Topic Modelling of Legal Documents via LEGAL-BERT

Raquel Silveira [1], Carlos G. O. Fernandes [2], João A. Monteiro Neto [3], Vasco Furtado [4], and José Ernesto Pimentel Filho [5].

[1] *Federal Institute of Education, Science and Technology of Ceará, Tianguá, Ceará, Brazil*
[2] *University of Fortaleza and Banco do Nordeste do Brasil S.A, Fortaleza, Ceará, Brazil*
[3] *University of Fortaleza Law School, and FUNCAP, Fortaleza, Ceará, Brazil*
[4] *University of Fortaleza, Fortaleza, Ceará, Brazil*
[5] *Federal University of Paraíba, João Pessoa, Paraíba, and FUNCAP, Fortaleza, Ceará, Brazil*

Abstract

Legal text processing is a challenging task for modeling approaches due to the peculiarities inherent to its features, such as long texts and their technical vocabulary. Topic modeling consists of discovering a semantic structure in the text. This way, it requires specific approaches. The relevant topics strongly depend on the context in which the legal documents will be presented. This work aims to describe and evaluate the use of BERTopic for topic modeling in legal documents. The authors have focused on a subset of landmark cases from the US Caselaw dataset to evaluate the impact of topic modeling, via domain-specific embeddings pre-trained from LEGAL-BERT. The research investigated different variations of generating sentence embeddings from the cases. Results here presented demonstrate that considering the references to statutory law (e.g. US Code) during the process of text embeddings improves the quality of topic modeling.

Keywords

Natural Language Processing (NLP); American Case Law; Contextualized Embeddings

1. Introduction

Topic Modeling has been successfully applied to Natural Language Processing (NLP) and it is frequently used when a huge textual collection cannot be reasonably read and classified by one person. Given a set of text documents, a topic model is applied to find out interpretable semantic concepts, or topics, present in documents. Topics represent the theme, or subject, of the text and can be used for the elaboration of high-level abstracts considering a massive collection of documents, research documents of interest, and also for grouping similar documents. [1].

The increasing volume of publicly available legal information has required a continuous effort in the field of automatic processing, intending to promote access to relevant information to the public. The authors have in mind the necessity of students, legal scholars, lawyers, judges, and court officials daily. In the case of long documents, succeeding in having a useful abstract with key information about its content and context is an important path to deliver legal services which will create an appropriate environment for improving the productivity in courts involving all due agents of such a process. Therefore, topic modeling may contribute to making efficient the analysis of legal documents since it reveals implied meanings, on the one hand. On the other hand, it performs the discovery of theme relations among different legal documents [2, 3].

Legal documents are often full of technical terminology. Students of law are commonly invited to make notes with particular views of the document because a series of opinions might be triggered as

RELATED - Relations in the Legal Domain Workshop, in conjunction with ICAIL 2021, June 25, 2021, São Paulo, Brazil
EMAIL: raquel_silveira@ifce.edu.br (R. Silveira); carlosgustavo@edu.unifor.br (C. G. O. Fernandes); joaoneto@unifor.br (J. A. Monteiro Neto); vasco@unifor.br (V. Furtado); jepf@academico.ufpb.br (J. E. Pimentel Filho)
ORCID: 0000-0001-7445-605X (R. Silveira); 0000-0003-0575-4509 (C. G. O. Fernandes); 0000-0002-0690-2449 (J. A. Monteiro Neto); 0000-0001-8721-4308 (V. Furtado); 0000-0002-7534-9405 (J. E. Pimentel Filho)

entries for understanding the case. A case is subject to interpretations that are not simple tasks. Therefore, it does give obstacles to create divisions and themes for the document. These features contribute to making topic modeling both challenging and resources consuming. [2] Contextualized text representations have been used to capture semantics. More recently, with BERT [5], these tasks have revolutionized natural language processing for many structured prediction problems [4]. As it happens in other specialized domains, the legal text (for instance, statutory law, lawsuits, contracts) has distinct characteristics in comparison to generic corpora, such as specialized vocabulary, particularly formal syntax, semantics based on extensive knowledge specific domain, you name it. [6] Thus, content representation of text from juridical documents is better processed when applied to a domain-specific model. [7]

In this article we investigate stochastic topic modeling approaches for legal documents, using BERTopic [8], a topic modeling technique that represents text from embeddings of BERT. Specifically, the proposed approach represents legal documents in content form using the pre-trained model LEGAL-BERT [7], a model pre-trained from legal data, intended to assist legal NLP research, computational law, and legal technology applications. The pre-trained contextual embeddings for the specific domain provide a more refined and semantic richer representation of the text. We then evaluate how much the quality of topic modeling of the document is influenced by citations to laws in the body of the text. We do this by extending the semantic representation of the document, with the insertion of text describing the United States Code cited in the document. The results of the different variations of this method have shown that adding the references to laws in embedding representation of the text improves the quality of topic modeling.

2. Related Work

Efforts have been made to apply Natural Language Processing and Machine Learning techniques to legal text. Latent Dirichlet Allocation (LDA) [9] has been used to model legal corpora [3, 10, 11]. The proposed approach by [10] uses LDA to model Extraordinary Resources received by the Supreme Court of Brazil. The data consist of a corpus of lawsuits annotated manually by the Court's specialists with thematic labels. The semantic analysis of topics shows that models with 10 and 30 topics were able to capture some of the legal matters discussed by the Court. In addition, experiments show that the model with 300 topics was the best text vectorizer and that the interpretable, low dimensional representations it generates achieve good classification results.

[11] qualitatively evaluates the performance of topic models to summarize and visualize British legislation, intending to facilitate the navigation and identification of relevant legal topics and their respective set of topic-specific terms. More specifically, Saffron models are evaluated (a software tool that can construct a model-free topic hierarchy), Non-Negative Matrix Factorization (NMF) [12], Latent Semantic Analysis (LSA) [13], LDA [9], and Hierarchical Dirichlet process (HDP) [14]. After evaluation Saffron has been consistently ranked as the most favorable of all models, as the aforementioned vocabulary pruning and usage of multi-word expressions has played a fundamental role in topic coherency.

To explore the possibilities of finding topics in case law documents, [3] evaluates the use of the LDA in extracting precise and useful topics and whether legal experts and people without legal training agree or not in their judgments about it. Experts evaluated Dutch case law documents, identifying that, for most documents, the model was unable to locate the main topic related to the subject of the document.

Until now, the LDA is still the preferred model for modeling topics. Despite its popularity, LDA has several weaknesses. To achieve optimal results they often require the number of topics to be known, custom stop-word lists, stemming, and lemmatization. Additionally, this method relies on the bag-of-words representation of documents which ignore the ordering and semantics of words. Distributed representations of documents and words are gaining popularity due to their ability to capture the semantics of words and documents [1]. Pre-trained language models based on [5] and its variants, have achieved state-of-the-art results in several downstream NLP tasks. This model is able to represent the text in a complex multidimensional space that has the property of capturing the characteristics of the language necessary for its comprehension. [7] release LEGAL-BERT, a family of BERT models for

the legal domain, pre-trained with EU and UK legislations, European Court of Justice cases, European Court of Human Rights cases, US court cases and US contracts.

[4] use generalized contextualized language models (BERT [5], GPT-2 [15], and RoBERTa [16]) for token-level contextualized word representations. These contextualized representations are used by the k-means algorithm to produce topics of the document in English from Wikipedia articles, Supreme Court of the United States legal opinions, and Amazon product reviews. These cluster models are simple, reliable, and can perform as well, if not better than the LDA topic models while maintaining the high quality of the topics.

[1] developed Top2Vec, a model that uses document and word semantic embedding to find topic vectors. Some of the characteristics of this model are: it does not require stop-word lists, stemming, or lemmatization, and it automatically identifies the number of topics. The resulting topic vectors are jointly embedded with the document and word vectors, the distance of which represents the semantic similarity between them.

BERTopic is a topic modeling technique that leverages models based on transformers to achieve robust text representation, HDBSCAN to create dense and relevant clusters, and class-based TF-IDF (c-TF-IDF) to allow easy interpretable topics, while keeps important words in topic descriptions [8].

The relevance of topics modeled in legal documents depends heavily on the legal context and the broader context of laws cited. Legal documents are of a specific domain: different contexts in the real world can lead to the violation of the same law, while the same context in the real world can violate different cases of law [2]. However, we are not aware of publications examining the topic modeling of legal documents considering the representation of the document from language models of the legal context.

3. Methodology

This section describes the approach used to identify and evaluate topics in legal documents. Initially, the paper presents the set of legal documents used to identify the topics, and after it indicates the methodology used for operating the topic modeling and evaluating activities.

3.1. Data Collection

We collected our primary set of legal documents from the Cornell Legal Information (Cornell LII)'s repository of Historic US Supreme Court Decisions representing the list of landmark court decisions in the United States. 314 legal cases were selected randomly and submitted to a cleaning process sweep all text associated with these cases available through the Cornell LII site. For each case, we then removed all HTML markup and editorial information and split the remaining text into paragraphs.

Landmark case categories and subcategories might be named in various ways with labeling processes derived from different criteria. Cases are often grouped by experts, organizations, or citizens to create a gallery of historic values of society. Data will be used both to represent the theme of the document and check the coherence of the modeled topics. For that, experts in the legal field analyzed each document, identifying two types of columns, the division and subdivision columns. This way, we aim to elect specific topics that are more useful for legal experts. Following patterns of historical analysis in which classification of documents must match the clustering purposes and fit previous experiences with the text itself, this proposed division and subdivision does create a strong meaning.

3.2. Topic Modelling

At its most basic level, topic modeling aims to capture the words that represent the concept of the document. Given a legal document dealing with capital punishment, the topic modeling algorithm can, for example, identify the following words "penalty, death, death penalty, punishment, capital punishment, execution, lethal, lethal injection, cruel, protocol" which can be the topic that the document represents.

Our data are defined in terms of documents $D = \{d_1, d_2, \ldots, d_N\}$ and of the paragraphs of each document $P_{di} = \{p_1, p_2, \ldots, p_n\}$. Thus, given a document composed of a set of paragraphs, in general, the objective is to cluster the paragraphs according to the contextual similarity (each cluster represents a topic) and then choose the topics that represent the main thematic of the document *di*. In addition, each paragraph is represented by a domain-specific contextual embedding; each topic is composed of a set of words that the approach identifies as most relevant to characterize it; and, the words of the topics chosen to represent the document identify the theme of the document. Therefore, the input of the approach is a text of the legal document and the output is the *k*-top words of the topic that represents the theme of the document.

We emphasize that our objective in this preliminary paper is not to discover the best architecture for this task but to provide a baseline to be used in future works.

Figure 1 shows the architecture of the topic modeling approach in legal documents used in this paper.

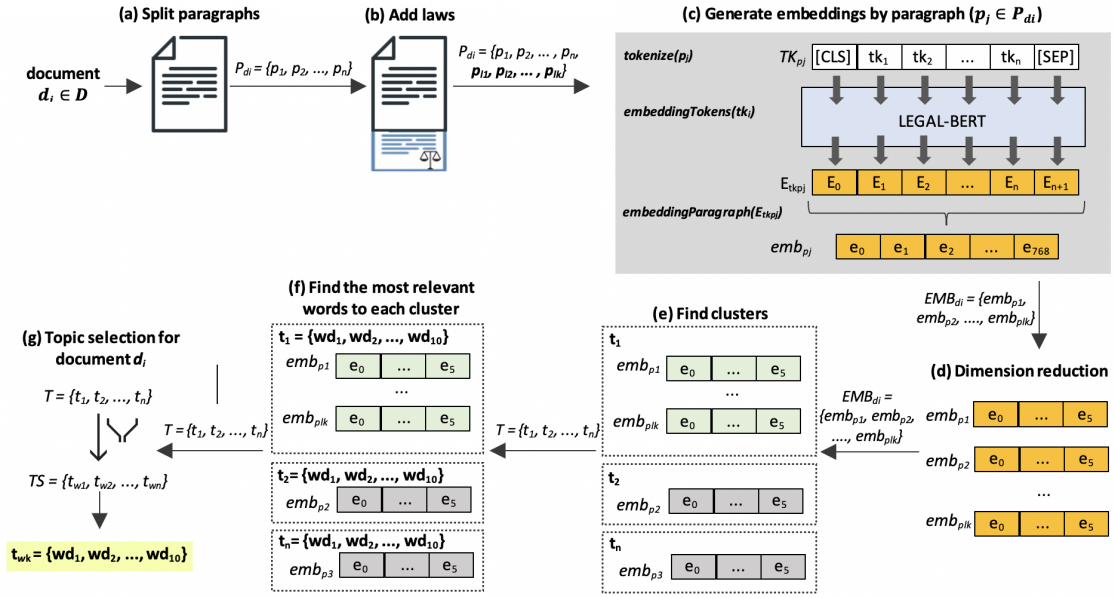

Figure 1: Architecture of the topic modeling approach in legal documents.

The approach presented in this paper identifies the document topics using BERTopic [8], a topic modeling technique that takes advantage of BERT embeddings [5], dimensionality reduction and clustering algorithms, as well as a class-based TF-IDF to create dense clusters, allowing interpretable topics from the extraction of the most important words from the clusters. In the following, we explain the topic modeling process, as well as some of the specific features of BERTopic.

We highlight that legal documents are known to be complex and written using a very peculiar structure and a specific set of words and expressions. They are often difficult to understand, are extensive, and can cite other cases and legislation. Initially, we split the documents into smaller units, that is, each document $d_i \in D$ is split into paragraphs $P_{di} = \{p_1, p_2, \ldots, p_n\}$, according to the original structure of the document (corresponding to "(a) Split paragraphs" in Figure 1).

One of our assumptions is that adding more information about the context of the documents increases the quality of the extracted topics. In the case of legal documents, citations to pre-existing cases and laws are as important as the content of the document itself. In this way, for each paragraph $p_j \in d_i$, we check that the paragraph contains a citation for the general and permanent laws of the United States (United States Code). If it contains, we add the text of the section of the laws cited to the set of paragraphs of the document $P_{di} = \{p_1, p_2, \ldots, p_n, \boldsymbol{p_{l1}}, \boldsymbol{p_{l2}}, \ldots, \boldsymbol{p_{lk}}\}$(step "(b) Add laws" in Figure 1). Finally, we remove any duplicate paragraphs.

Then, as shown in "(c) Generate embeddings by paragraphs ($p_j \in P_{di}$)" of Figure 1, we convert the elements of P_{di} in contextualized numerical vector representations of the legal domain, $EMB_{di} = \{emb_{p1}, emb_{p2}, \ldots, emb_{plk}\}$. We used the LEGAL-BERT [7] (an extended model of BERT pre-trained specifically for the legal domain) for this purpose, as it extracts different embeddings based on the context of the

legal texts. In this way, we obtain the vector representation, $emb_{pj} \in EMB_{di}$, for each paragraph $p_j \in P_{di}$, using equations (1) to (3) below:

$$TK_{pj} = tokenize(p_j),\ p_j \in P_{di} \quad (1)$$

$$E_{tkpj} = embeddingTokens(tk_i),\ tk_i \in TK_{pj} \quad (2)$$

$$emb_{pj} = embeddingParagraph(E_{tkpj}), \quad (3)$$

where, initially the *tokenize(p_j)* function adds the special tokens [CLS] and [SEP] at the beginning and end of the paragraph p_j, respectively, and splits it into subword tokens TK_{pj}, using the WordPiece algorithm [17], according to the LEGAL-BERT structure [7]. Then, the *embeddingTokens(tk_i)* function, vectorizes (embedding) each token $tk_i \in TK_{pj}$, considering the 768 hidden units of the hidden state of encoding last layer returned by the pre-trained model LEGAL-BERT [7]. Finally, the $embeddingParagraph(E_{tkpj})$function averages the embeddings for each embedding of token $emb_{tk} \in E_{tkpj}$, setting to emb_{pj} the embedding of the text of the entire paragraph (a vector representation of 768 units).

Before using a clustering algorithm, we first need to reduce the dimensionality of the embeddings of the paragraphs, since many clustering algorithms deal poorly with the high dimensionality. Dimension reduction allows for dense clusters of documents to be found more efficiently and accurately in the reduced space [1]. Among the dimensionality reduction algorithms, the Uniform Manifold Approximation and Projection (UMAP) [18] preserves the high-dimensional local and global structure in lower dimensionality and is capable of scaling to very large data sets. We used UMAP and reduced the dimensionality to 5 (corresponding to "(d) Dimension reduction" in Figure 1). UMAP has several hyperparameters that determine how it performs the dimension reduction. Possibly the most important parameter is the *number of nearest neighbors*. This parameter controls the balance between the preservation of global and local structures in the low dimensional embedding. We set the *number of nearest neighbors* to 15.

After having reduced the dimensionality of the embeddings of the paragraphs, we can cluster them, as shown in "(e) Find clusters" in Figure 1. The goal of density-based clustering is to find areas of highly similar embeddings in the semantic space, which indicate an underlying topic. This is performed on the UMAP reduced embeddings. We used Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [19] to find dense areas of embeddings without forcing data points into clusters, as we consider them outliers. The number of clusters has not been defined, while the minimum size of paragraphs in each cluster has been set to 5. In this way, the algorithm will try to find the ideal number of clusters, grouping similar paragraphs, whose clusters must represent the topics of the paragraphs.

Then, we identified a set of words that represent the content of each cluster (step "(f) Find the most relevant words to each cluster" in Figure 1). For this, a variant of the TF-IDF (*Term Frequency - Inverse Document Frequency)* is structured in clusters, named c-TF-IDF. The c-TF-IDF compares the importance of words to a specific cluster, revealing the most significant words in a topic, according to the TF-IDF score. The c-TFxIDF is calculated according to equation (4) below:

$$c - TF - IDF = \frac{f_i}{wd_i} \times log \frac{m}{\sum_j^n f_j}, \quad (4)$$

where the frequency of each word *f* is extracted from each cluster *i* and divided by the total number of words *wd* of cluster *i*. This action can be seen as a way of normalizing the frequency of words in the cluster. Then the number of clusters *m* is divided by the total frequency of the word *f* across all clusters.

To create a topic representation, we obtain the top-10 most representative words of each topic, based on their scores in c-TF-IDF. The higher the score, the more representative the word must be for the cluster, therefore for the topic.

After grouping similar paragraphs and identifying the most representative words for each topic, we selected the topics to characterize the document (corresponding to "(g) Topics selection for document d_i" in Figure 1). Each topic $t_k \in T$ receives a weight w_k. Our intuition is that the clusters with the largest number of paragraphs best represent the theme of the document since most paragraphs will be related to the main subject of the document, while a smaller number of paragraphs will be related to complementary subjects, but not the main subject. In this way, the weight w_k of the topic t_k is the number

of paragraphs clustered in the topic $|t_k|$. T_S represents the topics of T sorted (in descending order) by topic weight. Therefore, to represent the topics of the document, we select the n topics $t_k \in T_S$ until achieving a threshold.

3.3. Evaluation

The topic modeling approach described in this paper has been applied in a set of legal documents characterized as US landmark cases. These documents have divisions and subdivisions that suggest the main theme of the document.

Two variations of the approach were evaluated. These variations are associated with the representation of the document: (1) the document is represented only by the paragraphs that form it $P_{di} = \{p_1, p_2, \ldots, p_n\}$; (2) extending the semantic representation of the document, with the insertion of the text of the laws cited in the document to the set of paragraphs of the document, $P_{di} = \{p_1, p_2, \ldots, p_n, \boldsymbol{p_{l1}}, \boldsymbol{p_{l2}}, \ldots, \boldsymbol{p_{lk}}\}$.

The quality of the topic models can be evaluated in different ways. We carry out a qualitative assessment under the criterion of interpretability, that is, how the terms that define the topic from a consistent and coherent meaning can be understood by humans. For this, two experts in the legal field performed a manual inspection on the set of words most representative of the topics selected by the model (for example, the 10 most important words). From this inspection, the experts recorded whether there is a semantic correspondence of these words concerning the main thematic of each legal document analyzed (comparing them to the text and the division and subdivision of the document), indicating, if so, that the topics selected by the model represent the main theme of the document.

Then, the Kappa coefficient [20] was used to assess the degree of agreement between experts, calculated using equation (5) below:

$$Kappa = \frac{po - pe}{1 - pe}, \tag{5}$$

where *po* represents the observed proportion of concordances (sum of the concordant responses divided by the total); and *pe* represents the expected proportion of concordances (sum of the expected values of the concordant responses divided by the total).

Although there is no specific objective value from which the value of the Kappa coefficient should be considered as adequate, there are some suggestions in the literature that normally guide this decision, highlighting the proposal of [21], where Kappa < 0.40 indicates poor agreement; Kappa between 0.40 and 0.75, represents satisfactory to good; while Kappa > 0.75 represents excellent agreement.

4. Results and Analysis

In this section, we analyze the topics retrieved by the approach for each document. Initially, we evaluated the topics modeled according to the two variations of representation of the document. When using the representation of the document only with the paragraphs that compose it, in approximately 8% of the documents, the approach fails to model the topics. All of these documents have less than 100 paragraphs. Our observation is that the approach does not have enough information to group paragraphs according to the same semantics. By expanding the representation of the document with the insertion of the text of the aforementioned laws, we are adding more information on the subject of the document, thus, only 5% of the documents had no topics modeled. Thus, considering that the addition of laws in the representation of the document improves the quality of the theme modeling, the rest of the evaluation was carried out in this scenario.

From the qualitative evaluation carried out by the specialists, we obtain that 84.6% of the topics selected by the model correspond to the main theme of the document (considering the evaluations in which there is an agreement between the experts). We emphasize that the level of agreement of the evaluators, measured by the Kappa coefficient, is 0.78, qualitatively representing the level of agreement excellent, according to the approach of [21].

Table 1 shows the top-10 representative words (according to c-TF-IDF described in section 3.2 Topic Modeling) extracted by the topic modeling approach presented in this paper for a subset of the dataset with ten legal documents of different thematic. The words listed for each topic appear in

descending order, from the highest to least c-TF-IDF. More specifically, the documents are associated with the following themes: capital punishment, detainment of terrorism suspects, passengers and interstate commerce, federal native American law, Amish, freedom of speech and of the press, end of life, copyright/patents, federalism, and birth control and abortion, respectively.

Table 1
Topics modeled to legal documents.

ID	Division	Subdivision	Topics
D1	Criminal law	Capital punishment	death, execution, risk, id, injection, penalty, pain, lethal, punishment, protocol
D2	Criminal law	Detainment of terrorism suspects	court, jurisdiction, habeas, states, united, united states, courts, district, eisentrager, writ
D3	Equal Protection Clause	Passengers and Interstate Commerce	statute, interstate, state, commerce, court, passengers, led, states, sct, virginia
D4	Federal Native American law	Federal Native American law	indian, non indians, jurisdiction, non indian, try, courts, congress, tribes, indian tribes, try nonindians
D5	First Amendment rights	Amish	amish', 'education', 'children', 'religious', 'school', 'life', 'state', 'child', 'parents', 'compulsory
D6	First Amendment rights	Freedom of speech and of the press	sct, states, united states, present, led2d, danger, present danger, clear present
D7	Individual rights	End of life	new, suicide, treatment, medical, health, sct, york, new york, ann, patients
D8	Intellectual Property	Copyright/Patents	copyright, work, facts, original, works, protection, originality, act, author, telephone
D9	Tax Law	Federalism	direct, constitution, tax, taxes, apportioned, apportionment, cases, rule, present, indirect
D10	Women's rights	Birth control and abortion	abortion, procedure, state, fetus, court, medical, law, statute, dx, id

One way to evaluate topic modeling is to analyze how well the topics describe the documents. This assessment measures how informative the topics are to a user. Thus, when inspecting the topic model, we can confirm that some topics provide information about the document (D1, D3, D4, D5, D7, D8, D9, and D10, according to Table 1), that is, the words of the topic associated to the document are semantically related to the thematic of that legal document (represented by the division and subdivision), making it possible to identify the subject of the document. For example, the words "death, execution, risk, id, injection, penalty, pain, lethal, punishment, protocol" allow us to summarize the subject of "capital punishment". This example of a summary allows a user to identify the subject related to certain legal matters or simply summarize the content of a legal document by analyzing the topic of the document. It should be noted that the topics extracted for documents D2 and D6 do not provide information about the document. The authors assume that these documents are short, therefore presenting little information to cluster significant paragraphs with the theme of the document.

To illustrate the visualization of the topics generated by the approach, in Figure 2 we show the 2-dimensional t-space (reduction of dimensionality performed using UMAP) of the embeddings of the paragraphs of a legal document dealing with "capital punishment". Specifically, semantically similar texts must be close to each other in the vector space of embeddings, while different texts must be more distant from each other. In Figure 2, each circled area represents a cluster identified by the clustering algorithm (HDBSCAN). In this case, the document's paragraphs were clustered into 4 topics. The T1

topic has the largest number of paragraphs and is therefore chosen to represent the subject of the document. It is observed that the other topics (T2, T3, and T4) are distant from T1, capturing relatively different topics in the legal document. When applying c-TF-IDF, we obtain the following top-5 most representative words for the topic T1 "death, execution, risk, id, injection". While the following top-5 words for topics T2, T3 and T4, respectively, "503, 503 653, 653, 536 304, 536", "428 153, 153, 1994, 1976, 428" and "130 1879, 99 130 1879, 1879, 99 130, 99 ". Therefore, we emphasize that the words in the topic T1 (chosen to represent the document) are consistent with the theme of that document (capital punishment).

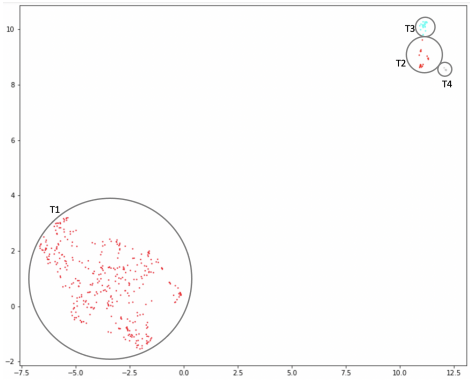

Figure 2: 2-dimensional projection of the vectorial space of the paragraphs of a legal document on the subject of the capital punishment.

The overview of the most significant words in the document topic enhances the understanding of the document's subject. A word cloud was also generated for the top-30 words of the topic T1 of the document shown in Figure 2, according to the c-TF-IDF, to observe the most important terms for the topic, as shown in Figure 3 below.

Figure 3: Most relevant words for the topic, according to c-TF-IDF.

Although the approach presented in this paper is still initial, it offers an attractive way to automate the summary of legal documents quickly. It can be useful when we have a large amount of text data and we want to identify the subject of a particular legal document. In this situation, we can classify and search a large number of documents more efficiently.

5. Conclusions

We propose the use of BERTopic to build thematic models of legal documents. The legal text has specific characteristics, such as specialized vocabulary, formal syntax, semantics based on an extensive specific domain of knowledge, and presents citations to other cases, statutory law, the Constitution, and amendments. In this way, we represent the text contextually from the LEGAL-BERT (pre-trained model for the legal domain) and provide information about the laws mentioned in the document. From a qualitative assessment, the approach presents good results, revealing topics consistent with the document's theme.

This preliminary approach can be used as a baseline for future papers. In the future, it is intended to explore different strategies for choosing the topics of a document, as well as to quantitatively evaluate the interpretability and coherence of the topics and to compare the proposed approach with other approaches of the state of the art. It is also intended to extend the approach to clustering documents according to the modeled topics.

6. References

[1] Dimo Angelov. Top2Vec: Distributed Representations of Topics. arXiv:2008.09470v1, (2020).

[2] A. Kanapala, S. Pal, R. Pamula. Text summarization from legal documents: a survey. Artificial Intelligence Review 51(3), 371–402 (2019).

[3] Ylja Remmits. Finding the Topics of Case Law: Latent Dirichlet Allocation on Supreme Court Decisions. Thesis. Radboad Universiteit, (2017).

[4] Laure Thompson, David Mimno. Topic Modeling with Contextualized Word Representation Clusters. arXiv:2010.12626v1, (2020).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics, (2019).

[6] Christopher Williams. Tradition and change in legal English: Verbal constructions in prescriptive texts, volume 20. Peter Lang, (2007).

[7] Ilias Chalkidis, Manos Fergadiotis. LEGAL-BERT: The Muppets straight out of Law School, arXiv:2010.02559v1, (2020).

[8] Maarten Grootendorst. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. doi: 10.5281/zenodo.4381785, (2020).

[9] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. J. Mach. Learn. Res. 3:993-1022 (2003).

[10] Pedro Henrique Luz de Araújo, and Teófilo de Campos. Topic Modelling Brazilian Supreme Court Lawsuits. JURI SAYS, 113, (2020).

[11] James O' Neill, Cécile Robin, Leona O' Brien, Paul Buitelaar. An Analysis of Topic Modelling for Legislative Texts. ASAIL 2017, London, UK, June 16, (2017).

[12] Daniel D. Lee, and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. Nature. 401 (6755): 788–791. (1999).

[13] Susan T. Dumais. Latent Semantic Analysis. Annual Review of Information Science and Technology. 38: 188–230, (2005).

[14] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet Processes. Journal of the American Statistical Association. 101 (476): pp. 1566–1581, (2006).

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 1(8):9, (2019).

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Man- dar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. ArXiv, abs/1907.11692, (2019).

[17] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin John- son, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rud- nick, Oriol Vinyals, Gregory S. Corrado, Macduff Hughes, and Jeffrey Dean. Google's neural machine translation system: Bridging the gap between human and machine translation. ArXiv, abs/1609.08144, (2016).

[18] L. McInnes and J. Healy. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, (2018).

[19] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. Journal of Open Source Software, 2(11), 205, doi:10.21105/joss.00205, (2017).

[20] Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-46, (1960).

[21] Fleiss, J. (1981). Statistical methods for rates and proportions (2th Ed.). New York: John Wiley & Sons, (1981).