Topic Similarities in Rights and Duties across European Constitutions using Transformer-based Language Models

Candida M. Greco[1], Andrea Tagarelli[1]

[1]Dept. Computer Engineering, Modeling, Electronics, and Systems Engineering (DIMES),
University of Calabria, 87036 Rende (CS), Italy

Abstract
The use of language models in the legal NLP field has brought significant advances in the use of AI systems to support legal professionals. However, most of the efforts so far have focused on processing documents such as legal cases, contracts and statutes. There are several types of legal resources that are still overlooked, and these include constitutions. A constitution establishes the basic principles, structures, functions and powers of a country's governance. Several portions of the constitutions are devoted to rights and duties of the citizens, which are essential to define and protect the status of citizens as individuals and as members of the society. To this regard, in this work we focus on the range of topics covered in the European constitutions that guarantee rights and duties to citizens. We present the first study providing lexical and semantic similarity analysis of the European constitutions, which especially takes advantage of using several Transformer-based models.

Keywords
European constitutions, topic similarity, legal language models, artificial intelligence and law

1. Introduction

A *constitution* is a fundamental document that serves as the supreme law of a country or an organization. It establishes the basic principles, rights, and rules that govern the functioning of the entity it applies to. The legal domain is currently one of the major fields of application of AI techniques for supporting experts in the analysis of documents, comprising mostly legal cases, contracts and statutes. Surprisingly, the current literature on the application of AI-based NLP to the processing and understanding of constitutions is quite limited. In [1], the authors carry out a comparative analysis between US constitution and a number of constitutions of four geographic regions (Africa, Asia, Europe, and Middle East), with the aim of detecting differences and similarities w.r.t. rights of citizens and the relationship between the major institutions of their governments; however, the study employs basic techniques used in text mining, discarding any reference to context-free as well as contextualized deep language models. Moreover, no insights into the European countries' constitutions are provided.

In this paper, we aim to fill this gap in the literature by conducting a similarity analysis of the European countries' constitutions with a focus on the *rights and duties of citizens*, which are

LIRAI'23: 1st Legal Information Retrieval meets Artificial Intelligence Workshop co-located with the 34th ACM Hypertext Conference, September 4, 2023, Rome, Italy

candida.greco@dimes.unical.it (C. M. Greco); andrea.tagarelli@unical.it (A. Tagarelli)

essential to define and protect the status of citizens as individuals and as members of the society. Our study is motivated by the opportunity of unveiling commonalities and differences in the constitutions of several European countries as a similarity search problem. In particular, we pursue two main research objectives: (i) understanding how much European countries agree (or differ) w.r.t. specific topics within the realm of citizen rights and duties and (ii) how pre-trained language models are able to discern the differences between topics, especially when they are conceptually related. To achieve this, we employ methods that involve lexical and semantic analysis of the text, with a specific focus on Transformer-based language models.

We believe our work can pave the way for further developments on AI-based solutions for NLP tasks involving the constitutions. This is supported by the evidence that similarity search is broadly employed in the legal AI as an essential means to address more complex tasks, such as statutory article retrieval [2], legal case retrieval [3], document review [4], legal judgement prediction [5], summarization [6], and many others.

In addition, the use of Transformer-based language models for solving legal tasks is a prevailing trend in recent years. For instance, in [7] such models are trained on a topic similarity task to predict the coherence among topics and to detect topical changes on legal texts. In [8], legal and general-purpose Transformer models are compared for legal document recommendation addressed as a similarity task. Nonetheless, our study is the first to analyze the semantics of European constitutions by leveraging Transformer-based language models.

Plan of this paper. The subsequent sections of the paper are organized as follows. In Section 2 we give a brief overview of the Transformer-based models used in this study. Section 3 provides a description of the dataset we built for supporting our study. In Section 4 we outline the specific objectives in our work. In Section 5 we discuss our experimental evaluation and provide an analysis of the achieved outcomes, whereas in Section 6 we summarize the main findings of our analysis, as well as discussing limitations and future perspectives.

2. Background on Transformers

Transformer models [9] have emerged as a dominant paradigm for designing outstanding deep learning models that have revolutionized the state-of-the-art in a wide range of challenging Natural Language Processing (NLP) tasks. The fundamental aspect of the Transformer architecture is the incorporation of attention mechanisms [10], which encompass all hidden states of a neural network at a time and assign suitable weights to capture the inter-dependencies among words. Currently, the Transformer paradigm is widely adopted in NLP, with a significant portion of state-of-the-art NLP models built upon this architecture.

In this work, we focus on a set of Transformer-based language models (TLMs) that is representative according to two key dichotomic aspects in our study: domain-generality vs. domain-specificity, suitability to similarity search tasks vs. task generality. This has led us to select *BERT* as domain-general model, *legal-BERT*s as domain-specific models, and *Sentence-Transformers* as models designed for similarity search tasks. In the following, we recall main characteristics of such models.

BERT. BERT [11] is widely recognized as the pioneering TLM that has revolutionized natural language understanding. The model's advantages encompass bidirectional unsupervised pre-training and a unified architecture that adeptly addresses a range of tasks. The bidirectionality is achieved through the Masked Language Modeling (MLM) task, which involves predicting masked input words from unlabeled text while considering both left and right context words. Moreover, BERT is pre-trained using the Next Sentence Prediction (NSP) task, which aims to determine if one sequence follows another in a given text. Over the years, BERT has played as a catalyst for extensive research, which provided several variants and enhancements of the model. This has culminated in a broad range of BERT-based models.

Legal-BERT. BERT and BERT-based models are primarily designed for general domains. However, their performance tends to degrade when applied to specific domains such as the legal one. To address this limitation, several strategies have been adopted to adapt the Transformer models to the legal domain. The main approaches include further pre-training the model or conducting pre-training from scratch using a legal corpus. Chalkidis et al. [12] were the first to propose both further pre-training and pre-training from scratch BERT on legal corpora, including EU and UK legislations, cases from the European Court of Justice (ECJ), cases from the European Court of Human Rights (ECHR), US court cases and US contracts. In particular, they developed `Legal-BERT-FP` models, obtained through further pre-training of BERT on different sizes of the training legal corpora, and `Legal-BERT-SC` model, the result of training BERT from scratch specifically for the legal domain.

Sentence-Transformers. S-BERT [13] is a variant of BERT that has been specifically tailored for tasks involving semantic textual similarity, clustering, and information retrieval through semantic search. The model employs a fine-tuned siamese network architecture, wherein two separate pre-trained BERT models, each dedicated to one input sentence, share tied weights that are updated during fine-tuning. The siamese architecture enhances the generation of sentence embeddings that encode meaningful semantic information. S-BERT is the core model from which a series of Sentence-Transformers have been developed over the years, representing the state-of-the-art for sentence embeddings.

3. Dataset

Data collection and structure. We retrieved the texts of the European constitutions from the portal www.constituteproject.org [14], which provides free and public access to constitutions from various countries around the world. The resources for the website come from the *Comparative Constitutions Project.*

The documents in the Comparative Constitutions Project were originally labeled according to a number of topics they identify on the constitutions. The topics are organized on three levels, whereby the labels correspond to the most specific topics (i.e., the third-level topics). Hence, given a label, it is possible to identify the topic hierarchy to which it belongs. We use the same labeling system to structure our dataset for the similarity analysis at multiple levels of depth. Specifically, we narrowed down on the first-level topic "rights and duties" and on the European

Table 1
Macro-topics and micro-topics related to Rights and Duties.

Macro-topics	Micro-topics
Physical Integrity Rights	Right to life; prohibition of slavery; prohibition of corporal punishment; ...
Social Rights	Access to higher education; protection of environment; right to health care; right to work; ..
Economic Rights	Protection from expropriation; Right to establish a business; Right to own property; ...
Citizen Duties	Duty to join a political party; duty to pay taxes; duty to serve in the military; duty to work
General Duties	Duty to obey the constitution; binding effect of const rights
Civil and Political Rights	Human dignity; freedom of opinion/thought/conscience; rights of children; right to privacy; ...
Legal Procedural Rights	Principle of no punishment without law; prohibition of double jeopardy; right to counsel; ...
Enforcement	Human rights commission; inalienable rights; ombudsman; ...
Equality, Gender and Minority Rights	Equality regardless of race/gender/religion; protection of stateless persons; right to culture ..

countries. We therefore conducted our similarity analysis of texts associated with third-level topics (hereinafter *micro-topics*, for short) and second-level topics (hereinafter *macro-topics*, for short) referring to rights and duties.

Table 1 shows the macro-topics and an overview of the corresponding micro-topics. Overall, European constitutions encompass 9 macro-topics and 111 micro-topics. Note that the same portion of a constitution can be assigned to multiple micro-topics, but also the same micro-topic can be associated to several, not necessarily contiguous parts of a constitution, in which case the texts are concatenated. In summary, the resulting dataset consists of text portions of constitutions, with each portion being hierarchically assigned the country name, the micro-topic and the macro-topic.

Data cleaning and chunking. Each text corresponding to a particular combination of country and micro-topic may span from one sentence to multiple parts of a constitution. Transformers generally have a maximum limit on the number of tokens they can process. To handle this, we divide the text into chunks so that the input does not exceed the 512 tokens limit imposed by BERT and BERT-based models. The chunking process was carried out so as to keep as many sentences together as possible and ensuring not to break sentences and paragraphs. In any case, just a few instances ended up to exceed 512 tokens and the splitting consisted of mostly 2 or 3 chunks. When possible, a long portion was chunked on the basis of the constitution' structure (e.g., if the portion is the concatenation of parts from different articles, the subdivision was carried out keeping together the sentences from the same article). The chunks were associated with the same country, micro-topic and macro-topic. In general, an instance of the dataset corresponds to one country, but if the text was divided into chunks, there is one instance per chunk. The resulting dataset size is about 2580 instances.

Moreover, a step of *anonymization* was carried out to debias the lexical analysis from specific terms, while preserving the essential meaning in the sentences. In particular, we introduced generic identifiers to replace particular occurrences in the text, such as: names of persons (e.g., royals, secretaries who drew up documents) with "person", inhabitants (e.g., Italians) with "European people", countries (e.g., Italy) with "geo-political European entity", locations (e.g., the Athos peninsula mentioned in the Greek constitution) with "location", organizations (e.g., the United Nations mentioned in the Croatia constitution) with "organization". Analogously, all legal references were replaced with a special token (*law_ref*).

4. Methodology

Lexical vs Semantic similarity across European countries. Firstly, we assess lexical and semantic differences among European countries on the parts of constitutions that share the same micro-topic. This is useful for understanding how much European countries agree (or differ) with respect to a specific theme of interest.

Multi-level semantic similarity analysis. A significant portion of our research efforts has been devoted to an extensive and multi-faceted examination of the similarity between texts extracted from European constitutions. The aim is to assess how well Transformer-based models can capture the closeness of texts that share the same topic, but more importantly how well they can discern differences between texts that deal with different topics (albeit discussing rights and duties) and whether and to what extent they are able to detect subtle nuances of texts from different but similar topics. In particular, we compare our selected Transformer-based models to conduct the following analysis tasks:

- **Topic similarity of texts having the same micro-topic**: the input corresponds to the instances concerning the same topic. The purpose is to compare the language models to assess the ability to detect similarities between countries.
- **Topic similarity of texts having the same macro-topic**: the input consists of all the instances that are associated with micro-topics belonging to the same macro-topic. The purpose is to compare the various language models to assess the ability to discern similar but not identical topics.
- **Topic similarity of texts across all the macro-topics**: the input consists of all portions of the countries dealing with micro-topics from all macro-topics. The purpose is to conduct an overall assessment on the generic topic of rights and duties. Specifically, given a micro-topic as a query, we evaluate the ability of the models to assign higher similarity scores to instances related to the query and, conversely, to assign lower scores with respect to instances related to other micro-topics. Ideally, the lower scores given to instances from the other micro-topics should still reflect whether or not they belong to the same macro-topic of the query, i.e., the scores given to instances from the same macro-topic should be higher than the scores given to instances related to other macro-topics.

5. Experimental evaluation

5.1. Settings

Topic similarity is measured as cosine similarity throughout all the conducted tests. For the lexical analysis, the term-frequency inverse-document-frequency term relevance function (TF-IDF) is adopted to get the document embeddings. In this case, since the sparse vectorial space poses no limit to the input length, the text was not divided into chunks. On the other hand, lemmatization[1] and stemming[2] operations were performed as pre-processing steps.

[1] https://spacy.io/api/lemmatizer

[2] https://www.nltk.org/howto/stem.html

For the semantic analysis, the text embeddings are obtained by each of our models, using the following implementations. `bert-base-uncased` is selected as the domain-general BERT. The selected legal-specific models are developed by [12] and available on Huggingface,[3] namely

- `legal-bert-base-uncased`, a BERT model pre-trained from scratch on legal corpora, which is referred to as `legal-bert-sc` in the original paper;
- `legal-bert-500k`, a BERT model further pre-trained on legal corpora, which is referred to as `legal-bert-fp` in the original paper;
- `bert-base-uncased-echr`, which is `legal-bert-fp` fine-tuned on ECHR cases;
- `bert-base-uncased-eurlex`, which is `legal-bert-fp` fine-tuned on EurLex.[4]

We did not consider other available legal models in [12] since they are specific to U.S. law, while our study is focused on European constitutions. Using all the aforementioned models, we obtain the sentence embeddings applying mean pooling strategy on top of the contextualized token embeddings.

Finally, we consider sentence-Transformer models since they are highly applicable to similarity-related tasks. Currently, several models are available through the `sentence-transformers` library[5] and they are ranked according to the quality of sentence embeddings, based on the performances achieved on different tasks and domains.[6]. Based on the ranking, we chose models that achieve good performance, but at the same time have a manageable size and a maximum length of 512 tokens. We therefore opted for the following models:

- `gtr-t5-large`,[7] based on T5 [15] and fine-tuned for semantic search,
- `all-mpnet-base-v1`,[8] based on MPNet model [16] and fine-tuned on different use-cases,
- `all-distilroberta-v1`,[9] based on a distilled RoBERTa model [17] and fine-tuned on different use-cases.

5.2. Lexical vs Semantic similarity

We first compute and analyze heatmaps of the similarity scores between `TF-IDF` embeddings produced for texts related to the same micro-topic. Each heatmap entry refers to an instance of the dataset dealing with the selected micro-topic and corresponds to a particular country. Note that, in all heatmaps shown throughout this paper, lighter colors correspond to higher similarity scores.

Figure 1 shows representative examples corresponding to selected micro-topics; for the sake of readability, we removed the country labels from the heatmaps. We notice different situations depending on the micro-topic, but in general, there is a low similarity among countries, as it is evident in, e.g., Figure 1a. In some cases, a number of countries show strong lexical similarity,

[3] https://huggingface.co/nlpaueb/legal-bert-base-uncased

[4] https://eur-lex.europa.eu/

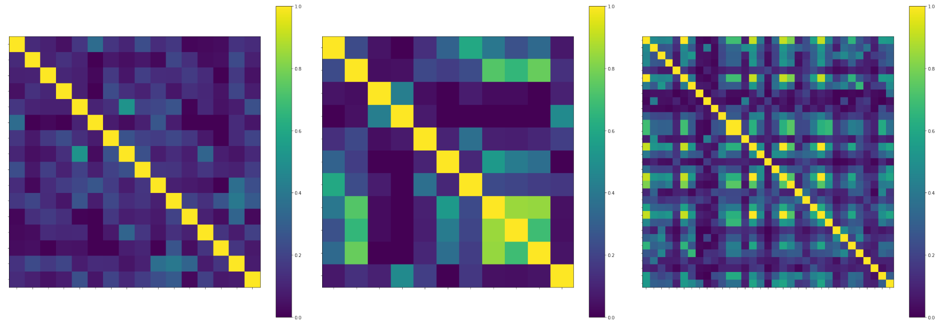
[5] https://www.sbert.net/index.html

[6] https://www.sbert.net/docs/pretrained_models.html

[7] https://huggingface.co/sentence-transformers/gtr-t5-large

[8] https://huggingface.co/sentence-transformers/all-mpnet-base-v1

[9] https://huggingface.co/distilroberta-base



(a) micro-topic “Limits on employment of children"

(b) micro-topic “Right to renounce citizenship"

(c) micro-topic “Prohibition of torture"

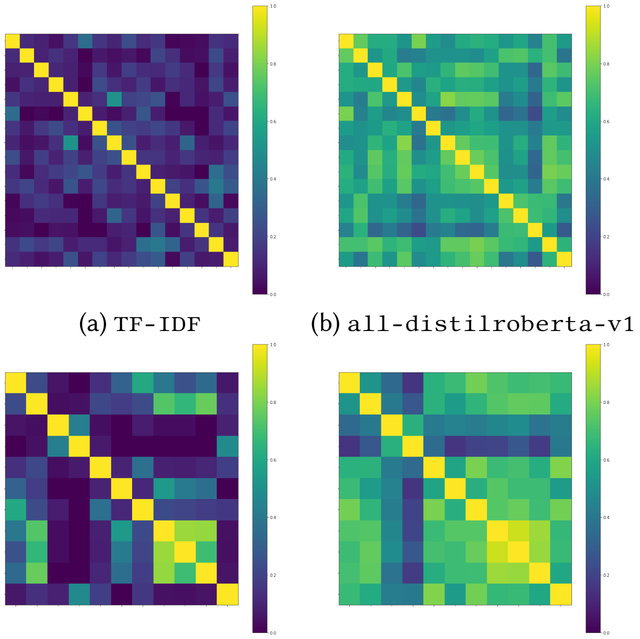
Figure 1: Lexical similarity using TF-IDF

Table 2
TF-IDF most similar examples w.r.t. the micro-topics (a) “*Limits on employment of children*”, (b) “*Right to renounce citizenship*”, and (c) “*Prohibition of torture*”. Given the query text (highlighted in bold), the list of the most similar candidates in descending order of score is shown.

Micro-topic	Similar examples
Limits on employment of children	Italy 1947 (rev. 2020): “**Working women are entitled to equal rights and, for comparable jobs, equal pay as men. Working conditions must allow women to fulfil their essential role in the family and ensure appropriate protection for the mother and child. The law establishes the minimum age for paid labour.** **The Republic protects the work of minors by means of special provisions and guarantees them the right to equal pay for equal work.**” Malta 1964 (rev. 2016): “*Minimum age for paid labour.* *The minimum age for paid labour shall be prescribed by law.*”
Right to renounce citizenship	Russian Federation 1993 (rev. 2014): “**A citizen of the Russian Federation may not be deprived of his (her) citizenship or of the right to change it.**” Serbia 2006: “*A citizen of the Republic of Serbia may not be expelled or deprived of citizenship or the right to change it.*” Slovakia 1992 (rev. 2017): “*No one may be deprived of the citizenship of the Slovak Republic against his will.*”
Prohibition of torture	Albania 1998 (rev. 2016): “**No one may be subjected to torture, cruel, inhuman or degrading punishment or treatment.**” Czech Republic 1993 (rev. 2013): “*No one may be subjected to torture or to cruel, inhuman, or degrading treatment or punishment.*” Kosovo 2008 (rev. 2016): “*No one shall be subject to torture, cruel, inhuman or degrading treatment or punishment.*” Moldova (Republic of) 1994 (rev. 2016): “*No one shall be subject to torture or other cruel, inhuman or degrading punishments or treatments.*” Portugal 1976 (rev. 2005): “ *No one shall be subjected to torture or to cruel, degrading or inhuman treatment or punishment.*” Montenegro 2007 (rev. 2013): “*No one can be subjected to torture or inhuman or degrading treatment.*”

such as in Figure 1b and 1c in which there are peaks of high scores and also peaks of maximum similarity. In Table 2, we report some examples of the most similar texts according to the similarity scores w.r.t. the micro-topics discussed in Figure 1. We can notice that for the micro-topics “*Right to renounce citizenship*” and “*Prohibition of torture*” the texts associated with the highest scores are structurally similar or almost identical.

Figure 2 compares the heatmaps obtained by pairwise similarity of lexical-based embeddings and pairwise similarity of semantic-based embeddings (generated by `all-distilroberta-v1`), corresponding to the micro-topics “*Limits on employment of children*” and “*Right to renounce citizenship*”. It can be noticed that the lexical and the semantic heatmaps have markedly different scores, with the former having significantly lower scores than the latter. In particular, in Figure 2(a-b), we can notice that `all-distilroberta-v1` reveals some similarity matches between countries which are absent in TF-IDF. We show some examples in Table 3. In Figure 2(c-d), the heatmaps have a similar shape but, again, the semantic model associates significantly higher



(a) `TF-IDF` (b) `all-distilroberta-v1`

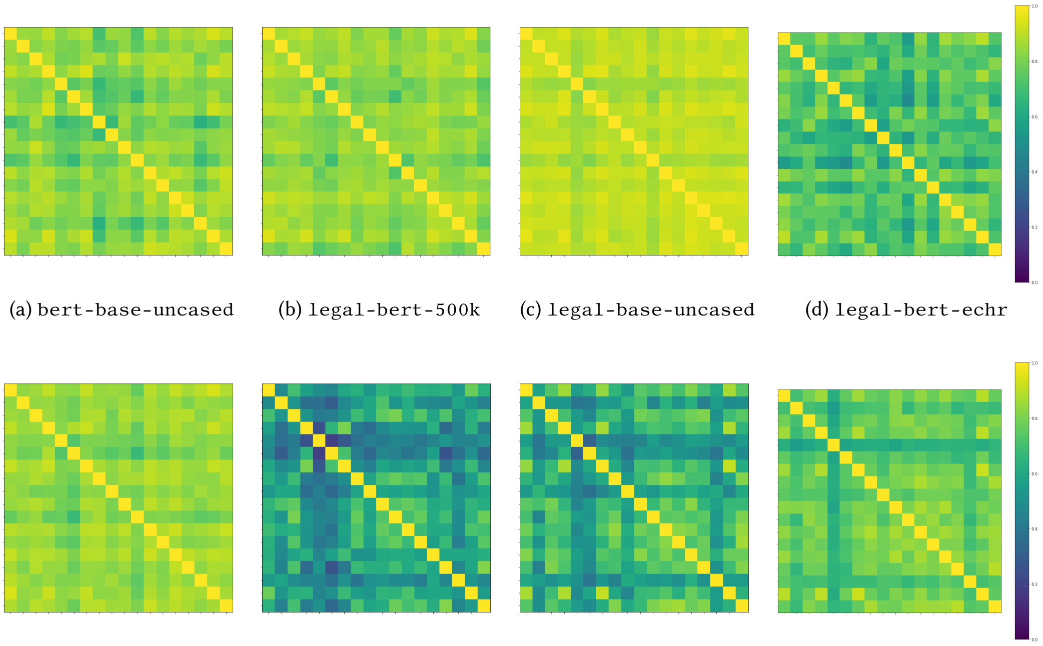
(c) `TF-IDF` (d) `all-distilroberta-v1`

Figure 2: Heatmaps corresponding to (a-b) micro-topic "*Limits on employment of children*" and (c-d) micro-topic "*Right to renounce citizenship*".

Table 3
Example pairs for the micro-topic "*Limits on employment of children*" that are regarded as similar by `all-distilroberta-v1` but dissimilar based on `TF-IDF` embeddings

Moldova (Republic of) 1994 (rev. 2016) "*All employees shall have the right to social protection of labour. The protecting measures shall bear upon the labour safety and hygiene, working conditions for women and young people, the introduction of a minimum wage per economy, week-ends and annual paid leave, as well as difficult working conditions and other specific situations. The exploitation of minors and their involvement in activities, which might be injurious to their health, moral conduct, or endanger their life or proper development shall be forbidden.*" Montenegro 2007 (rev. 2013): "*Youth, women and the disabled shall enjoy special protection at work. A child shall be guaranteed special protection from psychological, physical, economic and any other exploitation or abuse.*"
Albania 1998 (rev. 2016): "*Every child has the right to be protected from violence, ill treatment, exploitation and use for work, especially under the minimum age for work, which could damage their health and morals or endanger their life or normal development.*" Croatia 1991 (rev. 2013): "*Children may not be employed before reaching the legally determined age, nor may they be forced or allowed to do work which is harmful to their health or morality.*"

scores than the lexical model. By comparing the two models, it can be inferred that strong similarities are captured by both, but the semantic model is able to detect more adequately the common focus of the texts. High scores are, indeed, expected since the texts discuss the same micro-topic. By contrast, the lexical model often has very low scores even on texts of the same micro-topic, consequently it also fails to differentiate texts of the same micro-topic from texts of different micro-topics.



(a) `bert-base-uncased` (b) `legal-bert-500k` (c) `legal-base-uncased` (d) `legal-bert-echr`

(e) `legal-bert-eurlex` (f) `all-distilroberta-v1` (g) `all-mpnet-base-v1` (h) `gtr-t5-large`

Figure 3: Micro-topic "*Duty to pay taxes*".

5.3. Results at micro-topic level

As the lexical model revealed to be unable to adequately detect similarities between texts about the same micro-topic, hereinafter we focus on semantic similarity only, in the attempt of identifying the best model in capturing commonalities and differences between micro-topics.

A first step is at the micro-topic level, that is, evaluating which model is able to associate the highest scores for texts belonging to the same micro-topic. Figure 3 shows heatmaps of all the considered Transformer-based models (BERT, Legal-BERT models, and Sentence-Transformers) for the micro-topic "*Duty to pay taxes*". In general, they all provide high scores, but `legal-bert-base-uncased` exhibits extremely high similarities. Among all, `legal-bert-echr` and the Sentence-Transformers provide some slightly lower scores. The models exhibit a similar pattern across all micro-topics; we omit the heatmaps due to space limitation of this paper, nonetheless, in Table 4 we provide an overview of the models' behavior. More precisely, for each macro-topic, we provide the average values of the mean, minimum, maximum, and median similarity scores calculated on the same micro-topics. For instance, for the macro-topic "*Physical Integrity Rights*", `bert-base-uncased` provides, on average, a mean similarity score of 0.834 on texts related to the same micro-topic, while `all-distilroberta-v1` provides an average maximum similarity score of 0.878 for the macro-topic "*Civil and Political Rights*". We observe that legal models have the highest values and the smallest range between the minimum and maximum scores for most macro-topics. This may be due to an over-specialization knowledge of legal domain, leading to high scores for texts having strong

Table 4
Statistics on similarity scores within the same micro-topic, for each macro-topic.

Model	Physical Integrity Rights				Social Rights				Economic Rights				Citizen Duties				General Duties			
	Mean	Min	Max	Median	Mean	Min	Max	Median	Mean	Min	Max	Median	Mean	Min	Max	Median	Mean	Min	Max	Median
`bert-base-uncased`	0.834	0.670	0.976	0.842	0.850	0.720	0.944	0.859	0.836	0.670	0.968	0.845	0.808	0.658	0.926	0.822	0.855	0.736	0.945	0.857
`legal-bert-uncased`	0.876	0.773	0.976	0.884	0.912	0.842	0.967	0.917	0.899	0.798	0.983	0.907	0.899	0.835	0.954	0.903	0.903	0.824	0.958	0.911
`legal-bert-500k`	0.820	0.644	0.965	0.828	0.865	0.751	0.951	0.872	0.841	0.677	0.971	0.853	0.831	0.708	0.926	0.842	0.851	0.755	0.933	0.858
`legal-bert-eurlex`	0.829	0.668	0.969	0.833	0.854	0.745	0.942	0.859	0.836	0.693	0.965	0.844	0.825	0.713	0.926	0.829	0.845	0.762	0.916	0.850
`legal-bert-echr`	0.747	0.538	0.946	0.750	0.769	0.601	0.906	0.779	0.744	0.516	0.933	0.751	0.704	0.496	0.874	0.717	0.747	0.594	0.892	0.755
`all-distilroberta-v1`	0.612	0.327	0.899	0.606	0.569	0.283	0.820	0.568	0.564	0.226	0.883	0.570	0.486	0.188	0.803	0.480	0.569	0.327	0.815	0.566
`all-mpnet-base-v1`	0.629	0.336	0.923	0.625	0.612	0.337	0.860	0.614	0.590	0.239	0.917	0.592	0.542	0.279	0.740	0.537	0.590	0.407	0.749	0.581
`gtr-t5-large`	0.787	0.654	0.933	0.781	0.764	0.625	0.903	0.763	0.760	0.583	0.939	0.757	0.740	0.615	0.852	0.743	0.751	0.637	0.874	0.746

Model	Civil and Political Rights				Legal Procedural Rights				Enforcement				Equality, Gender and Minority Rights			
	Mean	Min	Max	Median	Mean	Min	Max	Median	Mean	Min	Max	Median	Mean	Min	Max	Median
`bert-base-uncased`	0.853	0.724	0.954	0.861	0.857	0.739	0.956	0.866	0.856	0.685	0.979	0.865	0.873	0.755	0.964	0.880
`legal-bert-uncased`	0.906	0.835	0.970	0.915	0.891	0.802	0.969	0.898	0.908	0.803	0.985	0.916	0.919	0.844	0.974	0.924
`legal-bert-500k`	0.856	0.738	0.952	0.866	0.834	0.695	0.948	0.845	0.864	0.732	0.978	0.871	0.874	0.764	0.960	0.879
`legal-bert-eurlex`	0.853	0.743	0.949	0.860	0.850	0.736	0.949	0.857	0.849	0.704	0.971	0.856	0.875	0.766	0.961	0.879
`legal-bert-echr`	0.763	0.582	0.914	0.771	0.744	0.552	0.914	0.751	0.785	0.581	0.943	0.797	0.808	0.654	0.935	0.815
`all-distilroberta-v1`	0.609	0.331	0.878	0.614	0.556	0.277	0.837	0.554	0.569	0.238	0.869	0.584	0.595	0.324	0.860	0.595
`all-mpnet-base-v1`	0.633	0.361	0.893	0.637	0.594	0.311	0.866	0.602	0.565	0.269	0.854	0.566	0.602	0.329	0.877	0.594
`gtr-t5-large`	0.781	0.652	0.914	0.779	0.758	0.619	0.899	0.759	0.756	0.573	0.926	0.755	0.784	0.659	0.918	0.780

legal concepts in common, or conversely to a poor ability to recognize differences.

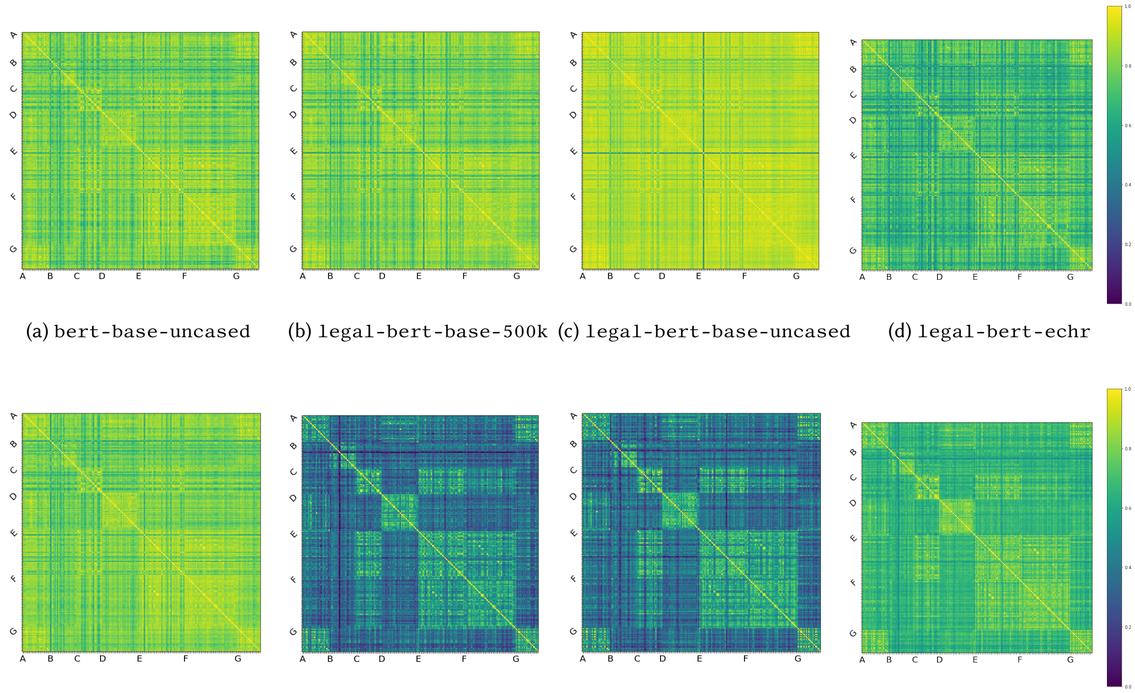
5.4. Results at macro-topic level

Considering the aforementioned behavior of the models on each macro-topic, we explore whether the high scores of legal models can be ascribed either to a robust understanding of the micro-topic or to a potential inability to discern subtle semantic distinctions.

For this purpose, we assess whether different micro-topics belonging to the same macro-topic are indistinguishable, that is, whether the models can distinguish related but not identical topics. In Figure 4, we show the heatmaps corresponding to the various micro-topics belonging to the macro-topic "*Economic Rights*". Note that each heatmap reports both the similarities between texts belonging to the same micro-topic (which are concentrated on the main diagonal) and the similarities between texts belonging to different micro-topics. For the sake of readability, we replace the names of the countries with letters corresponding to their respective micro-topics; when an explicit label is missing, it is inferred that the entry of the heatmap is associated to the preceding label.

It can be noticed that legal models (and even `bert-base-uncased`) are unable to perceive different degrees of similarity, which should be higher for texts of the same micro-topic and lower between texts of different micro-topics. On the contrary, the Sentence-Transformers (particularly `all-distilroberta-v1` and `all-mpnet-base-v1`) are able to distinguish the different micro-topics much more clearly. However, there are micro-topics that are nearly indistinguishable even for the Sentence-Transformers. Examining these challenging micro-topics, we observe that they often involve remarkably similar concepts. For example, in Figure 4g, there is a clear difficulty in distinguishing the micro-topics C, E, and F, which, however, have in common the aspect of addressing matters pertaining to property rights. Similarly, the micro-topic A and G are practically indiscernible even for the Sentence-Transformers.

The above is also evident in the boxplots in Figure 5. In this case as well, the micro-topics A and G encompass a closely related concept, namely aspects related to business. Once again, the behavior of the models is consistent across all macro-topics, with `all-distilroberta-v1` and `all-mpnet-base-v1` showing the best results. To provide an example, Figure 6 shows the



(a) `bert-base-uncased` (b) `legal-bert-base-500k` (c) `legal-bert-base-uncased` (d) `legal-bert-echr`

(e) `legal-bert-eurlex` (f) `all-distilroberta-v1` (g) `all-mpnet-base-v1` (h) `gtr-t5-large`

Figure 4: Macro-topic "*Economic Rights*": A - *Right to establish a business*; B - *Provisions for intellectual property*; C - *Right to transfer property*; D - *Right to choose occupation*; E - *Right to own property*; F - *Protection from expropriation*; G - *Right to competitive marketplace*

boxplots of `all-mpnet-base-v1` and `legal-bert-uncased` for the micro-topic "*Prohibition of slavery*" against all the micro-topics of its macro-topic ("*Physical Integrity Rights*"). It can be observed that, in the case of `all-mpnet-base-v1`, the boxplot corresponding to the texts of the micro-topic under examination (the first one from the left) has a higher mean compared to the other boxplots, which represent the similarity scores of texts from the "*Prohibition of slavery*" topic compared to texts from other micro-topics of its macro-topic. On the contrary, the boxplots of `legal-bert-uncased` demonstrate that the model does not perceive substantial differences between texts with different micro-topics compared to texts with the same micro-topic.

5.5. Overall results on Rights and Duties

Another crucial aspect concerns the models' capability to distinguish between micro-topics that belong to the same macro-topic, as opposed to micro-topics from different macro-topics. Given a micro-topic as a query, we shall compute similarity scores for all micro-topics. This can be seen as an overall evaluation of the Rights and Duties topic.

Our analysis can be grouped into three categories: (1) similarity between texts of the same micro-topic as the query, (2) similarity between texts of different micro-topics but within the same macro-topic as the one associated to the query micro-topic, and (3) similarity between

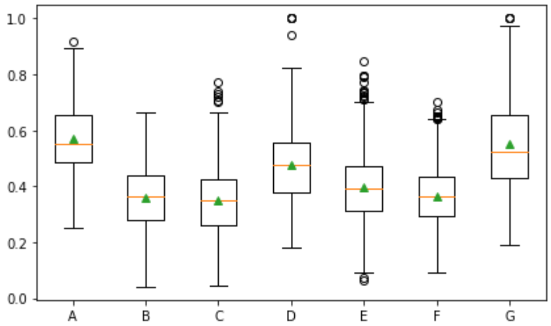
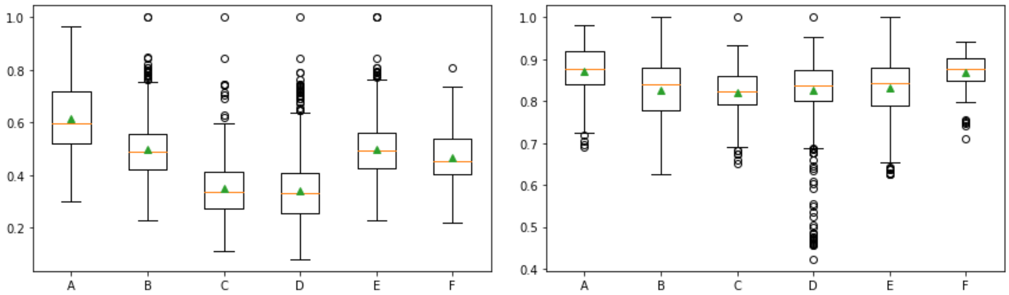


Figure 5: Boxplots on the micro-topic "*Right to establish a business*" vs the other micro-topics of the same macro-topic ("*Economic Rights*") using `all-mpnet-base-v1`. A - *Right to establish a business*; B - *Provisions for intellectual property*; C - *Right to transfer property*; D - *Right to choose occupation*; E - *Right to own property*; F - *Protection from expropriation*; G - *Right to competitive marketplace.*



(a) `all-mpnet-base-v1`

(b) `legal-bert-uncased`

Figure 6: Boxplots on the micro-topic "*Prohibition of slavery*" vs the other micro-topics of the same macro-topic "*Physical Integrity Rights*". A - *Prohibition of slavery*; B - *Prohibition of cruel treatment*; C - *Prohibition of capital punishment*; D - *Right to life*; E - *Prohibition of torture*; F - *Prohibition of corporal punishment.*

texts with both a different micro-topic and macro-topic w.r.t. the query. The expected behavior for the models is to assign high similarity scores for the first category and low scores for the other two categories, but the scores associated with the second category should be higher compared to the scores associated with the third category.

Table 5 summarizes the analysis conducted on all micro-topics, providing an overall mean value across the three categories. For instance, `bert-base-uncased` obtains, on average, a mean similarity score of 0.853 on texts related to the same micro-topic, a mean similarity score of 0.799 on texts related to different micro-topics but having the same macro-topic, and a mean similarity score of 0.762 on texts related to different micro-topics and different macro-topics. The difference between the values of the first and second categories (column $|\mu - M|$), is of 0.053, for the second and the third categories (column $|M - M'|$) is of 0.037, and for the first

Table 5
Overall evaluation of the models. The *micro-topic* column represents the average (μ) over the mean values of similarity scores between texts belonging to the same micro-topic. The *macro-topic* column represents the average (M) over the mean values of similarity scores between texts belonging to the same macro-topic but not the same micro-topic. The *other macro-topics* column represents the average (M') over the mean values of similarity scores between texts belonging to different macro-topics and different micro-topics. The remaining columns correspond to the difference between the values of the columns micro-topic and macro-topic ($|\mu - M|$), the difference between the values of the columns macro-topic and other macro- topics ($|M - M'|$), and the difference between the values of the columns micro-topic and other macro-topics ($|\mu - M'|$).

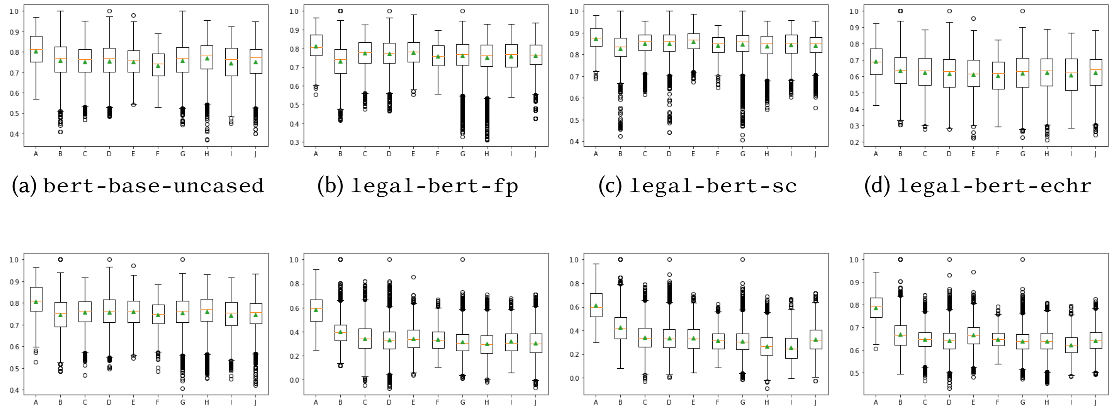
model	**micro-topic** (μ)	**macro-topic** (M)	**other macro-topics** (M')	$\lvert\mu - M\rvert$	$\lvert M - M'\rvert$	$\lvert\mu - M'\rvert$
`bert-base-uncased`	0.853	0.799	0.762	0.053	0.037	0.091
`legal-bert-uncased`	0.905	0.871	0.852	0.033	0.019	0.052
`legal-bert-500k`	0.853	0.798	0.766	0.055	0.031	0.087
`legal-bert-eurlex`	0.853	0.796	0.756	0.057	0.039	0.096
`legal-bert-echr`	0.765	0.687	0.643	0.078	0.044	0.122
`all-distilroberta-v1`	0.576	0.399	0.313	0.177	0.086	0.263
`all-mpnet-base-v1`	0.603	0.414	0.323	***0.189***	***0.090***	***0.279***
`gtr-t5-large`	0.769	0.682	0.643	0.086	0.039	0.126

and the third categories (column $|\mu - M'|$) is of 0.091. Once again, it can be observed that the generic `bert-base-uncased` and the legal models are the least effective in distinguishing between different micro-topics. Even `gtr-t5-large` demonstrates limitations in this regard, particularly in distinguishing between the second and third category. Among all the models, `all-mpnet-v1` achieves the most favorable results, demonstrating the largest disparity in all scenarios (columns $|\mu - M|$, $|M - M'|$ and $|\mu - M'|$), followed by `all-distil-roberta`.

As an illustrative example, we show the boxplots of all models for the micro-topic "*Prohibition of slavery*" in Figure 7. The differences across the boxplots are more evident with `all-mpnet-base-v1`. Indeed, the first boxplot (related to the first category) exhibits the highest and most uniform values, the second boxplot (related to the second category) is sufficiently lower than the first one but higher than the others (related to the third category).

6. Discussion

The experimental results unveil a number of major findings. In general, we have found that European constitutions share many topics in the context of Rights and Duties. Despite for most micro-topics the lexical similarities are generally very low, it also happens that few pairs of European countries apparently address the same micro-topic in a similar manner; however, lexical analysis is not sufficient to capture the similarities between countries at a fine grain. On the other hand, the semantic analysis unveils quite different behavior of the language models. In particular, it is evident that the legal models we consider are not directly applicable to similarity tasks, whereas the Sentence-Transformers demonstrate significantly better results, despite not being specifically trained on legal corpora. This is not actually surprising, since the constitutions are not written in a highly technical legal language, as they should be easily comprehensible even for non-experts. Secondly, the Sentence-Transformers are trained to generate sentence



(a) `bert-base-uncased` (b) `legal-bert-fp` (c) `legal-bert-sc` (d) `legal-bert-echr`

(e) `legal-bert-eurlex` (f) `all-distilroberta-v1` (g) `all-mpnet-base-v1` (h) `gtr-t5-large`

Figure 7: Boxplots on the micro-topic "*Prohibition of slavery*" vs the parts of constitutions referring to its macro-topic ("*Physical Integrity Rights*") and all the others macro-topics. A - *Prohibition of slavery*; B - *Physical Integrity Rights*; C - *Social Rights*; D - *Economic Rights*; E - *Citizen Duties*; F - *General Duties*; G - *Civil and Political Rights*; H - *Legal Procedural Rights*; I - *Enforcement*; J - *Minority Rights*.

embeddings, which capture the semantic meaning and context of the texts rather than relying on specific legal terminology. As a result, the Sentence-Transformers can effectively capture the similarities and nuances of the constitutions. Among them, `all-mpnet-base-v1` has proven to be the best performer, although `all-distilroberta-v1` follows closely behind; by contrast, `gtr-t5-large` performs significantly worse, likely due to its specialization in semantic search and lack of fine-tuning for other use cases.

This study has some limitations that leave room for future improvement. We notice that when the topics are highly similar, all the models faced difficulties in perceiving their differences. A fine-tuning phase on the constitution data would help recognize subtle nuances in meaning. Also, we are aware that there are many other legal models that could have been considered in the experimentation, and their inclusion could have provided further insights into real performances of legal architectures on similarity tasks. However, we opted to focus on a selected set of models that are widely recognized and representative of the current state-of-the-art in the field. Future research may include exploring a broader range of legal models to gain a comprehensive overview of their capabilities and limitations in detecting topic similarities among constitutions.

7. Conclusion

In this work, we presented a lexical and a semantic similarity analysis among the segments of European constitutions. We investigated the ability of legal and Sentence-Transformer models to recognize texts that share the same topic and to differentiate texts that cover different topics. We conducted a multi-faceted experimental evaluation and provided an analysis of the achieved outcomes, highlighting main findings, limitations and further perspectives.

References

[1] T. Bayrak, A comparative analysis of the world's constitutions: a text mining approach, Soc. Netw. Anal. Min. 12 (2022) 26.

[2] A. Louis, G. van Dijck, G. Spanakis, Finding the Law: Enhancing Statutory Article Retrieval via Graph Neural Networks, in: Proc. of the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023), Association for Computational Linguistics, 2023, pp. 2753–2768.

[3] Y. Shao, J. Mao, Y. Liu, W. Ma, K. Satoh, M. Zhang, S. Ma, BERT-PLI: modeling paragraph-level interactions for legal case retrieval, in: Proc. of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI 2020), 2020, pp. 3501–3507.

[4] S. Shaghaghian, L. Y. Feng, B. Jafarpour, N. Pogrebnyakov, Customizing contextualized language models for legal document reviews, in: Proc. of the IEEE International Conference on Big Data (IEEE BigData 2020), IEEE, 2020, pp. 2139–2148.

[5] N. Aletras, D. Tsarapatsanis, D. Preotiuc-Pietro, V. Lampos, Predicting judicial decisions of the European Court of Human Rights: a Natural Language Processing perspective, PeerJ Comput. Sci. 2 (2016) e93.

[6] A. Shukla, P. Bhattacharya, S. Poddar, R. Mukherjee, K. Ghosh, P. Goyal, S. Ghosh, Legal Case Document Summarization: Extractive and Abstractive Methods and their Evaluation, in: Proc. of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (AACL/IJCNLP 2022), Association for Computational Linguistics, 2022, pp. 1048–1064.

[7] D. Aumiller, S. Almasian, S. Lackner, M. Gertz, Structural text segmentation of legal documents, in: Proc. of the Eighteenth International Conference for Artificial Intelligence and Law (ICAIL 2021), ACM, 2021, pp. 2–11.

[8] M. Ostendorff, E. Ash, T. Ruas, B. Gipp, J. M. Schneider, G. Rehm, Evaluating document representations for content-based legal literature recommendations, in: Proc. of the Eighteenth International Conference for Artificial Intelligence and Law (ICAIL 2021), ACM, 2021, pp. 109–118.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS 2017), 2017, pp. 5998–6008.

[10] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: Proc. of the 3rd International Conference on Learning Representations (ICLR 2015), 2015.

[11] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (NAACL-HLT 2019), Association for Computational Linguistics, 2019, pp. 4171–4186.

[12] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, I. Androutsopoulos, LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, 2020, pp. 2898––2904.

doi:10.18653/v1/2020.findings-emnlp.261.

[13] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proc. of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019), Association for Computational Linguistics, 2019.

[14] Z. Elkins, T. Ginsburg, J. Melton, R. Shaffer, J. F. Sequeda, D. P. Miranker, Constitute: The world's constitutions to read, search, and compare, J. Web Semant. 27-28 (2014) 10–18.

[15] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, J. Mach. Learn. Res. 21 (2020) 140:1–140:67.

[16] K. Song, X. Tan, T. Qin, J. Lu, T. Liu, MPNet: Masked and Permuted Pre-training for Language Understanding, in: Proc. of the Annual Conference on Neural Information Processing Systems (NeurIPS 2020), 2020.

[17] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, CoRR abs/1907.11692 (2019).