A pipeline for data management, knowledge extraction and semantic analysis of unstructured legal judgments

Chiara Bonfanti[1,*], Michele Colombino[1], Giorgia Iacobellis[1], Rachele Mignone[1], Ivan Spada[1], Laurentiu Jr Marius Zaharia[1], Marinella Quaranta[1], Marianna Molinari[1], Susanna Marta[1], Ilaria Angela Amantea[1], Davide Audrito[1], Emilio Sulis[1], Luigi Di Caro[1] and Guido Boella[1]

[1] *Computer Science Department - University of Turin, Via Pessinetto 12, 10149, Torino, Italy*

Abstract

This paper describes a pipeline for data management, knowledge extraction and semantic analysis of unstructured legal judgments on a digital database. The research focuses on the storage of judgments, the processing of textual content through the use of Natural Language Processing and AI technologies and the advanced semantic navigation of the database. These results are obtained from the research group of the University of Torino in the NGUPP project.

Keywords

Legal informatics, Legal document classification, Legal document similarity, Principles of Law, Text embeddings

1. Introduction

The digitalization of justice concerns both the direct activity of judges and lawyers and the sources from which they draw information on precedents and laws. A more efficient exploitation of the stock of knowledge embodied in the decisions issued by the Courts implies a corresponding efficiency gain of the justice system as a whole. Legal informatics aims at providing a possible feasible solution to increase the efficiency of the justice system by unlocking its very own potential. This work describes a pipeline for processing judgment with the creation of a unified digital database for national Courts, through the adoption of a Web App, aimed at the storage of judgments, the processing of textual content through the use of Natural Language Processing / AI technologies, and the advanced semantic navigation of the database thus created.

Office for Trial. The Office for Trial (UPP) is an organizational structure made up of court assistants, operating in the judicial offices. The UPP aims of ensuring the reasonable length of the proceedings, through the innovation of organizational models, the increase in human resources and a more efficient use of technologies. Provided for in Article 16-octies of Decree-Law No. 179/2012, which firstly highlighted a link between technological innovation, organization and quality of justice; it has recently been revalued as a stable organizational structure, thanks to the Italian latest justice reform, and so destined to operate even after the achievement of the National Recovery and Resilience Plan (NRRP) objectives.

Research project. The Next Generation UPP project (NGUPP) aims at improving the efficiency of the judicial system in north-western Italy, by testing - throughout the 35 judicial offices involved - new collaborative schemes between universities and judicial offices in order to provide to UPP employees transversal skills to ensure the effective functioning of a contemporary judicial system and to provide support for the process of digitalization and technological innovation. NGUPP steams from the NRRP, by which Italy engaged with the European Commission in order to define actions and interventions to overcome the economic and social impact of the pandemic, acting on the country's structural nodes and successfully facing the environmental, technological and social challenges of our time. In an effort to identify feasible solutions for the fulfilment of the undertakings given to the European Union through a multidisciplinary approach, using legal, business and IT skills, our research led us to the implementation of a tool that would not only be up-to-date but could also be used by legal practitioners in post-project phases. This paper describes the results obtained from

Ital-IA 2023: 3rd National Conference on Artificial Intelligence, organized by CINI, May 29–31, 2023, Pisa, Italy

chiara.bonfanti@edu.unito.it (C. Bonfanti); michele.colombino@edu.unito.it (M. Colombino); giorgia.iacobellis@edu.unito.it (G. Iacobellis); rachele.mignone@unito.it (R. Mignone); ivan.spada@unito.it (I. Spada); laurentiu.zaharia@edu.unito.it (L. J. M. Zaharia); marinella.quaranta@unito.it (M. Quaranta); marianna.molinari@unito.it (M. Molinari); susanna.marta@unito.it (S. Marta); ilariaangela.amantea@unito.it (I. A. Amantea); d.audrito@unito.it (D. Audrito); emilio.sulis@unito.it (E. Sulis); luigi.dicaro@unito.it (L. D. Caro); guido.boella@unito.it (G. Boella)

0009-0007-8015-7786 (C. Bonfanti); 0009-0007-3248-1661 (M. Colombino); 0009-0003-1730-7711 (G. Iacobellis); 0009-0009-2699-8730 (R. Mignone); 0009-0002-0459-1189 (I. Spada); 0009-0002-3559-8367 (L. J. M. Zaharia); 0000-0003-2691-0611 (M. Quaranta); 0000-0003-1329-1858 (I. A. Amantea); 0000-0002-9239-5358 (D. Audrito); 0000-0003-1746-3733 (E. Sulis); 0000-0002-7570-637X (L. D. Caro); 0000-0001-8804-3379 (G. Boella)

CEUR Workshop Proceedings (CEUR-WS.org)

the research unit of the University of Torino. In the following, Section 2 introduces the background with related works, definitions, and dataset. Section 3 describes the methodology, while first results are detailed in Section 4. Section 5 concludes the paper.

2. Background

Related work. The present work follows the research approach of legal informatics [1], where computational methods and AI applications are increasingly relevant [2], especially in the area of e-Justice and analysis of judicial decisions. Judicial citations are approached with network analysis to address, for instance, the decisions of the CJEU [3, 4]. As concerns automatic judicial interpretation and prediction, a variety of supervised [5] and unsupervised [6] methodologies are applied, e.g. to assess public procurement fraud detection [7], paying attention to explainability [8]. Other research lines pursue the objective of extracting and classifying argumentative patterns in judgments [9] and to model the most effective standards [10] and design-ontological techniques [11] to represent legal text sources. Recently, a promising research domain is engaged with analyzing the process of harmonization of EU and domestic legislation [12].

Definitions. The present paragraph aims at defining terms and keywords on which the particular topic of this paper is based. A judgment (i.e. *Sentenza*) is identified by “code and year". The code is a sequential number released by the court when the judgment becomes definitive and is inserted in the court’s official records. Year, instead, determines the year in which the judgment was published into. NGR, which stands for “number of general register” and corresponds to a chronological number assigned to a specific case (and its files, including the judgment), is used to link and store all the acts and documents related to the case in a unique folder. The subject (i.e. *Materia*) pinpoints a Macro Area of the domain of the judgment, nonetheless the section of the court that created it. The label (i.e. *Voce*) discerns a specific subset of the Macro Area: Salary (i.e. *Retribuzione*), Contribution (i.e. *Contribuzione*), Individual dismissal (i.e. *Licenziamento individuale*) are different labels of the subject Work (i.e. *Lavoro*).

Dataset. The dataset used for the present work encompasses data extracted from Turin Court (i.e.*Tribunale*), which supplied a gross amount of 27,477 judgments concerning the labour law division (i.e. *Sezione lavoro*). The mentioned decisions were delivered in the following file formats: real-pdf, docx, doc, docm. A subset of 4,804 judgments was provided with a specific label. The total number of labels is 309. It’s important to notice how the distribution of judgments on the different labels is skewed, as shown in Figure 1.

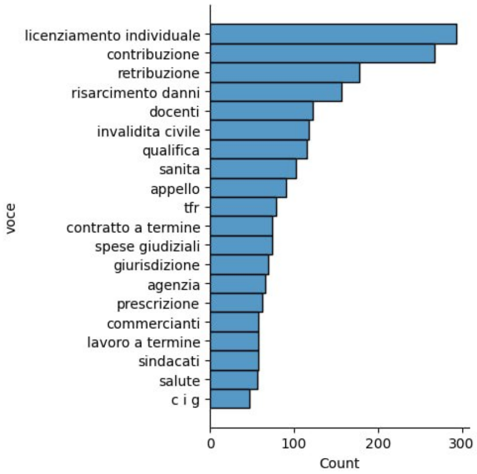

Figure 1: Distribution of the first 20 labels of Turin court’s dataset. The labels shown in the figure are a subset of the 309 labels of the Turin dataset used for classifying the judgments, in order: individual dismissal, contribution, salary, damage compensation, teachers, legal disability, qualification, healthcare, appeal, severance pay, fixed-term contract, litigation fees, limitation, traders, fixed-term employment, labour union, fixed-term job, trade union, health, unemployment benefit.

3. Methodology

In order to digitize legal archives and provide a system that can be easily used by Judges and UPPs, a platform is being developed to host the resources and processes them in a way that automatically catalogues and indexes the collection. Semantic information extraction allows navigation by metadata and similarity.

3.1. Information Retrieval and Segmentation

An important step towards the achievement of the various tasks discussed below is the automatic extraction and segmentation of text. The approach used to structure the data was to mirror the segmentation pattern used by domain experts. The following is a brief pipeline of the operations that involved this task: 1) Conversion of judgments to *.docx* format: we decided to converge files with different formats to a single data representation to facilitate the text extraction process. 2) Removal of less informative paragraphs: Stakeholder’s information was disregarded in order to perform classification tasks using clean data. 3) Structuring of textual content in JSON format.

The process of information extraction led to the definition of two different JSON representations for each judgment, by metadata and by content. The following metadata was collected: court, section, subject, judgment code-year, NRG code-year. The content is organized as follows: 1) *Oggetto*: The subject matter of the case addressed by the judgment. It is typically very informative about the subject to which the judgment belongs, 2) *Conclusioni*: Some indications about the conclusion of the proceedings concerning the parties, 3) *Svolgimento del processo*: The central part of the judgment where the facts of the case and the reasons for the decision made by the judge are addressed, 4) *P.Q.M.*: The final verdict, 5) *Voce*: the indication of the label, where present. We were able to obtain a labelled dataset on Turin judgments through a matching process. Given a list of indexes of items, matching was conducted by comparing the judgments' code-year and NRG code-year to those reported within the indexes. An index represents a file containing all references to a case organized by *voce.* Each case is associated with an NRG code which can be found in the *oggetto* section of the judgment.

3.2. Preprocessing

To enhance the quality of the data and preserve its privacy, it was necessary to perform a preprocessing pipeline, consisting of 1) pseudo-anonymization, 2) conversion to lowercase, 3) removal of special characters (accents, punctuation symbols and non-uft8 characters), 4) removal of URLs and HTML tags, 5) conversion of word numbers to their numeric form, 6) removal of stopwords, 7) lemmatization. The pseudo-anonymization phase overwrites proper names, surnames and tax codes. This phase allows us to use the dataset without directly processing this kind of personal data of the people involved in the judgments. In addition, the use of specific tags, that replace the data just mentioned, maintains the semantics of the sentence and the relationships between entities inside the text. Subsequently, the text was cleaned of irrelevant components so as not to compromise the previous phase, since some sensitive information includes stopwords, capital characters and punctuation symbols. The lemmatisation phase was performed using Morph-it! [13] to speed up the computation on the Italian dataset.

3.3. Classification

Datasets. One of the main tasks addressed in this work is the automatic classification of judgments. Considering the imbalance of the dataset, the tests on classification were conducted with a limited dataset; in fact, not all judgments from the Turin dataset were taken into account. Specifically, two main corpora were produced, which we will refer to below as “corpus_8_classes" and “corpus_15_classes", the first generated using 800 judgments distributed equally over 8 entries and the latter with 1,872 judgments distributed over 15 entries. The entries considered are, in order, the first 15 illustrated in Figure 1.

For the creation of the datasets we employed some kind of vector space modelling techniques. Starting from these representations we trained some models. For major details, results and discussion are visible in section 4.1. Data used in this paper for the creation of the datasets matches with the following content of the JSON fields: “Oggetto", “conclusioni", “svolgimento del processo", “P.Q.M" and “voce". Starting from these fields we defined 8 different datasets, 4 for each corpus. At the end of the preprocessing pipeline on the “corpus_8_classes", the use of TF [14] and TF-IDF [15] led us to define two sparse matrices of 23,618 x 800 dimension, while on the second corpus, the result of the TF and TF-IDF vectorization returned two 28,319 x 1,872 sparse matrices. To have a recent comparison regarding the state of the art on the embeddings representation, the remaining 4 datasets were created using the following resources:

- **Doc2Vec:** Doc2Vec [16] is an unsupervised neural network model that learns fixed-length feature vectors for representing textual data. The network architecture, like for word2vec [17], provides two different algorithms for the embeddings generation: “Continuous Bag of Words” (CBOW) e “Skip-Gram'(SG)”[17]. For the learning process, we considered the first one, CBOW, which implementation is visible in the python library: gensim.models.Doc2Vec[1]. The model, after a preprocessing step, specifically required for this implementation of the algorithm, was trained for 30 epochs with the following hyperparameters: vector_size = 300, negative=5, hs=0,min_count=2,sample=0, alpha=0.025, min_alpha=0.001.
- **Italian-Legal_bert**: Italian-Legal_bert [18] is a version of a pretrained BERT-BASED [19] model (ITALIAN XXL BERT[2]) trained on italian legal texts. The embeddings of this model are obtained running an additional round of training for 4 epochs on a 3,7GB preprocessed text from the National Jurisprudential Archive using the Huggingface PyTorch-Transformers library[3].

Models. Our classification work focused more on data representation than on the use of neural models and finetuning of networks. A first experiment has seen the use

[1] https://radimrehurek.com/gensim/models/doc2vec.html

[2] https://huggingface.co/dbmdz/bert-base-italian-xxl-cased

[3] https://huggingface.co/docs/transformers/index

of a multiclass SVM [20] as a baseline model. Assuming nonlinearly separable data, we trained the SVM model using an “rbf" kernel-trick[4]. In the second order, considering the dimensions of the datasets, we conducted some tests using a Logistic Regression[5] model with a “lbfgs" solver. In presence of sparse and poor data, these models tend to show the same behaviour. Furthermore, we considered a Random Forest classifier[21] with max 2,000 trees, which, instead, results more efficiently on datasets with a limited number of features. Finally, the same tests were repeated running an Ensemble Learning task with a simple Voting classifier[6] using all the previous models.

3.4. Similarity

Judgments contain a set of sections that describe the focal points of the document, specifically parts (i.e. *Parti*), subject matter (i.e. *Oggetto*), fact (i.e. *Fatto*), reasoning (i.e. *Motivi*) and decision (i.e. *Decisione*). These sections represent a substantial amount of information meticulously describing judgments, some of which share characteristics and suggest similarity and relatedness between judgments on multiple levels. Sections include citations (e.g. judgments, legal articles) that relate resources, especially judgments with the same (or similar) citations that can discuss similar issues and treat the fact in a similar manner. Citations can be considered differently depending on their position in the text, domain, and specific moment in time. These relationships between resources provided the input to develop an additional feature for the dataset treatment in order to provide additional functionality consisting of semantic similarity search within the online catalogue of judgments. The domain of application constrains the use of recurring structures and terminologies in judgments [22] that guides the treatment of data from an entropy perspective with the aim of finding the most relevant components in the text that constitute the discriminating features. A hybrid approach oriented to the analysis of know-how and reproduction of some methodologies applied by domain experts was opted for. The goal is the completion of the task by enriching it with an attempt to provide an explanation of the results provided by the system would allow greater transparency of the platform.

3.5. Principles of Law

In Case-law. Defining what can be considered a principle of law is not straightforward. Whereas the country considered in our analysis abide by a common or a civil law legal system we found an across-the-board shared definition, with a similar gauge. Principles of law *are used to fill the gap existing between what is defined by legal doctrine and reality.* The first can be an imposition [23] as it happened in many countries that were colonized, or [24] with sets of law written centuries before. In Italy, a Country following a civil law approach to legislation, principles of law are: *an official interpretation given by the Supreme Court (i.e.* Corte di Cassazione*), whose scope is to give a generalized interpretation and application of a rule.*

In Computer Science. In this project, as mentioned in the previous paragraphs in this section, we approached topics as *Classification* and *Similarity*. Our hypothesis is that given a correct set of methods to recognize the ways in which principles of law are expressed in a sentence, we are able to find new metadata, useful in the development of the tasks before mentioned.

4. Results

4.1. Classification

In this section, we will show in more detail all the results of our experiments. All data visualized in the following tables are derived by applying a 10-fold cross-validation method on the datasets and models defined in the previous section. Table 1 shows the results of the main evaluation metrics we considered: accuracy, precision, and recall. Reading the table by columns, as depicted, the Random Forest classifier (2,000 trees) is the model with the best results. The limited structure of these datasets has led to more performing results in that model which, in general, tends to decrease its performance in case of the number of classes and features increases. It is interesting to note from Table 1 how the dataset that responds with higher performance is the one obtained using *doc2Vec*, in fact, all the models applied to this dataset return high precision and recall values.

Table 2 describes the results of the models on the “corpus_15_classes". From a first observation it can be seen how the nature of this corpus has had a significant impact on the performance of the models which are decreased, compared to the previous test. All the results obtained from the different models, except for the dataset created by *doc2vec* embeddings, reflect our expectations about the decreasing of the performances. In both corpora, italian-legal-BERT reported the worst results, due to the excessive sparseness of the data, while *doc2vec* appears to guarantee excellent performance even with the baseline models.

[4] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html

[5] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

[6] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.VotingClassifier.html

Test 1- corpus 8 classes				
Dataset	**Random forest**	**SVM**	**Logistic Regression**	**Enseble Voting**
		Average accuracy		
TF	0.900	0.806	0.868	0.893
TF-IDF	0.925	0.906	0.900	0.906
Ita-legal BERT	0.856	0.793	0.887	0.85
Doc2Vec	0.981	0.981	0.975	0.981
		Average precision		
TF	0.895	0.822	0.871	0.894
TF-IDF	0.921	0.910	0.904	0.910
Ita-legal BERT	0.861	0.810	0.894	0.852
Doc2Vec	0.983	0.981	0.975	0.981
		Average recall		
TF	0.899	0.805	0.871	0.891
TF-IDF	0.923	0.904	0.899	0.903
Ita-legal BERT	0.865	0.809	0.890	0.857
Doc2Vec	0.980	0.980	0.975	0.980
		Average f1 score		
TF	0.893	0.811	0.868	0.891
TF-IDF	0.919	0.904	0.898	0.903
Ita-legal BERT	0.855	0.795	0.890	0.852
Doc2Vec	0.980	0.979	0.973	0.979

Table 1
Evaluation of the performances of the four datasets derived by the "corpus_8_classes"

Test 2 - corpus 15 classes				
Dataset	**Random forest**	**SVM**	**Logistic Regression**	**Enseble Voting**
		Average accuracy		
TF	0.784	0.776	0.802	0.816
TF-IDF	0.784	0.805	0.794	0.808
Ita-legal BERT	0.722	0.714	0.786	0.741
Doc2Vec	0.914	0.954	0.962	0.957
		Average precision		
TF	0.859	0.829	0.791	0.765
TF-IDF	0.865	0.859	0.837	0.853
Ita-legal BERT	0.773	0.835	0.766	0.853
Doc2Vec	0.943	0.966	0.972	0.965
		Average recall		
TF	0.730	0.723	0.785	0.788
TF-IDF	0.726	0.744	0.737	0.750
Ita-legal BERT	0.640	0.595	0.748	0.750
Doc2Vec	0.878	0.945	0.955	0.955
		Average f1 score		
TF	0.752	0.756	0.782	0.788
TF-IDF	0.745	0.773	0.751	0.768
Ita-legal BERT	0.660	0.602	0.752	0.768
Doc2Vec	0.898	0.954	0.962	0.955

Table 2
Evaluation of the performances of the four datasets derived by the "corpus_15_classes"

5. Conclusions and Future Work

In this paper, we presented a pipeline for providing a system that facilitates some of the activities of magistrates and UPP's relating to the automatic classification, semantic information research, and navigation of legal texts by metadata and similarity. We explored some baseline solutions focusing mainly on data representation than on the use of state-of-the-art neural models and fine-tuning of networks. Although the composition of the corpora and the lack of data, we obtained excellent results showing that it is possible to achieve good performance even using simple models, however in the future, there would be anything but baseline models to explore and evaluate. Another approach to the classification task could be a combination of similarity techniques and machine learning models we will consider in future work. In fact, the use of some similarity metrics could help us to outperform the classification models, if two judgments are more similar, it is more likely that they belong to the same category.

In regards to the principles of law, we speculate the possibility of identifying relationships of interest, useful to model the connection between entities explicitly stated in a legal text such as a judgment.

Acknowledgments

The 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 Ministry of Justice as a side of the PON Governance and Institutional Capacity 2014-2020 (Axis I – Action 1.4.1) and implemented in synergy with the interventions envisaged by the National Recovery and Resilience Plan (NRRP) in support to the justice reform.

References

[1] G. Contissa, F. Godano, G. Sartor, Computation, Cybernetics and the Law at the Origins of Legal Informatics, Springer, Cham, 2021, pp. 91–110. doi:10.1007/978-3-030-54522-2_7.

[2] L. Robaldo, S. Villata, A. Wyner, M. Grabmair, Introduction for artificial intelligence and law: special issue "natural language processing for legal texts", Artif. Intell. Law 27 (2019) 113–115. doi:10.1007/s10506-019-09251-2.

[3] M. Derlén, J. Lindholm, Is it Good Law? Network Analysis and the CJEU's Internal Market Jurisprudence, Journal of International Economic Law 20 (2017) 257–277.

[4] G. Sartor, P. Santin, D. Audrito, E. Sulis, L. Di Caro, Automated extraction and representation of citation network: A cjeu case-study, in: R. Guizzardi, B. Neumayr (Eds.), Advances in Conceptual Modeling, Springer, Cham, 2022, pp. 102–111.

[5] F. Galli, G. Grundler, A. Fidelangeli, A. Galassi, F. Lagioia, E. Palmieri, F. Ruggeri, G. Sartor, P. Torroni, Predicting outcomes of italian vat decisions 1, in: Legal Knowledge and Information Systems, IOS Press, 2022, pp. 188–193.

[6] R. A. Shaikh, T. Sahu, V. Anand, Predicting outcomes of legal cases based on legal factors using classifiers, Procedia Computer Science 167 (2020) 2393–2402. doi:10.1016/j.procs.2020.03.292.

[7] R. Nai, E. Sulis, R. Meo, Public procurement fraud detection and artificial intelligence techniques: a literature review, in: Symeonidou et al. (Ed.), EKAW, volume 3256 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3256/km4law4.pdf.

[8] R. Meo, R. Nai, E. Sulis, Explainable, interpretable, trustworthy, responsible, ethical, fair, verifiable AI... what's next?, in: S. Chiusano, T. Cerquitelli, R. Wrembel (Eds.), ADBIS 2022, Turin, Italy, September 5-8, 2022, Proceedings, volume 13389 of *LNCS*, Springer, 2022, pp. 25–34. doi:10.1007/978-3-031-15740-0_3.

[9] G. Grundler, P. Santin, A. Galassi, F. Galli, F. Godano, F. Lagioia, E. Palmieri, F. Ruggeri, G. Sartor, P. Torroni, Detecting arguments in CJEU decisions on fiscal state aid, in: Proc. of the 9th Workshop on Argument Mining, International Conference on Computational Linguistics, Korea, 2022, pp. 143–157. URL: https://aclanthology.org/2022.argmining-1.14.

[10] M. Palmirani, F. Vitali, Akoma-Ntoso for Legal Documents, Springer Netherlands, Dordrecht, 2011, pp. 75–100.

[11] D. Audrito, E. Sulis, L. Humphreys, L. Di Caro, Analogical lightweight ontology of eu criminal procedural rights in judicial cooperation, Artificial Intelligence and Law (2022) 1–24.

[12] E. Sulis, L. B. Humphreys, D. Audrito, L. D. Caro, Exploiting textual similarity techniques in harmonization of laws, in: S. B. et al. (Ed.), AIxIA 2021, volume 13196 of *LNCS*, Springer, 2021, pp. 185–197. doi:10.1007/978-3-031-08421-8_13.

[13] E. Zanchetta, M. Baroni, Morph-it, A free corpus-based morphological resource for the Italian language. Corpus Linguistics 1 (2005) 2005.

[14] H. P. Luhn, The automatic creation of literature abstracts, IBM J. Res. Dev. 2 (1958) 159–165.

[15] K. S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Documentation 60 (2021) 493–502.

[16] Q. V. Le, T. Mikolov, Distributed representations of sentences and documents, in: International Conference on Machine Learning, 2014.

[17] T. Mikolov, K. Chen, G. S. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, 2013.

[18] D. Licari, G. Comandè, ITALIAN-LEGAL-BERT: A Pre-trained Transformer Language Model for Italian Law, in: Symeonidou et al. (Ed.), EKAW, volume 3256 of *CEUR Workshop Proceedings*, CEUR, Bozen-Bolzano, Italy, 2022. URL: https://ceur-ws.org/Vol-3256/#km4law3.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: ACL: HLT, Vol. 1, ACL, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[20] B. E. Boser, I. M. Guyon, V. N. Vapnik, A training algorithm for optimal margin classifiers, in: Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144–152.

[21] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[22] X. Li, J. Gao, D. Inkpen, W. Alschner, Detecting relevant differences between similar legal texts, in: Proceedings of the Natural Legal Language Processing Workshop 2022, 2022, pp. 256–264.

[23] N. L. Mahao, Can african juridical principles redeem and legitimise contemporary human rights jurisprudence?, Comparative and International Law Journal of Southern Africa 49 (2016) 455–476.

[24] F. Galindo, Juridical principles for juridical applications. the derinfo methodology, in: D. Karagiannis (Ed.), Database and Expert Systems Applications, Springer Vienna, Vienna, 1991, pp. 425–430.