately but integrates them into an online platform with the objective of cataloguing, indexing and enabling semantic search by legal context. The first results demonstrate the ability to perform several tasks on the same legal domain, by addressing domain experts through a unified legal management system.

**Keywords:** Legal document classification · Taxonomy alignment · Augmented reading · Automated knowledge extraction · Legal case study · Principles of Law

## 1 Introduction

Legal informatics is increasingly applied in a large variety of legal fields, e.g. organizational management, documentation management, and decision-making. A relevant task concerns legal context analysis by automated reading and semantic comprehension. The definition of a context, interpretable by experts in the legal domain, grants the possibility of making systems interpretable by design. Several

---

\* This author contributed mainly to the analysis of Principles of Law.

\*\* This author contributed mainly to classification and alignment tasks.

\*\*\* This author contributed mainly to the analysis of the legal context.

Natural Processing (NLP) techniques allow to intercept semantically relevant information in a legal document so as to trace the semantic context within which it falls. This can be the basis for deeper analysis and comparisons between legal documents.

This paper describes a practical application of Artificial Intelligence (AI) techniques for legal context analysis in an Italian case study. We demonstrate the ability to apply different approaches by involving legal experts in a real case.

A first challenge of a typical operative framework concerns the conversion of legal judgments into a machine-readable format. Documents are semi-structured and stylistic variations bring complications in intercepting speech changes and identifying the positions where relevant information appears in the legal judgment. In a second step, we investigate the classification and analysis of legal documents, by defining a single shared hierarchy of subject classification labels. To achieve our goals, we introduce a taxonomy alignment pipeline focusing on judgments and classification headings from two different legal sources. Finally, this alignment will then be used to classify the available judgments with machine learning models.

In the following, Section 2 introduces the legal domain, Section 3 defines the related works, while Section 4 shows some definitions and the data used in our experiment. Details of the methods used to achieve our task are discussed in Section 5. Finally, Section 6 and Section 7 concludes the paper.

## 2 Legal domain

In recent years, computer technology and machine learning (ML) have reshaped knowledge in many areas, including law. The legal professions and the world of law are also approaching these issues, pushed by the recent paradigms dictated by digital innovation and automation. The digitisation of justice has a twofold impact: firstly, it affects the direct activities of judges and lawyers and the tools they use in carrying out their respective professions; secondly, it affects the sources from which legal practitioners draw information on precedents and laws.

As a matter of fact, the efficient use of existing sources and materials not only allows for time and cost efficient justice but also allows one to benefit from the experience of insights and resolutions already made and identified in previous legal applications. In this respect, the text of a court decision can be a rich source of information and data, but it will not be useful to a subsequent legal operator unless the data are organized, classified and analyzed.

The present work aims at identifying a methodology for manipulating legal texts to identify similarities in both the lexical and semantic aspects. Once identified, they can be automatically classified to make archiving and searching efficient. These developments impact and improve the entire judicial apparatus.

The research is based on a real case study within a large national project in Italy, i.e. the Next Generation UPP<sup>3</sup> (NGUPP project), which aims at improving the efficiency of the justice system in northwestern Italy.

---

<sup>3</sup> Next Generation UPP - <https://www.nextgenerationupp.unito.it/home>

The project involves collaborations between universities and judicial offices in order to provide the staff of the Trial Office with effective skills to ensure the understanding and functioning of a justice system for providing support to the process of digitization and technological innovation.

In an effort to identify feasible solutions for an effective legal management system, our contribution involves the creation of a set of analyses and tools able to be used by legal practitioners.

The present paper reports the results of processing legal documents in their textual content, through the use of NLP and, more generally, AI techniques applied to a database system, e.g. advanced semantic navigation, classification, and taxonomic alignment of judgments.

### 3 Related work

In the legal informatic scenario, many contributions have explored topics concerning the development of computational methodologies and techniques [10, 24, 30], as well as the applications of AI techniques for manipulating legal texts [23]. Recent efforts in legal informatics have focused on automatic classification [5], information extraction [26], construction and alignment of legal ontologies [4], automatic legal text synthesis [25], or analysis of judicial decisions.

Our contribution combines different methodologies for the manipulation of legal texts using lightweight NLP techniques. In this context, a relevant task concerns the alignment of taxonomies that share similarities both lexically and semantically. The concepts covered by the alignment are set within hyperonymic relationships which complicates the research for similar features between different structures. In the latest years, the alignment task is typically approached by employing ontology matching techniques, [32] hierarchical clustering [6] or rule-based methods [8]. Ontology matching is definitely the most popular approach. Nevertheless, recent studies have considered the use of ML and deep learning techniques to improve the accuracy of alignment [9, 12]. In the case study here proposed, we based on a previous work [7] in which the alignment task was conducted between a taxonomy and a list of unstructured labels.

Since legal documents are considered to be semi-structured, when performing information extraction in the legal domain, the first step is the standardization of the text and its paragraph structure, which allows the documents to be machine-readable and homogenizes documents extracted from different sources. This step can be performed either by using the document’s formatting [1] or by performing topic modelling [29].

These kinds of representations allow for a deeper analysis of the documents which can highlight otherwise “hidden” information such as the document’s contextualization within the jurisprudential setting. Therefore, it is relevant to investigate key information and the citations within the document. In literature, citations are often treated following a graph-based approach which, by building a network of documents, explicit their correlation [31, 27].

In the proposed approach, we extract citations and other relevant information leaving the final user (i.e. legal actor) the choice of weighting each of them according to some personalized criteria. The legal context in which the documents are located becomes somewhat customizable and can vary based on the actor and/or the task. Each piece of information, as well as each kind of citation (i.e. case law or jurisprudential), can have a different relevance, which can also depend on the paragraph where they appear.

This is possible because each paragraph of a legal document, if well structured, contains information relevant to a specific context and hence translates part of its meaning to its content. Furthermore, in literature, the paragraph is often treated as an input document, in order to reduce the task complexity [21].

## 4 Data and sources

In this section, we outline the data and the composition of the datasets used for the different tasks, discussed in detail in Section 5. In particular, the data used came from two different sources:

- **Court of Turin:** a set of 27,477 first- and second-instance judgments granted by the Court of Turin, Labor Section, which we will refer to throughout this paper as *turin-set*.
- **Leggi d'Italia archive:** another set of judgments, varying in quantity depending on the reference task, was extracted from *Leggi d'Italia*<sup>4</sup> (a web-based archive of judgments made public, distributed throughout Italy). The retrieval of these legal judgments was done by applying scraping techniques through the python library *scrapse*<sup>5</sup>.

All legal judgments were obtained in various textual formats: real-PDF, DOCX, DOC, DOCM, and HTML. To achieve the various methods of analysis and processing of legal texts, discussed in this paper, the judgments were rendered in JSON format. Extraction into such a machine-readable format was achieved with the support of domain experts, so as to select the most relevant textual content elements, discernible in the judgments in our possession.

Specifically, the obtained JSON file contains the following attributes: filename, Court of jurisdiction, Section (i.e. *Sezione*), matters, judgment identifier (code-year), NRG (code-year, with which each case is associated), judgment type (first- or second-instance) and segmented paragraphs.

### 4.1 Classification and taxonomy alignment

In order to achieve tasks concerning the classification and taxonomic alignment of judgments (more details in sections 5.1 and 5.2), we will refer as *LI-set* to the set of 21,562 legal judgments belonging to the Labor Section, automatically

<sup>4</sup> [https://pa.leggiditalia.it/#mode=home,\\_\\_m=site](https://pa.leggiditalia.it/#mode=home,__m=site)

<sup>5</sup> <https://pypi.org/project/scrapse/>

extracted from *Leggi d’Italia*, in addition to the *turin-set* data. Specifically, in the *turin-set*, only a subset of 4,804 was labelled, thus useful for the purposes of a supervised classification task.

As can be seen, analyses performed on a dataset of such size can be biased, especially due to the nature of the data and its sampling. To overcome the problematic lack of data, we introduced an initial step of taxonomic alignment and sentence enrichment, whose results are visible in Section 5.2.

## 4.2 Understanding the juridical context

The context analysis was tested on a subset of 5,059 legal judgments sampled on *LI-set*, which we will call *LI-set-8*. This subset was extracted by taking legal judgments belonging to the eight most populated matters and with separate Facts and Decisions paragraphs so as not to work on unsupervised segmented documents.

## 4.3 Principles of Law extraction

Within the context of our research - conducted in the framework of the Italian legal context - principles of law have been defined as focus points of universalizable individual decisions provided by the Supreme Court of Cassation (i.e. *Suprema Corte di Cassazione*). Thus, since - according to its monophyletic role - the Supreme Court offers generalized interpretations and applications of rules. In this sense, it is important to note that a principle of law does not serve as a source of law, nor does it encode a detailed rule accompanying the interpreted one. Rather an integration of the law by the interpreter, thanks to which the individual decision is brought under a general rule intended to be applied not only to the same cases but also to similar or comparable ones. While it may be tempting to describe the relationship between Supreme Court citations and principles of law as a two-way relationship, such an approximation may prove inaccurate from a juridical standpoint. In fact - as experienced - citations of principles of law may vary, even within the same reporting judgment: when such stated in the jurisprudential panorama, they can be implicit and so normally just mentioned; otherwise, they can be directly correlated to citations of rulings of the Supreme Court, and so explicit. Not by chance, these principles stem from legal interpretation: from single rules, from more or less vast sets of rules, or from the legal system as a whole [11]

*Extraction.* With regard to the extraction of principles of law [17], we explore a dataset created from Turin-set with subsequent annotations made by domain experts, hereafter reported as POL-set. During these annotations, the domain experts have been asked to highlight the explicit citations of Supreme Court rulings, when indicative of a principle of law (i.e. cfr. Cass. no. 32500 of 2018 cit., whereas ”cfr. Cass.” is the keyword, and ”32500” is the id of the judgment issued in 2018).

## 5 Methodology

The wide set of methods exploited for processing legal texts concerns automatic classification techniques, taxonomic alignment, legal context analysis, and principles of law extraction.

### 5.1 Classification

*Datasets.* A recent work [1] focused on a tentative of automatic classification on first - and second - instance judgments, from the labour section of the *turin-set*. The aim was to show how it was possible to achieve good performance on the data at our disposal without using state-of-the-art neural models, which require a lot of computational resources in terms of space and execution time. In this particular case, the *turin-set*, has a strongly unbalanced distribution of labels. In fact, as mentioned in 4, only a subset of the judgments was considered for our experiments, i.e the 15 most populated labels. Finally, we defined 2 corpora of judgments:

- **corpus\_8\_classes:** generated using 800 judgments equally distributed among the 8 most populated labels, which are: "*individual dismissal*", "*contribution*", "*salary*", "*damage compensation*", "*teachers*", "*legal disability*", "*qualification*" and "*healthcare*".
- **corpus\_15\_classes:** generated using 1,872 unequally distributed judgments across the 15 most populated entries, which are: "*individual dismissal*", "*contribution*", "*salary*", "*damage compensation*", "*teachers*", "*legal disability*", "*qualification*", "*healthcare*", "*appeal*", "*severance pay*", "*legal fees*", "*fixed-term contract*", "*jurisdiction*", "*agency*" and "*limitation*".

To define our dataset, we applied vector space modelling techniques. In particular, the following 4 resources were used: TF, TF-IDF, Italian-Legal\_Bert [15] and Doc2Vec [13].

*Models.* To tackle a supervised classification task, we exploited 3 models: multiclass SVM [2], Logistic Regression <sup>6</sup>, Random Forest classifier [3]. Finally, we repeated the tests by performing ensemble learning of type Voting classifier <sup>7</sup>. All tests were performed in 10-fold cross-validation.

*Results.* From the results in [1], the best performing model is Logistic Regression run on the dataset constructed using Doc2Vec embeddings. This combination gave us accuracy values close to 96% for the *corpus\_15\_classes*. On the other hand, on the *corpus\_8\_classes*, the Random Forest classifier turns out to be the model that summarily returns the highest performance of around 98% accuracy. The results are impressive in that with the adoption of simple models, we have achieved considerable performance by overcoming the limits of data lacking.

<sup>6</sup> [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LogisticRegression.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html)

<sup>7</sup> <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html>

## 5.2 Taxonomy alignment

*Pipeline.* Considering the unbalance of the distribution of the labels in the *turin-set* and being aware that any other task on this set might suffer from the same bias, we approached a first attempt at aligning legal taxonomies [7]. Looking at the distribution of the judgments in the 2 sets, we noticed how on the *LI-set* we have a well-defined taxonomy of labels, while on the *turin-set* we have no structure. More in detail, this is a work of taxonomy transfer and not just alignment, carried out once again with the use of streamlined techniques. The aim is twofold: we define an alignment of entries, as well as we realize a data enrichment. To achieve this result, we defined a pipeline:

- **Label comparison:** a lexicographic comparison between the names of the labels. This comparison between the labels of the 2 sets is performed considering the labels of all the levels of the *LI-set* taxonomy. We use the *Reinforced Edit Similarity* [7], a similarity metric that combines the edit distance and the cosine similarity on the counter-vectorized representation of the labels. Finally, we select the pair of labels with the highest similarity score on the various levels of the *LI-set* taxonomy.
- **Semantic similarity:** in this phase, we transform the judgments into Doc2Vec embeddings. Then we recover an equal number of judgments from both sources for each match of labels that we consider at the end of the previous step. To evaluate the semantic similarity, we compute the cosine similarity between Doc2Vec’s centroids of the clusters of the candidate labels. Finally, we select the pair of labels with the highest score.
- **Validation with classification:** At the end of the alignment process, classification work was performed. The results were then compared with those obtained before alignment, as a check on the goodness of the method adopted.

*Datasets.* For the taxonomic alignment task, we considered a different composition of the datasets, compared to the classification task. As a consequence of the fact that at the end of the pipeline, a subset of 11 entries of the *turin-set* was correctly aligned with a subset of 10 entries of the *LI-set*, from this result, we generated 2 corpora:

- **corpus\_11\_labels\_torino:** generated using a set of 318 judgments distributed among 11 entries of the *turin-set* dataset, resulting from the alignment process. The list of labels of this corpus is the following: “agency”, “subordinate work”, “collective dismissal”, “social allowance”, “proof”, “dismissal”, “assistance”, “notification”, “injunction”, “sickness benefit ” and “severance pay”.
- **corpus\_10\_labels\_LI:** created using a total of 7308 judgments distributed among 10 entries of the *LI-set* dataset. This set of labels contains the following ones: “AGENCY (contract)”, “SUBORDINATE WORK (rapport of)”, “SOCIAL SECURITY/Civil invalids”, “PROOF GENERALLY IN CIVIL MATTERS/Burden of proof”, “SUBORDINATE WORK/Dismissal”, “WORK AND SECURITY/Social security and assistance”, “TAXES AND

*TAXES IN GENERAL/Notification of documents”, "INJUNCTION/Injunction order”, "SUBORDINATE WORK/Allowance, in general” and "PENSIONS/Severance or severance pay”.* As it can be seen, the labels *"subordinate work”* and *"collective dismissal”* of the *turin-set* were merged in a single label called: *"subordinate work”*.

*Results.* To obtain objective feedback on the performance of the pipeline, we executed a classification work before and after the alignment. The aim is to show that the performance of the models used for classification is not significantly diminished. This second classification task used the same models illustrated in the 5.1 section. Based on the results obtained on the *turin-set*, we preferred to use only Doc2Vec as a resource for the generation of sentence embeddings. Hence we used the **corpus\_10\_labels\_LI** as the training set and the **corpus\_11\_labels\_torino** as the testing set. In the results section of [7], we illustrate how the performance of the models decreases significantly as the number of labels increases. For the first 6 labels, only the SVM model obtains an accuracy score of 95%, a value that drops to 80% when considering the total 11 labels.

### 5.3 Understanding the juridical context

This part of the contribution focuses on setting up a flexible system for customized context analysis. Legal domain experts argue that the reading, analysis, and comparison of legal judgments follow a partially standard procedure, which is supplemented by the goals and know-how of the specific legal actor. End-users can give as much relevance as they see fit to each paragraph and each type of citation (legislative or case law, first and second instance, or Supreme Court) depending on their goals and what they consider most important.

A major step in understanding a legal document is its contextualization through the extraction and processing of semantically relevant information. The extraction of such context allows for more in-depth analysis and explainable, human-readable results.

**Document segmentation** Legal judgments and scientific papers have some similarities with each other in terms of (1) structure (sequence of paragraphs describing the development of the legal case or scientific experiment), (2) use of specific language, and (3) different importance of information in the paper depending on location. In the field of scientific papers, the *scite* index [19] takes the latter consideration into account. In fact, a citation has a different importance if it appears in the introduction or in the discussions section.

First and second instance judgments follow a semi-structured schema. The document is divided into core parts (header, facts, reasoning, and decisions), which sometimes may be merged. Being able to segment the judgments by core parts, often corresponding to the standard paragraphs, allows for a more precise



analysis that can attribute different relevance to the extracted information depending on its position in the text. In this way, the end-user is provided with a customized tool that can accurately meet his personal needs.

Given a dataset of legal judgments with the paragraphs well separated (or properly segmented through topic analysis technologies), it is clear how each section describes a different core point of the legal case. In this perspective, end-users may be interested in, for example, a facts-oriented or decision-oriented analysis. Similarly, the analysis can be carried out on multiple paragraphs weighted by importance, usefulness, and impact on the task.

A limitation of this approach is that some legal judgments are not written by splitting each specific core part into paragraphs, and the use of unsupervised systems would go on to introduce errors during document analysis.

**Information extraction** Since the document follows a semi-structured outline, relevant information can be sought in recurring positions which are usually standard but may vary depending on who redacted the document. The information extracted from the document by scraping the header and segmented paragraphs is located as follows:

- **Court and Section:** found either in the metadata or the header of the document and extracted while building the dataset.
- **Matter:** can be found in multiple places such as metadata, object, or decisions section (where it can be expressed explicitly or implicitly). In most cases, it was extracted when building the dataset, while in others it was necessary to automatically classify the document as described in section 5.1
- **Outcome:** usually stated in the *P.Q.M.* paragraph. While it is often written explicitly, it may be necessary to resort to further analysis in order to extract it.
- **Citations:** present throughout the document though their role can change significantly depending on the type (legislative or case law) and paragraph in which they are found [18].

**Building a weighted context** Augmented reading, performed on dataset *LI-set-8*, defines the context in which the legal case is set. Automatic knowledge extraction is enriched by the weight given to each paragraph; hence, the relevance of the extracted information is defined by its position in the document. The weight assigned to each portion of the document depends on the criterion chosen for context definition and on the subjectivity of the legal actor. Paragraphs can be optionally nullified, with weight 0, or made equal, with weight 1.

Specifically, we can construct the semantic context of a legal judgment by extracting the following information: Court, Section (macro-topic label), matter (micro-topic label), paragraphs (Facts, Decisions, etc.), legislative and case law citations, and outcome. Starting with a coarse-grained description of context (defined by Section and Matter), we go deeper until we reach a fine-grained level of detail that is useful for more specific analyses.

Such context is constructed by assigning to each piece of information retrieved a weight that determines its importance for a subsequent analysis or processing phase. By doing so the information can be extracted in bulk for the document and can be used in different ways and adapted to specific use cases.

Furthermore, as claimed by legal domain experts, identifying the paragraph in which some information occurs can provide further semantic evidence that would not be available otherwise. For example, a citation extracted from the facts paragraph of a second-instance document will likely be referring to a topic covered in the first-instance judgment of the same trial.

#### 5.4 Principles of Law

With the scope of furthering our understanding of the semantic connotations of the case law citations in the text of the ruling, we defined in our previous works a regular expression extraction methodology as a baseline to tackle the task. To improve our results we explored the possibilities of giving the extracted data a more structured form.

*Extraction.* Taking into consideration the challenge of having different formats and patterns of even the most common information (e.g. dates), it has been given much more importance to the pre-processing stage, compared to our previous approaches. In our experiment, we compared our previous results on *POL-set* with our new more structured approach. Our findings are shown in Table 1 which represents the following values of the validation of our regex extraction:

- **Data:** Performing our calculation on *POL-set* (a subset of the Turin set generated from our domain experts’ annotations), we performed two different experiments whose results are represented as the rows in this table. *Unstructured data* is the raw in which can be found results of the regex extraction of citations from the original dataset. Meanwhile *Structured data* is the result of a novel approach comprehending more pre-processing and regular expressions used. The data in this set are organized as a JSON object with the following fields: "text", "context", "number", and "date".
- **Retrieved and Relevant set:** The outcome of the regular expression pattern matching has been stored in a list called "*Retrieved*" that represents a collection of retrieved citations of Supreme Court rulings. The cardinality of this list is considered in our calculations. Similarly, the extraction of the ground truth represented by *POL-set* from *.docx* documents, annotated by legal domain experts, have been extracted and stored in a list called "*Relevant*", which, like the "*Retrieved*" list, is taken into account for result analysis and metric calculations.
- **Intersection and Threshold:** representing important information for the subsequent calculations of metrics, the *intersection* has been obtained differently for *structured* and *unstructured* data. For the former, there has been a more extensive pre-processing that led to one on one similarity intersection between a unique identification key made of "text" and "number" fields. For

the latter, which had been through only mild pre-processing, the intersection has been calculated between *relevant* and *retrieved* citations, and it has been obtained with the use of an Edit Distance.<sup>8</sup> The *threshold* percentage value represents the similarity that has been used to create the *intersection* set.

- **Precision, recall, f1-score:** metrics such precision, recall [28], and f1-score [22] have been calculated to further compare our findings. Better results, correlated as well with the threshold value, are recognizable when the data is structured.

Table 1: This table shows the results obtained during the principles of law extraction process using a set of regular expressions.

Supreme Court citations			<i>Metrics and evaluations</i>				
<b>data</b>	<b>retrieved</b>	<b>relevant</b>	<b>intersection</b>	<b>threshold</b>	<b>precision</b>	<b>recall</b>	<b>f1score</b>
unstructured_data	5	10	4	70%	0.8	0.4	0.533
structured_data	8	13	8	100%	1.0	0.615	0.761

*Evaluation.* Our results have been evaluated using common metrics, whose calculations have been made starting from a similarity set by a simple intersection between the *relevant* and *retrieved* sentences. It has been found that structured data with the appropriate ID have better precision and recall.

## 6 Future work

This contribution aims to delve deeper into the three main tasks and the technologies employed from different perspectives.

First, expanding the dataset to more Sections, beyond the Labor Section already covered, would allow for a greater overview and testing on a broader landscape. Similarly, covering more Courts and Sections implies dealing with documents written by more legal domain experts providing more stylistic differences in the writing of legal judgments.

The datasets used are from Italian Courts, the use of multilingual platforms such as EurLex<sup>9</sup> would allow the technologies to be tested on multiple languages and use those that are available or perform better in one, such as English. Since legal judgments are translated across multiple languages, the analysis performed on one translation can be moved to the others. In an indirect way, it is possible to perform analysis on the English version and then provide the results for the other languages that may not benefit from the same performing tools.

<sup>8</sup> Edit Distance - <https://www.nltk.org/api/nltk.metrics.distance.html>

<sup>9</sup> EurLex - <https://eur-lex.europa.eu/homepage.html> accessed 23.06.2023

Regarding the taxonomy alignment pipeline, our techniques can also be introduced in an international context. Future developments will involve both the refinement of the methodologies used to study similarities between different taxonomies and an analysis of the application of our techniques to multilingual legal cases.

With regard to principles of law, our research focused on the framework of the Italian legal context and, in particular, on principles of law, stated by the Supreme Court of Cassation, when reported in first or second-instance judgments. The next steps might investigate if there could be room for comparable experimentations at an international level, such as within citations by the Court of Justice of the European Union (CJEU) or by the European Court of Human Rights (ECHR). In this perspective, it is crucial to consider the inherent differences in the definition of these principles of law, from a juridical standpoint, as well as the patterns in which they are represented in rulings.

In addition, future approaches on the topic may comprehend the usage of technologies such as legal expert systems. According to [14] the concept of an expert system is attractive. It carries the connotation of assisting individuals to access justice, enhancing the overall legal process [20]. There have been previous approaches to the usage of legal expert systems in a NLP context such as [16]. Once extracted, principles of law provide a good metric and it has been speculated that they can constitute a knowledge base onto which conduct inference potentially aiding in jurisprudential prediction. Especially in cases in which jurisprudence, with its principles of law, is called to make up for, even in the absence of legislation (referred to as "lawless cases"). In the Italian legal context, these cases frequently revolve around personal rights.

In addition, the use of multilingual systems would make it possible to expand the overview by distributing it geographically and generalizing it to the European level.

## 7 Conclusions

In this paper, we presented a practical application of different methods for processing legal judgments by delving into the definition of legal context on several levels. The goal is to propose lightweight tools that can be integrated into an online platform having the objective of cataloguing, indexing and enabling semantic search by context.

We explored an approach of transposition and alignment of a not-hierarchic structure in a well-defined taxonomy, using a pipeline of different approaches. With the legal judgments of two different sources and their labels, we first defined a lexical similarity on the labels with a new similarity metric, resulting from the combination of existing techniques. Hence, going down to the semantic level, we applied *cosine similarity* on the centroids in the groups of judgments we identified as similar in the previous step. Finally, as a check on the validity of our method, we trained some machine learning models, and then evaluated the performance on the data before and after the alignment. With his final check,

we were assured that the alignment did not lead to a loss of information in the newly constructed groups of judgments.

From the perspective of defining the legal context, we have seen how to consider the semi-structured document as a sequence of parts/paragraphs. In this way, it is possible to customize the analysis by choosing only the relevant sections of the document. We found the following main advantages: (1) specialization of the analysis by going to reduce noise, (2) attribution of specific meanings to the extracted information depending on its position in the text, (3) weighting of paragraphs, and (4) reduction of computational load by going to ignore the parts of the document that are not affected.

One of the main challenges is to test the systems on multilingual archives so as to assess the effective generalization of the contribution at the European level.

## 8 Acknowledgments

The research work has been funded in Next Generation UPP<sup>10</sup> project supported by the European Union, National Operational Program Governance and Institutional Capacity 2014-2020, European Social Fund and European Regional Development Fund. The Next Generation UPP project is part of the "Unitary project for the dissemination of the Office for Trial and the implementation of innovative operating models in the judicial offices for the disposal of the backlog", promoted by the Italian Ministry of Justice and implemented in synergy with the interventions envisaged by the National Recovery and Resilience Plan (NRRP) in support to the justice reform.

## References

- [1] C. Bonfanti et al. "A pipeline for data management, knowledge extraction and semantic analysis of unstructured legal judgments". In: *(In press) Proceedings of Conference Ital-IA 2023*. 2023. URL: <https://www.ital-ia2023.it/submission/41/paper>.
- [2] Bernhard E. Boser, Isabelle M Guyon, and Vladimir Naumovich Vapnik. "A training algorithm for optimal margin classifiers". In: *Annual Conference Computational Learning Theory*. 1992.
- [3] L. Breiman. "Random Forests". In: *Machine Learning* 45 (2001), pp. 5–32.
- [4] Cristian Cardellino et al. "Ontology Population and Alignment for the Legal Domain: YAGO, Wikipedia and LKIF". In: *International Workshop on the Semantic Web*. 2017.
- [5] Haihua Chen et al. "A comparative study of automated legal text classification using random forests and deep learning". In: *Inf. Process. Manag.* 59 (2022), p. 102798.

<sup>10</sup> Next Generation UPP - <https://www.nextgenerationupp.unito.it/home>

- [6] Patrick Clerkin, Pádraig Cunningham, and Conor Hayes. *Ontology discovery for the semantic web using hierarchical clustering*. Tech. rep. Trinity College Dublin, Department of Computer Science, 2002.
- [7] Michele Colombino et al. “Organizing the Unorganized: A Novel Approach for Transferring a Taxonomy of Labels into Flat-Labeled Document Collections.” In: *(In press) Proceedings of ASAIL 2023, 6th Workshop on Automated Semantic Analysis of Information in Legal Text*. 2023. URL: [https://drive.google.com/file/d/1vUbmPY073rqgSqCizB9UT80JV9gjfnqa/view?usp=drive\\_link](https://drive.google.com/file/d/1vUbmPY073rqgSqCizB9UT80JV9gjfnqa/view?usp=drive_link).
- [8] Susel Fernández, Juan R Velasco, and Miguel A López-Carmona. “A fuzzy rule-based system for ontology mapping”. In: *Principles of Practice in Multi-Agent Systems: 12th International Conference, PRIMA 2009, Nagoya, Japan, December 14-16, 2009. Proceedings 12*. Springer. 2009, pp. 500–507.
- [9] Anna Giabelli et al. “WETA: Automatic taxonomy alignment via word embeddings”. In: *Comput. Ind.* 138 (2022), p. 103626.
- [10] Guido Governatori et al. “Thirty years of Artificial Intelligence and Law: the first decade”. In: *Artificial Intelligence and Law 30* (2022), pp. 481–519.
- [11] Riccardo Guastini. “Principi di diritto e discrezionalità giudiziale”. In: *Diritto pubblico* 3 (1998), pp. 641–660.
- [12] Yuan He et al. “BERTMap: A BERT-based Ontology Alignment System”. In: *AAAI Conference on Artificial Intelligence*. 2021.
- [13] Quoc V. Le and Tomas Mikolov. “Distributed Representations of Sentences and Documents”. In: *ArXiv* abs/1405.4053 (2014).
- [14] Philip Leith. “The rise and fall of the legal expert system†† Previously published in Leith P., ‘The rise and fall of the legal expert system’, in *European Journal of Law and Technology*, Vol 1, Issue 1, 2010. View all notes”. In: *International Review of Law, Computers & Technology* 30 (Sept. 2016), pp. 94–106. DOI: 10.1080/13600869.2016.1232465.
- [15] Daniele Licari and Giovanni Comandè. “ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law”. en. In: *EKAW*. Ed. by Symeonidou et al. Vol. 3256. CEUR Workshop Proceedings. Bozen-Bolzano, Italy: CEUR, Sept. 2022. URL: <https://ceur-ws.org/Vol-3256/#km4law3> (visited on 11/19/2022).
- [16] Inari Listenmaa et al. “An NLG pipeline for a legal expert system: a work in progress”. In: *ArXiv* abs/2107.02421 (2021).
- [17] I.A. Amantea M. Molinari C. Bonfanti. “Principles of law: approaching a functional extraction”. In: *(In press) Proceedings of AI4LEGS*. 2023.
- [18] R. Mignone et al. “Augmented reading and Similar Case Matching: from legal domain experts’ modus operandi to a computational pipeline”. In: *(In press) Proceedings of KM4LAW 2023, 2nd international workshop on Knowledge Management and Process Mining for Law*. 2023.
- [19] J. M. Nicholson et al. “scite: a smart citation index that displays the context of citations and classifies their intent using deep learning”. In: *bioRxiv* (2021).

- [20] Adaora Obayi et al. “Advancement in E Recruitment Towards Expert Recruitment System (ERS)”. In: (Nov. 2020).
- [21] K. Raghav, Krishna Reddy, and V. Balakista Reddy. “Analyzing the Extraction of Relevant Legal Judgments using Paragraph-level and Citation Information”. In: 2016.
- [22] C. J. Van Rijsbergen. *Information Retrieval*. 2nd. USA: Butterworth - Heinemann, 1979. ISBN: 0408709294.
- [23] Livio Robaldo et al. “Introduction for artificial intelligence and law: special issue “natural language processing for legal texts””. In: *Artificial Intelligence and Law* 27 (2019), pp. 113–115.
- [24] Giovanni Sartor et al. “Thirty years of Artificial Intelligence and Law: the second decade”. In: *Artificial Intelligence and Law* 30 (2022), pp. 521–557.
- [25] Reshma Sheik and S. Jaya Nirmala. “Deep Learning Techniques for Legal Text Summarization”. In: *2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)* (2021), pp. 1–5.
- [26] Francesco Sovrano, Monica Palmirani, and Fabio Vitali. “Legal Knowledge Extraction for Knowledge Graph Based Question-Answering”. In: *International Conference on Legal Knowledge and Information Systems*. 2020.
- [27] Emilio Sulis et al. “Exploiting co-occurrence networks for classification of implicit inter-relationships in legal texts”. In: *Inf. Syst.* 106 (2022), p. 101821. DOI: 10.1016/j.is.2021.101821.
- [28] Kai Ming Ting. “Precision and Recall”. In: *Encyclopedia of Machine Learning*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA: Springer US, 2010, pp. 781–781. ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8\_652. URL: [https://doi.org/10.1007/978-0-387-30164-8\\_652](https://doi.org/10.1007/978-0-387-30164-8_652).
- [29] Raquel de V. Silveira et al. “Topic Modelling of Legal Documents via LEGAL-BERT1”. In: 2021.
- [30] Serena Villata et al. “Thirty years of artificial intelligence and law: the third decade”. In: *Artificial Intelligence and Law* 30 (2022), pp. 561–591.
- [31] Rupali Wagh and Deepa Anand. “Application of citation network analysis for improved similarity index estimation of legal case documents : A study”. In: *2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC)*. 2017, pp. 1–5. DOI: 10.1109/ICCTAC.2017.8249996.
- [32] Yuanzhe Zhang et al. “Ontology Matching with Word Embeddings”. In: *China National Conference on Chinese Computational Linguistics*. 2014.