Monitoring Sustainable Development Goals in European Legislation using Hybrid AI

Michele Corazza
Monica Palmirani
Franco M. T. Gatti
Salvatore Sapienza
michele.corazza2@unibo.it
monica.palmirani@unibo.it
franco.gatti3@unibo.it
salvatore.sapienza@unibo.it
CIRSFID ALMA-AI, University of Bologna
Bologna, Italy

ABSTRACT

There is an urgency to detect in legal acts the corresponding provisions where a policy is implemented, to track its evolution over time, to measure the effectiveness of the norms, and to evaluate the impact on society. From this perspective, the Sustainable Development Goals program (SDG) provides a fundamental instrument for monitoring ground basis pillar of the world wide policies. In this work, we propose a method which leverages both the structural nature of legislative documents in AKN-XML and unsupervised machine learning to perform a match between individual articles and definitions and the 2030 Agenda for Sustainable Development Goals, yielding a more fine-grained annotation of individual articles and definitions, instead of the preexisting document-level annotation. Our work provides better traceability of the SDGs policies in the EU legislation permitting the legislator to detect the articles where the association is weakest. During the legal drafting, our tool could be integrated into the editor to suggest better legal definitions for improving the implementation of the SDGs.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; **Natural language processing**; • **Information systems** → **Digital libraries and archives**.

KEYWORDS

Akoma Ntoso, LegalXML, Policy tracking, Sustainable Development Goals

ACM Reference Format:
Michele Corazza, Monica Palmirani, Franco M. T. Gatti, and Salvatore Sapienza. 2024. Monitoring Sustainable Development Goals in European Legislation using Hybrid AI. In *17th International Conference on Theory and Practice of Electronic Governance (ICEGOV 2024), October 01–04, 2024, Pretoria, South Africa.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3680127.3680223

This work is licensed under a Creative Commons Attribution International 4.0 License.

ICEGOV 2024, October 01–04, 2024, Pretoria, South Africa
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1780-2/24/10
https://doi.org/10.1145/3680127.3680223

1 INTRODUCTION

In many deliberative institutions there is an urgency to detect in legal acts the corresponding provisions where a policy is implemented, to track its evolution over time, to measure the effectiveness of the norms, and to evaluate the impact on society. From this perspective the Sustainable Development Goals programme (SDG) provides a fundamental instrument for monitoring ground basis pillar of the world wide policies. On the other hand, one of the emerging applications of artificial intelligence in the legislative domain aims to find the correspondence between the law and the policies defined by a deliberative body. In 2017, the European Commission's Joint Research Centre started a policy mapping of a large set of Juncker Commission's actions, identifying their linkages to the targets composing the 17 goals included in the 2030 Agenda for Sustainable Development Goals (SDG) [9]. These EU legislative documents were manually mapped and further textual analyses were carried out by a group of experts, and at the end of the process the database obtained was finalised in cooperation with the Directorate-General pertaining to each policy initiative. Further reviews, and the implementation of text mining and natural language processing techniques, led to an automatisation of the policy mapping process that was implemented starting from the documents of the Von Der Leyen Commission (2020-2024). In 2022 the authors performed a project in collaboration with the European DG Informatics and a database of EU Legislation was built with the support of the European Publication Office. The database consists of 1568 regulations and directives between 2010-2019 converted into XML using the legal standard Akoma Ntoso (AKN). Among the documents available in this database, we decided to consider only documents produced during the *Juncker Commission* (2015-2019) i.e. documents that have been manually annotated by experts, in order to have a comparison with the ground truth dataset.

However, this dataset classifies the European legislation at the document level, while legal experts are interested in knowing precisely the article or the portion of the text connected to the SDGs in order to track policies and, in case of modifications, to detect the improvements over time. Secondly, if the legislator intends to

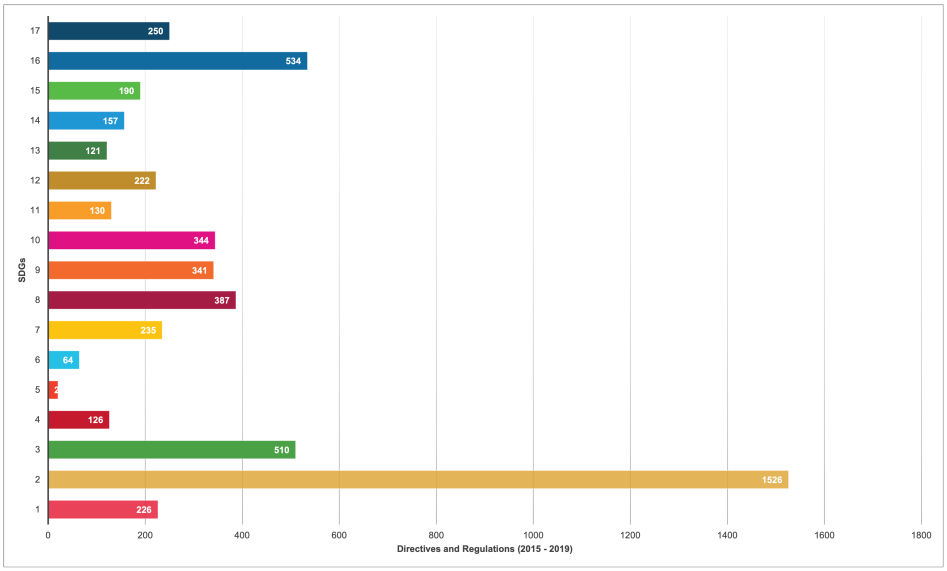

Figure 1: Number of documents per each SDG manually mapped by JRC.

reinforce the SDGs indicators inside the legal text, it is important to know exactly what provisions are not sufficient or "weak" in order to reinforce the strategy. Thirdly, legislative documents include modifications/derogations to other documents and the quoted text influences the classification of SDGs. On the other hand, we could foster some qualified parts of the legislative document that are more descriptive of the scope and objective, like legal definitions or the first articles. Finally, AI methods are usually applied at a document level. Given limitations in token-handling, legal documents are usually segmented in chunks, with no difference between parts (preamble, definition, articles, final provisions, etc.). This implies that the hierarchical structure of the legal document is lost in the AI-friendly representation of the legal text. For these reasons, the main research questions are the following:

- RQ1: which partitions of the legislative document are connected with the targets of the SDGs taxonomy? We intend to reach a better granularity of the classification, in order to provide an accurate instrument to the legislator.
- RQ2: what target of SDGs are associated with legal definitions? We intend to detect how legal definitions are connected to SDGs for providing a mapping between the European legislation and the United Nations strategies.
- RQ3: is the structured text, especially in AKN-XML, better for SDGs classification? We intend to exploit the structure of the legislative text for connecting relevant portions of the text like the articles, which would usually be split in chunks in AI methods, losing the hierarchy.

We use Akoma Ntoso[1] because it is the official standard in EU institutions for modelling the legislative documents (AKN4EU[2]) and similarly is the standard for the documents in the UN (AKN4UN[3]).

[1]http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html
[2]https://op.europa.eu/it/web/eu-vocabularies/akn4eu
[3]https://unsceb.org/unsif-akn4un, https://unsceb-hlcm.github.io/

2 RELATED WORK

The landscape of Natural Language Processing (NLP) has seen rapid advancements in recent years due to the role of transformer models[24], and in particular BERT[7]. In the legal domain, more specialised models have been created[3], which are further fine-tuned from BERT or or other encoder models using text from the legal domain. One example is LEGAL-BERT[4], which is trained on EU, UK and US legislative documents, as well as court documents, including ones from the European Court of Justice. Another such model is custom LEGAL-BERT[27], trained on the Harvard Law Library case law corpus. The PoL model was trained on the Pile-of-Law dataset, obtained from 35 different sources for the English language [12]. Beside pre-trained models, a multitude of tasks and models have been applied to the legal domain. Most of these are related to the judiciary, such as models performing outcome prediction for court rulings. In [8] the authors use a transformer to classify the case's violated articles on a global consistency graph. In [26], the authors propose a pre-trained model for judgment prediction, similar case retrieval, legal question answering and legal reading comprehension in Chinese. In [16] a new dataset of 35k Indian Supreme Court rulings is proposed, while a multitude of baselines is applied to assess the initial performance of machine learning models on the task of predicting and explaining the outcome of a ruling. Another category of tasks is related to retrieval task, where a model is asked to retrieve a relevant document or portion of the document given a query. In this category, we can find question-answering models, legal case retrieval, etc. Interestingly, this category of tasks has been the object of multiple evaluation campaigns in the Competition on Legal Information Extraction/Entailment. In the last edition of 2023[11], four tasks have been proposed, regarding legal case retrieval, legal case entailment, statute law retrieval and legal textual entailment. Among the more successful models, THUIR[15] incorporates structural knowledge in the model in a pre-trained language model for legal case retrieval, and it is based on SAILER[14],

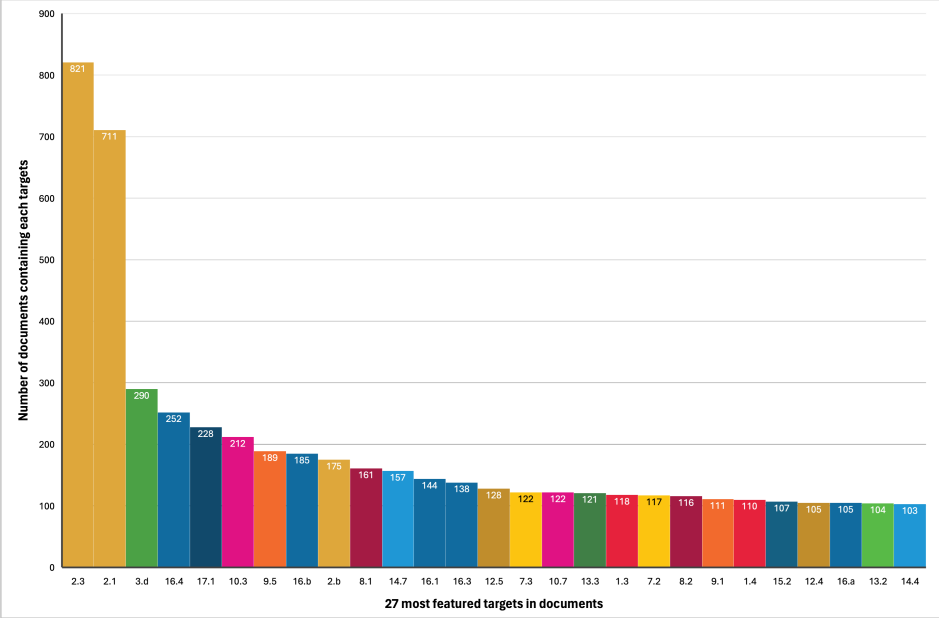

Figure 2: The 27 most featured targets in our document collection DOC1.

a custom encoder-decoder model which is structure-aware that was previously applied to Chinese case law and the previous COLIEE challenges. CAPTAIN[17] uses a multitude of approaches, including a re-training procedure for the Mono-T5 encoder-decoder architecture, which allows it to cast the retrieval questions as a specific input template for the model. Finally, JNLP[2] participated to multiple challenges with different methods, including using the alpaca large language model to perform zero-shot classification and augmenting the training set using Elasticsearch to retrieve relevant paragraphs for legal case retrieval.

In the relevant work, most of the approaches are based on Case Law and data from the judiciary in general, while only some approaches are applied to legislative documents. In this field, however, we can leverage the extensive work on the creation of machine-readable representations for legal documents in general, including legislative ones. In particular, our work leverages the Akoma Ntoso standard[20, 25], which is adopted by multiple international institutions around the world [6, 10, 18, 19, 21]. This format allows machines to manage, among other things, the hierarchical structure of legislative documents, the normative references, the legal definitions, the modifications, and the temporal parameters. For this reason, it is interesting to investigate if the legislative documents in AKN-XML could foster their annotations in an integrated manner with AI and NLP techniques.

3 DOCUMENT COLLECTION

In order to conduct this research we built two different document collections. The first one is considered the gold standard, as it is based on the manual policy mapping carried out by the JRC starting from 2017, performed on documents produced during the Juncker Commission mandate (2015-2019). The dataset consists of the list of the documents with their corresponding SDG. The dataset comes from the JRC portal. Since we were interested in analysing only regulations and directives, we filter the whole database of the mapped policies, obtaining the first dataset, consisting of 2791 documents. According to the manual policy mapping, each document was paired with at least one SDG, and multiple documents were found connected to multiple SDGs. Figure 1 shows how many documents were considered pertaining each SDG by the manual mapping.

The 17 SDGs are displayed on the y-axis, while the number of documents lies on the x-axis. SDG number 2, which aims to reach "zero hunger" is the most featured goal in documents, being mapped in 1526 documents. On the other side, the least featured one is goal number 5 - the "gender equality" goal - which has been mapped in just 20 documents. The second document collection we used in this research was built starting from a European legislation database EUR-LEX (2010-2021), converted in Akoma Ntoso (AKN) by our group. The AKN documents are annotated to contain their structure, references, and also marking legal definitions. We first downloaded the entire database and performed a matching between the 15284 documents contained in it and the documents of the first dataset. The final corpus we worked on was therefore made of the 2791 documents featured in both datasets, in Akoma Ntoso format. While working on documents in Akoma Ntoso, the tree structure of XML can be exploited to reach specific information within texts, considering as a basic legal unit the "article". After the article, we processed information also at legal definition levels of the legal acts and used it to feed the model. Articles are larger texts that in some cases might contain definitions (usually in the first articles of the Law), so many other articles don't include legal definitions. Moreover, in some cases, definitions might be contained by articles, and in some others they might lie outside of them. We chose definitions and articles because we consider these specific parts of texts as crucial for understanding the content and the

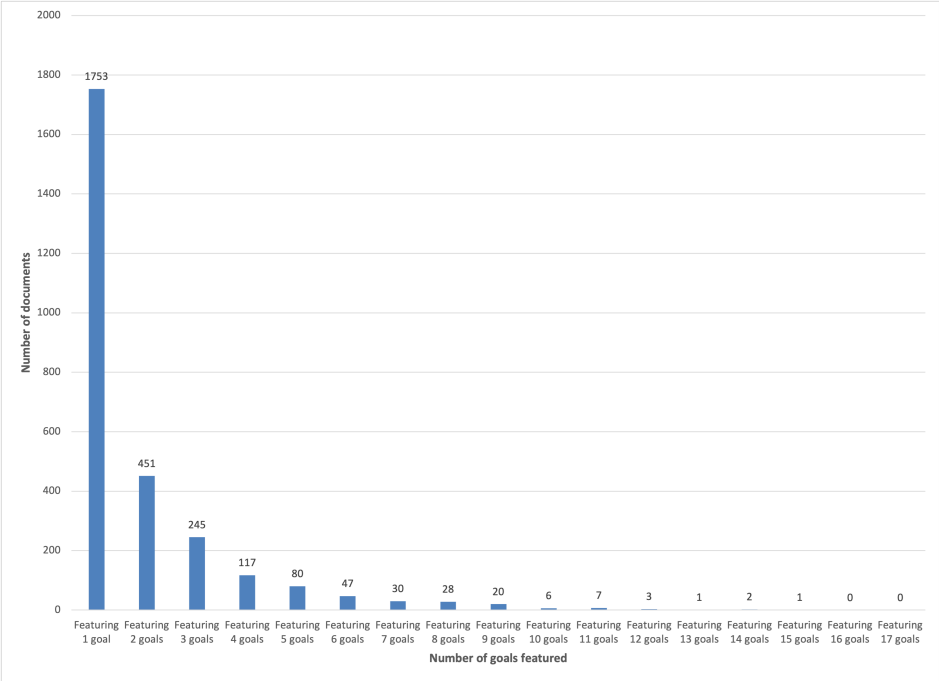

Figure 3: Number of goals featured in our document collection DOC1.

context of each one of the legal documents contained in the corpus. Especially the legal definitions provide the foundational taxonomy of the legal concepts belonging to a Law. It's important to underline that each SDG is further divided into a certain number of targets, which represent specific tasks the United Nations aim to complete in order to fulfil the goal they refer to. There are 169 total targets for 17 SDGs.

Out of 169 total targets included in the 17 SDGs, 35 are not featured in any of the 2791 documents of the dataset we built (DOC1). In addition, 47 targets out of 169 are featured in a number of documents spanning from 1 to 9, and 60 targets out of 169 are featured between 10 and 99 times across all documents. The remaining 27 targets, which are the most featured ones, are displayed in Figure 2. The two most featured targets (2.3 and 2.1) are both related to SDG number 2 featuring, respectively, in 821 and 711 documents. The documents mapped by the JRC can feature one or more goals, according to the topics related to them. Figure 3 shows a count of how many of the 2791 documents contain only one goal and how many contain more of them. More than a half of the documents collection is supposed to feature only one goal (1753 documents), while the remaining 1038 documents contain more goals.

4 METHODS

Since our training data only contains document-level annotations, while our goal is to produce a more fine-grained detection of the SDGs in European legislative documents, our approach can only be unsupervised. For this reason, our method is based on the ideas pioneered by Sentence BERT [22], which uses a siamese network to train a BERT model that can produce semantically informed embeddings, which can be compared using cosine similarity as a metric. In particular, we selected the *all-distilroberta-v1* model from the Sentence Transformers framework, due to the fact that it is the best performing model which supports longer contexts (512 tokens). This model in particular is derived from distillRoBERTa[23] and it is fine-tuned to produce embeddings that are more similar (in terms of cosine similarity) if they represent semantically similar sentences. In particular, the model fine-tuning procedure uses the pre-trained distillRoBERTa model M_d and a training set of matched pairs of sentences a and b, where a_i and b_i are semantically related, and the exact type of relation depends on the dataset used to train the model. Then, in order to obtain the contrastive loss used for *all-distilroberta-v1*, the normalized vector representations for all sentences is obtained by normalizing the outputs of the model M:

$$\bar{A} = \frac{M(a)}{|M(a)|} \quad \bar{B} = \frac{M(b)}{|M(b)|} \tag{1}$$

This results in two matrices $v_a, v_b \in \mathbb{R}^{b_s \times e_s}$, where b_s and e_s represent the batch size and the output size of the model, respectively. In order to obtain the cosine similarity between all embeddings in v_a, v_b, it is possible to only apply a matrix multiplication between the two sentences:

$$S = \bar{A}\bar{B}^T \tag{2}$$

Since $S_{ij} = A_i \cdot B_j$, meaning that each cell corresponds to the dot product between two embeddings, and since these vectors have unit norm from the normalization in equation 1 this corresponds with their cosine similarity. In order to produce the loss, then, two categorical cross entropy losses are applied, in a method pioneered in [5]. In order to do so, a diagonal matrix is produced, to be used

as the labels of the loss:

$$y = \text{diag}(1, 2, \ldots, b_s) \tag{3}$$

Meaning that each vector in the embedding matrices A, B must be matched with the corresponding element from the other half of the batch. Finally, the complete loss can be expressed as:

$$\mathcal{L}(S, y) = -\left(\frac{1}{b_s} \sum_{i=0}^{b_s} \log \frac{\exp S_i}{\sum_{j=0}^{b_s} \exp S_j} y + \frac{1}{b_s} \sum_{i=0}^{b_s} \log \frac{\exp S_i^T}{\sum_{j=0}^{b_s} \exp S_j^T} y \right) \Bigg/ 2 \tag{4}$$

With this training procedure, the model is able to produce semantically aware, general purpose embeddings, which can be used for our task. One of the goals of this experiment is to compute a similarity score between SDG sub-goals and articles, as well as between SDG sub-goals and definitions in the context of European legislative documents. The length of article and definitions, then, is one of the main obstacles to the application of a sentence transformer, which can only encode sequences of up to 512 tokens. In this situation, one might be tempted to split the articles by using a sentence-based approach, where each sentence is used to produce its embedding. Then, the article embedding would be derived by averaging the single sentence embeddings. In the legislative domain, however, we argue that the structure of the document is even more indicative of the boundaries between portions of text that need to be considered in relation with each other. The usage of Akoma Ntoso is then crucial, as it allows us to traverse an XML tree that denotes the structures that are present in many legislative documents (lists, paragraphs, etc). By leveraging the XML tree, we propose two different approaches in order to produce the embeddings: an approach that uses the XML hierarchy to aggregate embeddings in a hierarchical way, and a flat approach, which uses Akoma Ntoso tags to split the document but then proceeds to average the resulting embeddings without considering their structure (see Figure 4).

The two strategies have different merits: using the tree averaging strategy each structural component of the legislative documents has its own embedding assigned, and all components share the same importance when composing the final vector. On the other hand, the flat strategy assigns the same importance to each textual component, meaning that even smaller portions of the text (eg a single point in a list) have a relatively higher importance when producing the final results. The final dataset described in section 3 is the one containing information about the SDGs. In particular, since each SDG is composed of multiple sub-goals, we obtained the list containing a brief description of each sub-goal. Due to their brevity, no further processing was performed and they were used to obtain their respective embeddings. Once we obtained the two different sets of vectors, we produce the cosine similarities between each SDG-article pair belonging to the two datasets. Then, we produce a ranking of the most pertinent SDGs for each document, allowing us to analyse the results.

5 RESULTS

While we do have a form of gold standard in the form of the document-level SDG annotations from the JRC, these annotations are not fine-grained enough to allow a trivial application of standard metrics for classification tasks or information retrieval. For this reason, and since the output of the model is a ranking of the SDGs that the model evaluates as semantically closer to articles and definitions. We cannot *a priori* determine whether an article or definition is related to one or more than one SDGs or whether an article or definitions is not related to any SDG at all. The latter case occurs when they are concerned with legal aspects of the document, such as the entry into force.

In order to address these issues, then, we make some reasonable assumptions about the nature of the relation between articles and SDGs. In particular, we assert that if a document is related to one or more SDGs according to the JRC, some of its articles should be related to said SDGs. Therefore, in terms of the model, some articles should have the matching SDG targets in the top rank obtained from the similarities. For this reason, we devise a framework in order to compare the results to a random baseline, where the rankings obtained from the model are completely random. In this case, we can first express the probability that one of the gold standard SDGs for a given document is in the first rank for an article of that document, which we denote as $P(a_i)$. As we are only interested in the first rank of a randomly generated list, this is equivalent to a uniform distribution:

$$P(a_i) = \frac{|s_i|}{164} \tag{5}$$

Where s_i is the set of SDG targets related to a given document, and 164 is the total number of SDG targets, while the i index represents the i-th article of the document. In order to evaluate the model, then, we use a Poisson Binomial distribution to test our null hypothesis, or the fact that the number of articles that match the "correct" SDG sub-goal is the outcome of a random process. In particular, we use the python package fast-poibin, which implements two methods to calculate the Probability mass function (PFM) of a Poisson Binomial distribution, one based on a discrete Fourier transform[13] and another based on a divide-and-conquer approach[1]. In our experiments we opted to use the more precise, but slower, divide-and-conquer approach and we modified the library to use higher precision floating point numbers.

By using the fast-poibin package, we are then able to produce the p-value for the right tailed tests, meaning the probability that the number of observed successes happened randomly. Given a random variable X that counts the number of successes, the package produces the cumulative distribution function:

$$F_X(x) = P(X \leq x) = \sum_{i=0}^{x} P(X = i) \tag{6}$$

In order to obtain the right-tailed p-values, then, we can apply the following procedure:

$$T_X(x) = P(X \geq x) = 1 - P(X \leq x - 1) \tag{7}$$

The result of this evaluation is shown in Table 1, where in addition to the two document splitting strategies we show a random baseline, including its expected number of matches and the resulting p-value. One peculiar aspect of this evaluation is the fact that the p-values obtained from the two results are exactly identical, despite the fact that the number of correct matches is different between the two splitting strategies. This situation is due to the sum in Equation 6, where the lower-valued probabilities (for high values of i in the equation) are so small that they cause numeric cancellation. For this

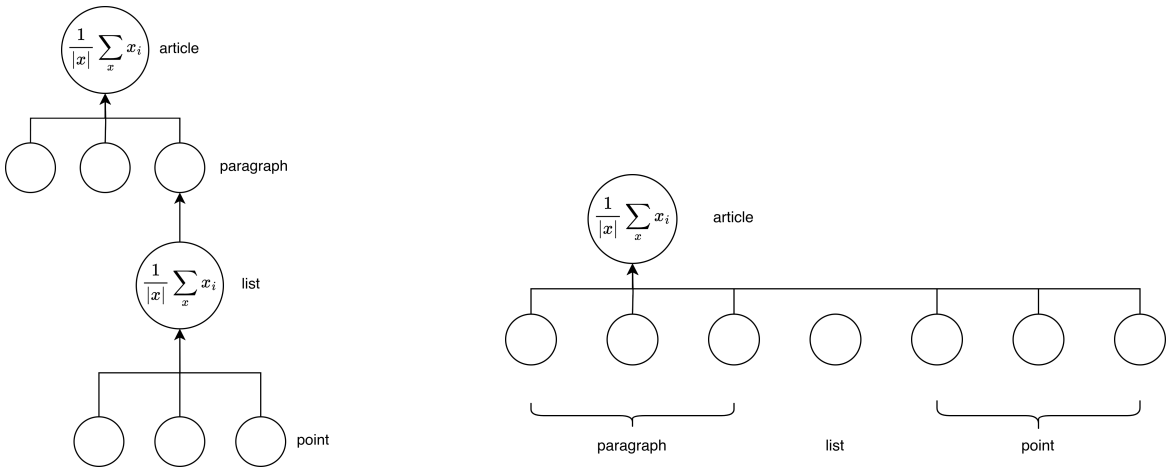

Figure 4: From left to right, the tree and flat strategies to average the embeddings, applied to an article with three paragraphs, one containing a list of points. Each node in the XML tree that directly contains text is associated with its own embedding vector from the model (the list, which contains no text, is ignored in the flat strategy). The vectors are then aggregated using the two strategies, and the sum in some nodes denotes the mean between vectors. The arrows denote the flow of the (aggregated) vectors in the tree.

reason, we express this p-value as an inequality, as the real value is less than the one indicated by the numerical result due to the fact that the sum exibiting the numeric cancellation is subtracted to 1 in Equation 7.

Split Stategy	No. Matches	P-value	Correct matches
Flat	3613	$< 2.71 * 10^{-19}$	0.165
Tree	3738	$< 2.71 * 10^{-19}$	0.171
Random Baseline	826	0.51	0.038

Table 1: Comparison between the two splitting strategies and a random baseline, including p-values for the right tailed tests.

As the Table shows, then, the unsupervised model is better than our random baseline by a very significant amount. Additionally, the tree strategy, which uses the hierarchical structure of the Akoma Ntoso documents seems to show a higher degree of accuracy, which indicates that the hierarchical nature of legislative documents was useful when forming embeddings, even when dealing with very long sequences of tokens.

Another possibility to assess the accuracy of the model is to compare the first-ranking SDG targets that are matched with the overall document annotation with those first-ranking targets that are not. In Figure 5 we show the two histograms for the similarities of the SDGs in the gold standard and those not belonging to the gold standard. By observing the two sets of histograms, it is clear that both splitting strategies result in higher similarities for the gold standards SDGs, and the difference appears to be significant. By comparing the two images, then, it is possible to note that the discrepancy between the two histograms appears to be more pronounced in the tree splitting strategy, which would suggest an advantage of the more structured approach to splitting documents.

While the discrepancies between non-gold standard SDGs and gold standard SDGs are clear in the histogram and they seem significant, it is necessary to test this hypothesis in a more rigorous manner. In order to assess the statistical significance of the discrepancies, then, we use two Welsh t-tests to compare the two distributions, under the null hypothesis that they have the same mean, and the alternative hypothesis that the SDG targets that are in the gold standard have a higher similarity than those not belonging to it. We use the Welsh t-test and not the Student's t-test as it does not assume the same variance for both distributions.

Split Strategy	Gold Standard SDGs	Non-Gold Standard SDGs	P-value
Flat	0.377 ± 0.06	0.353 ± 0.05	$6.84 * 10^{-105}$
Tree	0.381 ± 0.06	0.357 ± 0.05	$5.61 * 10^{-113}$

Table 2: Welsh t-test comparing the distributions of top ranking SDG targets that are and are not in the gold standard, respectively, for both splitting strategies.

The results of the aforementioned Welsh t-tests are shown in Table 2 and they confirm that the gold standard SDGs appearing as the most similar to any article follow a distribution that has a higher mean than the non-gold standard SDGs. However, we can observe that, while there is a difference in the mean similarities for both methods, this difference is relatively small. Finally, the tree splitting strategy tends to produce higher similarity values, while showing a higher degree of separation between the two classes.

As a final evaluation step, we annotated some of the documents in the corpus at the article (N=100) and definition (N=217) level, in order to assess the performance of the model. Since our model cannot determine whether a SDG target is relevant for a given article or definition, instead producing a ranking of the most relevant ones, we opted to compute the top 5, 3 and 1 recalls for both

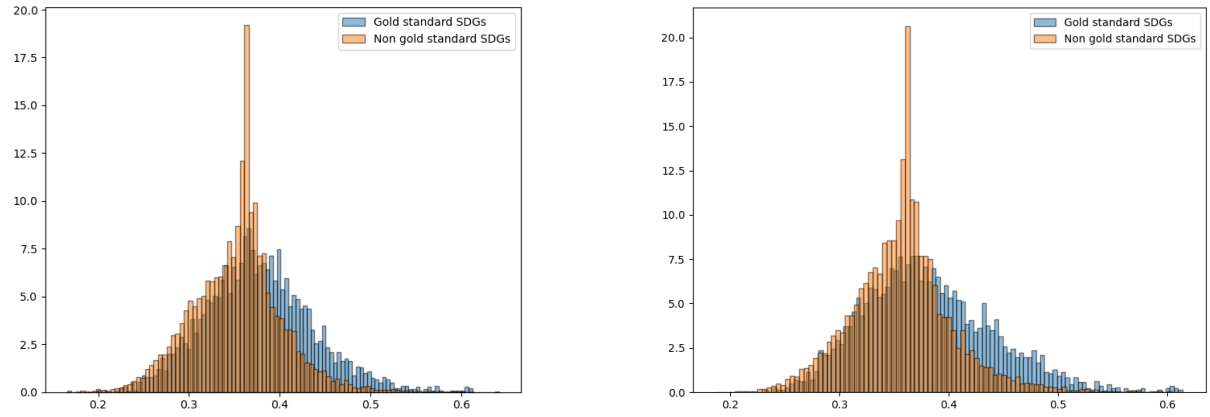

Figure 5: The histograms of the similarities from SDG targets in the gold standard and those that are not for the flat split strategy (on the left) and the tree split strategy (on the right).

Target	Strategy	Top 5 Recall	Top 3 Recalll	Top 1 Recall
Articles	Flat	**0.559**	0.364	0.128
Articles	Tree	0.549	**0.374**	**0.143**
Definitions	Flat	0.919	0.658	0.257
Definitions	Tree	**0.921**	0.658	**0.259**

Table 3: Top 5,3,1 recall for articles and definitions using both splitting strategies.

articles and definitions, using both the flat and xmltree strategies for splitting the document. In this context, the recall used is the one for information retrieval tasks, which measures the number of relevant SDG targets that have been retrieved divided by the total number of relevant SDG targets in the gold standard. The results of this evaluation, shown in Table 3 show a couple of interesting trends:

- The models are very successful when it comes to determining the relevant SDG targets for a given definition, while they can find less targets with respect to articles. Intuitively, this discrepancy could be due to the semantic similarity of definitions to SDG targets, which contain significant keywords that delimit the goal/target. In other terms, since SDGs are highly descriptive, legal definition tends to semantically match them.
- The difference between the two splitting strategies is negligible, with a very slight advantage for xmltree, which could be due to random chance. However, the xmltree strategy appears to produce a clearly better separation between the gold standard and non-gold standard SDGs, thus highlighting the necessity of preserving the hierarchical structure of the legal document.

6 DISCUSSION AND CONCLUSIONS

The results provide answers to the research questions that are the core of this paper:

- RQ1: we are now able to precisely detect the classification of the SDGs article by article. We have more success for the first articles or for the articles at the end that are very regular. In articles where there is only a normative reference, we should navigate the citation for detecting the text of the destination and include it in the model. The mapping works better for the legislative documents that are focused on a topic closely related with SDGs. For the legislative documents focused on the European legal procedure, the mapping is very generic and unspecified (e.g., 1.4, 16.10).
- RQ2: what target of SDGs are associated with the legal definitions? The recall using definitions i s significantly better than the one for the articles. This means that the legal definitions are instruments for understanding the main scope of the legislation. Figure 6 shows that the best results of true positives are found in the first four articles that are usually dedicated to presenting the scope, objectives, and definitions. In particular, usually the article 3 or 4 is devoted to present the definitions.
- RQ3: is the structured text, especially in AKN-XML, better for SDGs classification? The Tables 2, 3, 4 demonstrate that the solution using the XML knowledge produces better results than the plain text.

In conclusions we have used a database of JRC already annotated by human experts with the SDGs classification and an AKN database of legislation for assigning the same classification at the lower granularity (e.g., article vs. document; definition vs. article). We have discovered the following findings:

- structured documents in AKN-XML work better than unstructured ones;

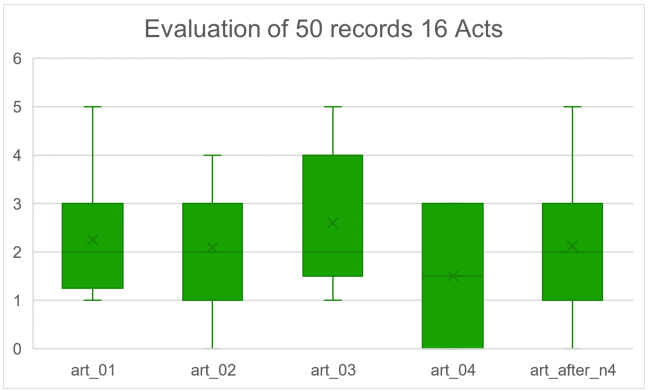

Figure 6: The mean, min and max of the number of true positives from the manual annotation of 16 Acts, for the first four articles and the rest of the document.

- thematic documents linked with the main topics of the SDGs work better than documents that describe legal normative procedures of the European or legislative system (e.g., implementation of legislation, modifications, relationship with Member States, derogations, etc.);

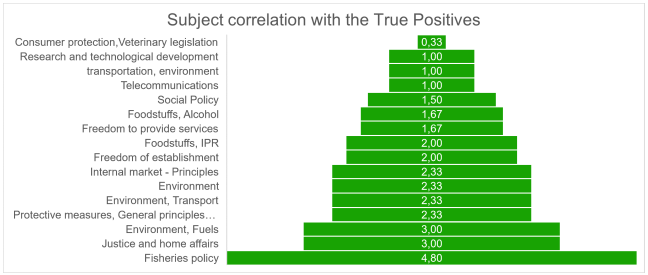

Figure 7: The true positive means per subjects caluclated per articles.

- documents annotated using definitions work better than using articles;
- articles describing the scope, definitions, objectives work best;
- articles that include few sentences and many normative references produce worse results, because of a lack of information.

Considering these output the future work will be focused on the following tasks:

- to use EUROVOC for creating a first signal on the good candidates of the SD goals;
- to use definitions and the first five articles for refining the mapping of the targets;
- to navigate the normative references in order to include text from the cited document;
- to navigate the normative references to the updated version, when it is available, allowing us to update the mapping with the evolution of the normative system;
- to map the other articles (from the fifth) using the target only in instances where there are additional SDGs targets as good candidates.

Our work provides better traceability of the SDGs policies in the EU legislation permitting the legislator to detect the articles where the association is weakest. During the legal drafting, our tool could be integrated into the editor to suggest better legal definitions for improving the implementation of the SDGs.

ACKNOWLEDGMENTS

This project is conducted with the support of the European Commission funds within ERC HyperModeLex. Grant agreement ID: 101055185.

REFERENCES

[1] William Biscarri, Sihai Dave Zhao, and Robert J Brunner. 2018. A simple and fast method for computing the Poisson binomial distribution function. *Computational Statistics & Data Analysis* 122 (2018), 92–100.

[2] Q Bui, Dinh-Truong Do, K Le, Dieu-Hien Nguyen, H Nguyen, T Pham, and L Nguyen. 2023. JNLP COLIEE-2023: Data Argumentation and Large Language Model for Legal Case Retrieval and Entailment. In *Proceedings of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE 2023)*. Braga, Portugal, 17–26.

[3] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[4] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

[5] Yihao Chen, Xianbiao Qi, Jianan Wang, and Lei Zhang. 2023. DisCo-CLIP: A Distributed Contrastive Loss for Memory Efficient CLIP Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Vancouver, BC, Canada, 22648–22657.

[6] Andrija Cvejić, Katarina-Glorija Grujić, Aleksandar Cvejić, Marko Marković, and Stevan Gostojić. 2021. Automatic Transformation of Plain-text Legislation into Machine-readable Format. In *The 11th international conference on information society, technology and management (ICIST 2021)*. Information Society of Serbia - ISOS, Kopaonik, Serbia, 50–55.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Jill Burstein, Christy Doran, and Thamar Solorio (Eds.). Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[8] Qian Dong and Shuzi Niu. 2021. Legal Judgment Prediction via Relational Learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (, Virtual Event, Canada,) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 983–992. https://doi.org/10.1145/3404835.3462931

[9] European Commission and Joint Research Centre, S Borchardt, G Barbero Vignola, D Buscaglia, M Maroni, and L Marelli. 2023. *Mapping EU policies with the 2030 agenda and SDGs – Fostering policy coherence through text-based SDG mapping*. Publications Office of the European Union, Luxembourg. https://doi.org/10.2760/110687

[10] Amelie Flatt, Arne Langner, and Olof Leps. 2023. *Model-Driven Development of Akoma Ntoso Application Profiles: A Conceptual Framework for Model-Based Generation of XML Subschemas*. Springer Nature, Cham, Switzerland.

[11] Randy Goebel, Yoshinobu Kano, Mi-Young Kim, Juliano Rabelo, Ken Satoh, and Masaharu Yoshioka. 2023. Summary of the Competition on Legal Information, Extraction/Entailment (COLIEE) 2023. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law* (<conf-loc>, <city>Braga</city>, <country>Portugal</country>, </conf-loc>) *(ICAIL '23)*. Association for Computing Machinery, New York, NY, USA, 472–480. https://doi.org/10.1145/3594536.3595176

[12] Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems* 35 (2022), 29217–29234.

[13] Yili Hong. 2013. On computing the distribution function for the Poisson binomial distribution. *Computational Statistics & Data Analysis* 59 (2013), 41–51.

[14] Haitao Li, Qingyao Ai, Jia Chen, Qian Dong, Yueyue Wu, Yiqun Liu, Chong Chen, and Qi Tian. 2023. SAILER: Structure-aware Pre-trained Language Model for Legal Case Retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Taipei</city>, <country>Taiwan</country>, </conf-loc>) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 1035–1044. https://doi.org/10.1145/3539618.3591761

[15] Haitao Li, Weihang Su, Changyue Wang, Yueyue Wu, Qingyao Ai, and Yiqun Liu. 2023. THUIR@COLIEE 2023: Incorporating Structural Knowledge into Pre-trained Language Models for Legal Case Retrieval. In *Proceedings of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE 2023)*. Braga, Portugal, 1–6.

[16] Vijit Malik, Rishabh Sanjay, Shubham Kumar Nigam, Kripabandhu Ghosh, Shouvik Kumar Guha, Arnab Bhattacharya, and Ashutosh Modi. 2021. ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 4046–4062. https://doi.org/10.18653/v1/2021.acl-long.313

[17] Chau Nguyen, Phuong Nguyen, Thanh Tran, Dat Nguyen, An Trieu, Tin Pham, Anh Dang, and Le-Minh Nguyen. 2023. Captain at coliee 2023: Efficient methods for legal information retrieval and entailment tasks. In *Proceedings of the Tenth Competition on Legal Information Extraction/Entailment (COLIEE 2023)*. Braga, Portugal, 7–16.

[18] Monica Palmirani. 2018. Akoma Ntoso for Making FAO Resolutions Accessible. In *Knowledge of the Law in the Big Data Age, Conference 'Law via the Internet 2018' (Frontiers in Artificial Intelligence and Applications, Vol. 317)*, Ginevra Peruginelli and Sebastiano Faro (Eds.). IOS Press, Florence, Italy, 159–169. https://doi.org/10.3233/FAIA190018

[19] Monica Palmirani. 2020. Lexdatafication: Italian Legal Knowledge Modelling in Akoma Ntoso. In *AI Approaches to the Complexity of Legal Systems XI-XII - AICOL International Workshops 2018 and 2020: AICOL-XI JURIX 2018, AICOL-XII JURIX 2020, XAILA JURIX 2020, Revised Selected Papers (Lecture Notes in Computer Science, Vol. 13048)*, Víctor Rodríguez-Doncel, Monica Palmirani, Michal Araszkiewicz, Pompeu Casanovas, Ugo Pagallo, and Giovanni Sartor (Eds.). Springer, Cham, Switzerland, 31–47. https://doi.org/10.1007/978-3-030-89811-3_3

[20] Monica Palmirani, Roger Sperberg, Grant Vergottini, and Fabio Vitali. 2018. *Akoma Ntoso Version 1.0 Part 1: XML Vocabulary.* Technical Report. OASIS Standard. http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part1-vocabulary.html

[21] Monica Palmirani, Fabio Vitali, Albano Bernasconi, and Luca Gambazzi. 2014. Swiss Federal Publication Workflow with Akoma Ntoso. In *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference, Jagiellonian University (Frontiers in Artificial Intelligence and Applications, Vol. 271)*, Rinke Hoekstra (Ed.). IOS Press, Krakow, Poland, 179–184. https://doi.org/10.3233/978-1-61499-468-8-179

[22] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[23] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108 [cs.CL] https://arxiv.org/abs/1910.01108

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017), 6000–6010.

[25] Fabio Vitali, Monica Palmirani, Roger Sperberg, and Véronique Parisse. 2018. *Akoma Ntoso Version 1.0. Part 2: Specifications.* Technical Report. OASIS Standard. http://docs.oasis-open.org/legaldocml/akn-core/v1.0/akn-core-v1.0-part2-specs.html

[26] Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. 2021. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open* 2 (2021), 79–84. https://doi.org/10.1016/j.aiopen.2021.06.003

[27] Lucia Zheng, Neel Guha, Brandon R. Anderson, Peter Henderson, and Daniel E. Ho. 2021. When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law* (São Paulo, Brazil) *(ICAIL '21)*. Association for Computing Machinery, New York, NY, USA, 159–168.