Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).
doi:10.3233/FAIA230972

Assessing Ocean's Legal Protection Using AI: A New Dataset and a BERT-Based Classifier

Youssef AL MOUATAMID [a,d,1], Jihad ZAHIR [a,b], Marie BONNIN [b] and Hajar MOUSANNIF [a]

[a] *LISI Laboratory, Cadi Ayyad University, Marrakesh, Morocco*
[b] *UMMISCO, IRD France Nord, Bondy F-93143*
[c] *IRD, Univ Brest, CNRS, Ifremer, LEMAR, Plouzane F-29280, France*
[d] *Univ Brest, LEMAR, Plouzane F-29280, France*

ORCiD ID: Jihad Zahir https://orcid.org/0000-0002-5344-8277, Marie Bonnin https://orcid.org/0000-0003-4140-2439, Hajar Mousannif https://orcid.org/0000-0002-1307-4215

Abstract. This paper aims to address the challenge of using artificial intelligence for empirical legal studies. We introduce a new annotated dataset on French marine environmental law dealing with definitions, bans, sanctions, and controls on living (turtles and seabirds) and non-living (plastic bags and straws) subjects. The annotation has been produced by law students and validated by a legal expert. Based on the developed dataset, we train a CamemBERT-based classifier which accurately predicts the class of a given legal article according to the pre-defined classes within the dataset we have created. The proposed training set and the resulting trained model provide a better interpretation and accessibility of legal texts to specialists and the general public, based on findings from legal studies and on natural language processing techniques.

Keywords. Classification, Machine learning, datasets, CamemBERT, marine environmental law, empirical legal studies

1. Introduction

Recently, we have been witnessing an effervescence in the field of environmental law due to the implementation of new, international and national measures, to cope with the exacerbation of anthropogenic threats weighing on the natural environment[1]. The increased legal protection of marine environments brings forth two main challenges: limited text accessibility and complex spatio-temporal analysis. Despite the online availability of legal documents, their complexity and challenging interpretation hinder accessibility for non-experts, resulting in a gap between existing laws and their practical implementation. This has prompted several international organizations to call for the development of legal indicators[2][3]. However, to inform these, researchers typically rely on classical empir-

[1] Corresponding Author: Youssef AL MOUATAMID, youssef.almouatamid@univ-brest.fr

ical techniques applied to extensive legal document collections, demanding substantial time and effort. Recent advances in Artificial Intelligence (AI), particularly in natural language processing (NLP), are opening unprecedented opportunities for better interpretation and accessibility of law. Our interdisciplinary work takes a pioneering step in automating the assessment of environmental legal protection, providing a new paradigm in conducting empirical legal studies. The main objective of this research is to make use of NLP techniques to extract information from positive environmental law. In order to analyze the marine environmental protection measures foreseen by the law, it is necessary to ascertain its existence and assess its features. That is why we have carefully selected a set of legal documents and sought in their articles to identify the characteristics of the analyzed legal rule. We therefore obtained a dataset of annotated articles which we used to train a classifier that predicts the class of the articles constituting a legal document.

Multiple researches have applied text classification to legal texts [4,5]. Different machine learning algorithms and methods have been used in legal text classification, including CNN [4] and BERT [6,7]. Variants of BERT that are specifically pre-trained on legal texts have also been proposed in the literature. For instance, JuriBert [7] was pretrained on legal texts, from *Légifrance*, and Court's decisions and Claimant's pleadings from the Court of Cassation. However, existing works on supervised classification in the legal domain have generally focused on case law [8] and none of them has tackled environmental law nor used the same classes as those proposed in this article.

As for datasets, resources addressing French legal texts are very limited. Existing datasets [9,10,11] are mainly focused on jurisprudence, and to the best of our knowledge, a dataset of annotated articles on marine environmental law in French does not exist.

This paper makes two main contributions. Firstly, we introduce a new in-house dataset of annotated articles: a prerequisite for training models using supervised classification. Secondly, we use CamemBERT[12] language model to build a classifier which accurately categorizes articles into one of five classes *(Definition, Ban, Sanction, Control or Miscellaneous)*.

2. Methodology

2.1. Dataset

Our objective is to build a dataset reflecting legal protection of marine environment. Therefore, we focused on legal documents addressing both living (turtles and seabirds) and non-living (plastic bags and straws) aspects of the issue. In our search for legal documents, we focused on those from French-speaking countries and those that use French as an administrative language, as outlined in Table 1.

2.1.1. Source Datasets / Dataset Collection:

The overall process begins with data collection following these steps:

1. **Theme-based search of documents** in *Faolex* database [2] (e.g., plastic bags).
2. **Download and OCR**: We use Optical Character Recognition (OCR) on downloaded documents to convert documents in pdf and images to text.

[2] https://www.fao.org/faolex

3. **Articles segmentation**: We extract textual content from the document using regular expressions to identify potential articles based on patterns, and concatenate the sections of text that represent an article.
4. **Articles filtering:** If the theme is present in the document title, all articles are retained. Otherwise, a regular expression is used to select theme-relevant articles. This step proves essential for cases where legal texts encompass a broader scope, such as fishing codes, with only select articles related to the analyzed theme.

Collected legal documents include statutes, decrees, dahirs and orders. Table 1 provides an overview of these documents.

Table 1. The number of documents and articles per country and theme in the datset.

Theme	Countries	Num of docs	Num of articles
Plastic bags	Belgium, Burkina Faso, Cameroon, Democratic Republic of the Congo, Djibouti, France, Gabon, Ivory Coast, Madagascar, Mauritania, Monaco, Morocco, Republic of the Congo, Senegal, Tunisia	73	298
Plastic straws	France, Monaco	3	14
Turtles	France, French Guiana, Guadeloupe, Guinea, Madagascar, New Caledonia, Senegal, Tunisia	11	48
Seabirds	Belgium, Burundi, Cameroon, Central African Republic, Djibouti, France, Madagascar, Mauritania, New Caledonia, Republic of the Congo, Senegal	18	58

2.1.2. Labels:

The choice of labels stems from legal reflection aimed at identifying potential indicators for addressing this intricate matter. Since the ultimate objective is to evaluate the role of the law in protecting the oceans, we have identified five rules, which are detailed below.

- Definition: Explicitly states and clarifies the object and scope of the ban.
- Ban: Legal prohibition on environmentally damaging human activities.
- Sanction: A penalty for breaking environmental laws.
- Control: The process to verify the existence and compliance, and to inspect, oversee, restrict or prohibit as well as impose penalties.
- Miscellaneous: Includes articles that do not fit the previous categories.

2.1.3. Annotation Strategy:

Our annotation strategy consists of having each article doubly-annotated by two law students, and only keeping those where both annotators agree on the chosen label. After this initial step, the resulting dataset was subsequently validated by a legal expert.
In the end, we obtained a labeled benchmark dataset of 418 articles. Table 2 shows the number of articles per label of our dataset.

2.2. Models

In this work, we use CamemBERT[12], a French version of BERT[6], to build a classifier which takes, as input, an article and classifies it into one of the five pre-defined classes.

Table 2. The number of articles by class and theme in the dataset.

Theme	Definitions	Sanctions	Bans	Control	Miscellaneous	Total
Plastic bags	83	68	66	21	60	298
Plastic straws	2	1	9	2	0	14
Turtles	1	5	21	3	18	48
Seabirds	0	25	18	15	0	58
	86	99	114	41	78	418

Conveniently, with BERT models, we can fine-tune and train the model on a relatively small dataset to get decent results, and this suits our case perfectly.

2.3. Pre-processing

Before feeding the articles into the model, we perform the following pre-processing steps:

- Article number removal: We removed the article number (*Art1:*), which appears at the beginning, since we are only interested in the content.
- Special character removal: We deleted newline characters, carriage returns, non-breaking space and single, curly, and double quotes.
- Multiple whitespaces: We replaced consecutive whitespace characters with a single space and removed leading and trailing whitespace from the entire text.

3. Experiment

To provide a better understanding and evaluation of the proposed model, we use different metrics: accuracy, recall, precision, and F1-score, in addition to the Matthews Correlation Coefficient (*MCC*), which makes up for the shortcomings in the other four metrics.

3.1. Evaluation

In order to test and evaluate the classifier's performance, we confronted it with a dataset, that the model has never seen. It was provided by a legal expert who was working on legal articles related to plastic bags (not exclusively in French). Thus, we selected those that were in French and that corresponded to our predefined classes *(definitions, bans, sanctions, and controls)* to validate the proposed model.
To assess the model's performance on themes that are different from those used in training *(plastic, turtle and seabird)*, we added to the test set a selection of articles on Canadian pesticides presented in Table 3

Table 3. The number of articles by class and theme in the test set.

Theme	Definitions	Sanctions	Bans	Control	Miscellaneous	Total
Plastic bags	12	14	4	16	30	76
Pesticides	0	47	81	0	0	128
	12	61	85	16	30	204

3.2. Results and discussion

The model was trained for 25 epochs with a batch size of 32, and a maximum token length of 512. In these settings, the model reached a validation accuracy of 91%. We have compared the proposed classifier with three other models: SVM, Random Forest, and LightGBM. The proposed model performs reasonably well on test set, achieving an accuracy of 95%, as well as an MCC of 0.92, outperforming the other baseline classifiers in all classes. In fact, the results of the confusion matrices fig 1, and tables 4 and 5 lead

Table 4. The accuracy and the MCC of the test set

Model	Accuracy	MCC
SVM	0.48	0.45
Random Forest	0.39	0.38
LightGBM	0.46	0.45
Our Model	**0.95**	**0.92**

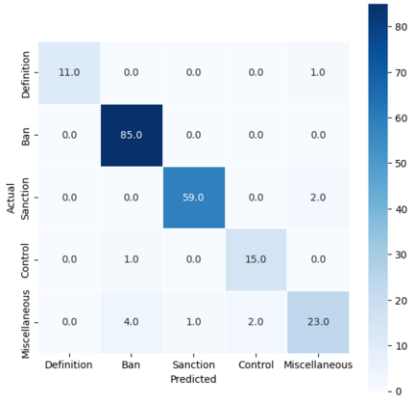

Figure 1. The confusion matrix of test set

us to believe that the proposed model performs well in all classes, regardless of theme, except for the *"Miscellaneous"* class where performance is comparatively lower.
This is linked to the nature of the *"Miscellaneous"* class and to the subjectivity of the annotator. For instance, an article dealing with a conditional permission may be considered by one annotator as belonging to the *"Miscellaneous"* class, while another will consider it to fall under the *"Ban"* class because of the implicit prohibition it implies. On the other hand, an article may be written in a complex way and raise some ambiguity as to the label, even for specialists. As previously stated, certain articles were assigned different labels depending on the annotator, and only those agreed upon by both law students and validated by the expert were retained.
Furthermore, the labeled articles in our dataset pertain to a specific topic. However, we may find other articles that match our labels within a legal document, such as the *"Fisheries Code"*. Although these articles do not align with the themes we have explored so far, they may be relevant for future work, thus highlighting the utility of our model and marking an initial step towards a legal observatory of the marine environment.

Table 5. The test set evaluation

	SVM			Random Forest			LightGBM			Our model			
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	Support
Definitions	1.00	0.25	0.40	1.00	0.25	0.40	1.00	0.25	0.40	**1.00**	**0.92**	**0.96**	12
Bans	0.94	0.16	0.28	0.75	0.03	0.06	0.89	0.09	0.16	**0.94**	**1.00**	**0.97**	85
Sanctions	**1.00**	0.73	0.84	**1.00**	0.65	0.79	**1.00**	0.70	0.82	0.98	**0.97**	**0.98**	61
Controls	0.50	**1.00**	0.67	0.67	**1.00**	0.80	0.50	**1.00**	0.67	**0.88**	0.94	**0.91**	16
Miscellaneous	0.22	**0.94**	0.36	0.19	**0.94**	0.32	0.21	**0.94**	0.34	**0.88**	0.77	**0.82**	30

4. Conclusion

In this paper, we have presented two main contributions, a new annotated dataset of marine environmental law, and a classifier which is able to differentiate definition, ban, sanction and control articles.
Due to the paucity of datasets on marine environmental law, especially those in French, we were urged to create one dealing with plastic bags and straw, turtles and seabirds in some French-speaking countries, which we annotated with the help of experts into five classes *(definitions, bans, sanctions, controls and miscellaneous)*. Following the dataset's creation, we built a CamemBERT-based model to classify the articles as per the predefined classes, which we validated using a test set, and the results were satisfactory.
Moreover, as outlined in the previous section, the encouraging results on various themes, including those that were not part of our training, represent a first step towards an observatory of marine environmental law.

5. Acknowledgments

This work has been supported by AIME project, and ISblue project, Interdisciplinary graduate school for the blue planet (ANR-17-EURE-0015) and co-funded by a grant from the French government under the program "*Investissements d'Avenir*" embedded in France 2030.

References

[1] Jouffray JB, Blasiak R, Norström AV, Österblom H, Nyström M. The Blue Acceleration: The Trajectory of Human Expansion into the Ocean. One Earth. 2020;2(1):43-54.
[2] Thirteenth meeting of the Conference of the Parties, 4-17 December 2016, Decision XIII/28, Indicators for the Strategic Plan for Biodiversity 2011-2020 and the Aichi Biodiversity Targets; 2016.
[3] Official Journal of the European Union L164/19 of 25.06.2008, p. 19-40, Art. 5 and 10; 2008.
[4] Hammami E, Akermi I, Faiz R, Boughanem M. Deep Learning for French Legal Data Categorization. In: Schewe KD, Singh NK, editors. Model and Data Engineering, Medi 2019. vol. 11815. Cham: Springer International Publishing Ag; 2019. p. 96-105. ISSN: 0302-9743 WOS:000567294500007.
[5] Cardellino C, Alemany LA, Teruel M, Villata S. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In: Proc Int Conf Artif Intell Law. New York, NY, USA: Association for Computing Machinery; 2017. p. 9-18. Journal Abbreviation: Proc Int Conf Artif Intell Law.
[6] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv; 2019. ArXiv:1810.04805 [cs].
[7] Chalkidis I, Fergadiotis M, Malakasiotis P, Aletras N, Androutsopoulos I. LEGAL-BERT: The Muppets straight out of Law School. In: Findings of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics; 2020. p. 2898-904.
[8] Mok JR, Mok WY, Mok RV. Sentence Classification for Contract Law Cases: A Natural Language Processing Approach. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law. ICAIL '21. New York, NY, USA: Association for Computing Machinery; 2021. p. 260–261.
[9] Louis A, Spanakis G. A Statutory Article Retrieval Dataset in French. arXiv; 2022. ArXiv:2108.11792.
[10] Niklaus J, Chalkidis I, Stürmer M. Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark. arXiv; 2021. ArXiv:2110.00806 [cs].
[11] French Monolingual legal corpus from Official Journal of France; 2019. Available from: http://data.europa.eu/88u/dataset/elrc_2501.
[12] Martin L, Muller B, Ortiz Suárez PJ, Dupont Y, Romary L, de la Clergerie É, et al. CamemBERT: a Tasty French Language Model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics; 2020. p. 7203-19.