# Organizing Portuguese Legal Documents through Topic Discovery

Daniela Vianna
Edleno Silva de Moura
dvianna@icomp.ufam.edu.br
edleno@icomp.ufam.edu.br
Instituto de Computação – Universidade Federal do Amazonas
Manaus, Brazil
Jusbrasil
Salvador, Brazil

## ABSTRACT

A significant challenge in the legal domain is to organize and summarize a constantly growing collection of legal documents, uncovering hidden topics, or themes, that later can support tasks such as legal case retrieval and legal judgment prediction. This massive amount of digital legal documents, combined with the inherent complexity of judiciary systems worldwide, presents a promising scenario for Machine Learning solutions, mainly those taking advantage of all the advancements in the area of Natural Language Processing (NLP). It is in this scenario that Jusbrasil, the largest legal tech company in Brazil, is situated. Using a dataset partially curated by the Jusbrasil legal team, we explore topic modeling solutions using state of the art language models, trained with legal Portuguese documents, to automatically organize and summarize this complex collection of documents. Instead of using an entire legal case, which usually is composed of many pages, we show that it is possible to efficiently organize the collection using the syllabus (in Portuguese, ementa jurisprudencial) from each court decision as they concisely summarize the main points presented by the entire decision.

## CCS CONCEPTS

• **Information systems** → **Information retrieval**.

## KEYWORDS

law tech, legal cases, topic model, language models

**ACM Reference Format:**
Daniela Vianna and Edleno Silva de Moura. 2022. Organizing Portuguese Legal Documents through Topic Discovery. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3477495.3536329

## 1 INTRODUCTION

The legal sector is well known for its document-intensive routine that involves an overabundance of text information varying from legal rulings to past cases and contracts. Keeping track and making sense of this huge collection of data is time-consuming and costly. With the advances in Artificial Intelligence (AI), more specifically in the Machine Learning and Natural Language Processing fields, the way the legal sector operates is rapidly changing. AI technologies can be applied to support tasks such as legal case retrieval [11, 12, 16], legal judgment prediction [4, 20], and text classification in the legal domain [5].

Brazil, a Portuguese speaking country, has a very large and complex judiciary system with over 25 millions new legal cases in the year of 2020 and over 75 millions cases pending decision [1]. Driven by the COVID-19 pandemic, the vast majority of the new cases in 2020, around 97%, were submitted to the courts electronically, with 65% of the Brazilian courts with 100% of their new cases electronically submitted. With a huge corpus of electronically submitted legal cases, intelligent methods geared towards legal data are extremely necessary to foster a fast and fair judiciary system. It is in this scenario that Jusbrasil[2], the largest legal tech company in Brazil, with around 2 millions access per day, operates.

As a important step towards building efficient intelligent tools for legal data, we tackle the challenging problem of organizing and summarizing a complex legal document collection. Our approach to this problem relies on topic discovery techniques, a research area that have resurfaced in recent years boosted by advances in neural topic models and contextual representations for text data, such as the transformer-based autoencoders methods. More specifically, in this work, we explore three different approaches for topic modeling, CombinedTM (CTM) [3], Top2Vec [1], and BERTopic [7].

One of the challenges of working with legal documents, is the complexity and uniqueness of the legal language. Legal documents are full of specific words and phrases that have meaning only as legal terms, in addition, well known words can take on different or new meanings when used by legal practitioners. Hence, the legal language demands its own language model. To this end, we evaluate different text representation models including Doc2Vec and transformer-based language models pre-trained using Portuguese text, BERTimbau [13] and BERTikal [10]. Experiments performed

[1] https://www.cnj.jus.br/pesquisas-judiciarias/justica-em-numeros/
[2] https://sobre.jusbrasil.com.br/ in Portuguese only.

*SIGIR '22, July 11–15, 2022, Madrid, Spain*

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3536329

over a curated collection of legal documents show that topic discovery techniques outperform topic model approaches in our legal scenario, with Top2Vec and BERTikal greatly improving the quality of topics created.

## 2 RELATED WORK

One of the most relevant problem in the legal domain is the retrieval of similar legal cases. Badenes-Olmedo et al. [2], presents an approach to solve cross-lingual document similarity through multilingual probabilistic topic model (MuPTM) [17]. In this process, the topic model was an important tool to represent multilingual documents without the requirement of dictionaries or a machine-translation system. In a quest to compute similarity between query and documents, Nanda et al. [9] used topic models to group documents into cluster of topics such that each document has a topic vector which can be compared to that of a query, selecting the initial top-n most similar documents to the query. The top-n documents can be passed through a semantic similarity model that will identify the most relevant documents to the given query. Mandal et al. [8] uses a topic model to generate a vector representation of the documents before cosine similarity can be used to measure the similarity between documents or parts of the document. Legal Classification is another task that have applied topic modeling with success. In the work presented by Fang et al. [6], a topic model approach is adopted to automatically annotate clusters of documents facilitating experts access to cluster topics without having to read verbatim, resulting in a faster and efficient data annotation process. Those are some of the ways topic modeling can be used to support challenging tasks in the legal domain. In our work, we study a combination of topic modeling/topic discovery approaches, combined with the state-of-the art solutions to NLP, to organize a large collection of Brazilian Portuguese legal documents with the intentions of serving tasks as the ones mentioned here.

## 3 EXPERIMENTAL METHODOLOGY

This section presents the experimental methodology we adopted. We start by describing the three topic discovery models together with a description of the embedding-based text representations. Finally, we present the legal dataset and the metrics used to evaluate the methods.

### Models

**CombinedTM (CTM):** an extension of Neural ProdLDA [14] that uses contextualized embeddings by combining BoW with BERT-like models. In combination with CTM we have evaluated two different transformer-based models: BERTimbau and BERTikal. BERTimbau is a BERT-like model pre-trained on the BrWaC (Brazilian Web as Corpus), a large Portuguese corpus [18]. BERTikal is a language model resulting from continue pre-training BERTimbau on portuguese legal documents [10].

**Top2Vec:** differently from CTM, top2vec is a topic discovery approach that tries to find dense clusters of documents from an embedded representation of words and documents [1]. In this work, we use Doc2Vec to jointly create embedded document and word vectors that later will be used by the clustering algorithm HDBSCAN to find dense areas in the document space.

**BERTopic:** similar to Top2Vec, it also creates dense clusters by leveraging embeddings and c-TF-IDF, a class-based TF-IDF procedure [7]. BERTopic was evaluated with three different embeddings approaches: Doc2Vec, BERTikal, and BERTimbau.

### Datasets

We evaluate the models on a dataset composed of 2864 Brazilian court decisions from the Jusbrasil document collection. From the 2864 documents, we do not have any previous knowledge about the topics that comprises 2439 of those documents. The remaining 425 documents are distributed between six different reference collections with topics defined in advanced by legal specialists from Jusbrasil staff. Those reference collections are used to support the evaluation of the quality of the topics found by the models. The reference collections are: collection 1 (RC1), composed of seven legal decisions about the legality of companies to outsource their main activities; collection 2 (RC2), composed of eight legal decisions about an institute of Civil Law, the abusive use of legal personality; collection 3 (RC3), composed of eight legal decisions about public employees retirement system; collection 4 (RC4), composed of 18 legal decisions about a specific tax charged on properties located in rural areas of Brazil, the ITR, which stands for "Imposto sobre Propriedade Territorial Rural" and can be translate to "Tax on Rural Land Property"; collection 5 (RC5), composed of 184 legal decisions related to moral and material damages caused by delays and cancellations of flights; and finally, collection 6 (RC6), composed of 200 legal decisions associated to Brazilian environment issues, such as the deforestation of the Amazon rainforest. In this work, we focus on the syllabus (in Portuguese, *ementa jurisprudencial*) from each court decision as they concisely summarize the main points presented by the entire decision. RC6 is the only reference collection that does not provide syllabi. To circumvent this issue, we use a predefined list of relevant words provide by legal specialists to select text snippets in the decisions that together emulate a syllabus.

### Data pre-processing

Considering the characteristics of each model evaluated and the range of possibilities brought by traditional procedures of preprocessing, in this work we explore the following parameters: to remove or extract content from URLs that appear in the input data – all syllabus in our dataset contains one or more URL carefully curated to represent relevant laws and decrees –, to ignore terms that have a document frequency lower or higher than a given threshold, to use a standard or augmented stopword list, and finally to use different token patterns that will preserve or not digits when part of a word. Combining all those parameters we got a total of 22 different combinations for our data pre-processing step. We also apply some basic cleaning procedures such as removing punctuation, removing extra white spaces, lower-casing terms, and stripping accents. The exception goes to transformer-based models, where we keep stopwords and do not lowercase terms. Notice that transformer-based models usually flourish with minimal data cleaning.

### Metrics

**Precision ($p$):** this is a trick metric to use in this work since it relies on the knowledge of true positives, and, in our scenario, we can not affirm that a document assigned to a topic, that is not a part of a known collection, is a false positive. With that in mind, we use precision as a tool to raise flags when the number of documents assigned to a topic surpass the number of documents in a reference collection supposed to be assigned to the same topic. Those cases can be further investigated by a specialist.

**Recall ($r$):** is an important metric to point out if the expected documents from a reference collection were correctly assigned to the same topic. It cannot be used as a hard metric, since the topic modeling solution can find multiples topics for one reference collection when a growing number of topics is generated. However, they can definitely guide our analysis.

**Modified $F_1$-score ($F_1$):** as mentioned before, precision is not a trustworthy metric in our scenario. We rely more on recall as an indication of topic quality. With that, we modified the traditional F1-score giving more weight to recall. The Modified F1-score is: $F_1 = (1+\beta^2)\frac{p \times r}{(\beta^2 \times p)+r)}$, where $p$ stands for precision and $r$ for recall. We adopted $\beta = 2$.

**Weighted modified F1-score ($wF_1$):** is a weighted version of the modified $F_1$ score, where all reference collections are treated equally when evaluating the overall performance of the topic modeling solution.

**External word embeddings topic coherence ($\alpha$):** follows Bianchi et al. [3]. Initially, we compute the average pairwise cosine similarity of the word embedding of the top-10 words in a topic; then, the average of those values for all topics is computed.

**Inverted Rank-Biased Overlap ($\rho$):** follows the approach adopted by Bianchi et al. [3] to compute topic diversity. RBO, proposed by Webber et al. [19] and Terragni et al. [15], compares the top-10 words of two topics allowing disjointedness between lists of topics and using weighted ranking to penalize lists that shares the same words according to the position of those word in the topic.

## 4 EVALUATION

In this section we evaluate the three approaches for topic modeling and topic recovery: CTM, Top2Vec, and BERTopic. The models will be evaluated using the data collection, pre-processing parameters, and metrics presented in Section 3. We first analyze the models performance in regards to one reference collection, CR5. Then, we will evaluate the models performance when considering all six reference collections. All the results refer to the average of three executions with different seeds, 10, 500, and 2021.

For this first study, we evaluated the performance of topic modeling/discovery approaches, CTM, Top2Vec, and BERTopic, when considering only the reference collection RC5 as our golden set. In combination with those models, three text representation models were also evaluated: Doc2Vec, BERTimbau, and BERTikal. The dataset is comprised by the documents from RC5, 184 legal decisions related to the topic "moral and material damages caused by flight delays and cancellations", plus 2439 documents from unknown topics. For the topic model approach, CTM, the parameter number of topics varied from 5 to 20. The topic discovery approaches do not required the number of topics to be predefined, which can be an advantage with one less parameter to be tuned.

Table 1 presents a compilation of the best results for each of the three models. In the first column we have the embedding adopted by each model. The following columns show the number of topics learned (#t), average recall (Avg $r$), average precision (Avg $p$), average F1-score (Avg $F_1$), average coherence (Avg $\alpha$), and finally average topic diversity (Avg $\rho$). The best values are highlighted in bold. CTM was evaluated using BERTimbau and BERTikal. CTM with BERTikal outperformed CTM with BERTimbau for all 22 combination of parameters, with some small variation in performance between different combinations of parameters. Top2Vec was evaluated using Doc2Vec, while BERTopic was evaluated using Doc2Vec, BERTimbau, and BERTikal. Both Top2Vec and BERTopic performed very well across all combination of parameters, with Doc2Vec being the best text representation for those two models. As far as coherence ($\alpha$) is concerned, Top2Vec is superior to CTM and BERTopic, which was also confirmed by a careful analyses conducted by a group of specialists. Regarding topic diversity ($\rho$), all models show similar values, with Top2Vec presenting a slight higher value than the other models.

Analyzing the results obtained with this first study, we noticed that in general, Top2Vec were finding between 20 and 25 topics for the majority of the pre-processing parameters. This value was even higher for BERTopic. While for CTM, we were varying the number of topics from 5 to 20, not including 20. With that in mind, we ran a few experiments increasing this range to 40, including 40. We noticed a small improvement on the quality of CTM, mainly regarding to precision; however, we still observed that CTM tends to find a smaller number of topics than Top2Vec and BERTopic.

For our second study, we extend the evaluation to include all six reference collections which implies a total of 2864 Brazilian court decisions, including 425 documents for the six reference collections and the 2439 documents with unknown topics. For this study, we analyze CTM with BERTikal, Top2Vec with Doc2Vec, and BERTopic with Doc2Vec and BERTikal. For CTM, we consider the number of topics varying from 5 to 40, including 40. In regards to the metrics, we concentrate on the average of weighted F1-score computed based on three runs with seeds, 10, 500, and 2021. For the best results, we also compute topic diversity and topic coherence. Those results are then evaluated by specialists. Table 2 presents the embedding, total number of topics (#t), number of unique topics for all reference collections (#rct), the average of weighted F1-score (Avg w_F1), the average of topic coherence ($\alpha$), and average of topic diversity ($\rho$) for the combination of models and pre-processing parameter that achieved the higher average F1-score.

Overall, considering the number of unique topics for all reference collections (#trc), BERTopic was more consistent across all combinations of pre-processing parameters, finding, for most cases, one topic for each reference collection evaluated. The CTM model repeatedly found five different topics for the six reference collections, with RC1 and RC3 sharing a topic. The same happened with model top2vec, with RC1 and RC2 also sharing a topic. Differently, BERTopic was able to group the majority of documents from each individual collection in unique topics. Even though Top2Vec were more consistent across the set of pre-processing parameters, in regards to average weighted F1-score, BERTopic was the model that achieved the best value for this metric. In regards to topic coherence and topic diversity, Top2Vec is the model with higher coherence and diversity.

Legal specialists carefully evaluated some of our most promising results. By their analysis, CTM usually does not find words that are very informative about a collection theme, the words are often

**Table 1: Avg. of recall, precision, F1-score, coherence, and diversity for the best approaches and collection RF5.**

| model | embedding | #t | Avg $r$ | Avg $p$ | Avg $F_1$ | Avg $\alpha$ | Avg $\rho$ |
|---|---|---|---|---|---|---|---|
| CTM | BERTikal | 11 | 0.873 | 0.624 | 0.808 | 0.021 | 0.927 |
| Top2Vec | Doc2Vec | 24 | **0.946** | 0.761 | 0.915 | **0.072** | **0.994** |
| BERTopic | Doc2Vec | 32 | 0.938 | **0.940** | **0.939** | 0.058 | 0.977 |
| BERTopic | Doc2Vec | 35 | 0.938 | **0.940** | **0.939** | 0.065 | 0.984 |

**Table 2: Avg weighted F1-score, avg. coherence, and avg. diversity for the best approaches and all reference collections.**

| model | embedding | #t | #rct | Avg wF1 | Avg $\alpha$ | Avg $\rho$ |
|---|---|---|---|---|---|---|
| CTM | BERTikal | 12 | 5 | 0.805 | 0.039 | 0.909 |
| Top2Vec | Doc2Vec | 23 | 5 | 0.876 | **0.078** | **0.997** |
| BERTopic | BERTikal | 38 | 6 | **0.880** | 0.043 | 0.981 |

generic and do not bring much knowledge when combined. The exception was RC5, the one examined in Table 1, that refers to moral and material damages caused by delays and cancellations of flights. Words such as 'moral', 'damage', 'flight', 'delay', 'liability', and 'consumer' were usually present in the topic representation. Top2Vec and BERTopic were able to find more meaningful terms, with BERTopic, that allows for ngrams in the list of terms per topic, outperforming all the other models. For instance, for RC5, BERTopic found relevant bigrams such as 'material damage', 'moral damage'. 'air transport', adding to relevant unigrams, 'flight', 'delay', 'baggage', and 'cancellation'. All models underperformed for collection RC2, being unable to find terms that would help identifying the underlying theme of this RC. This could be explained by the quality of the documents in the reference collection and the fact that they do not address a very distinctive theme. Also, the collection is very small which contribute to the bad results.

In regards to the pre-processing parameters, we observe that extracting content from URLs usually leads to results with better quality. Also, removing isolated digits and terms with a document frequency higher than 75%.

In conclusion, Top2Vec with Doc2Vec and BERTopic with BERTikal have shown to be the best choices to organize the collection of Portuguese legal documents evaluated in this work, with BERTopic with bigrams finding more meaningful terms from the point of view of the legal specialists, and Top2Vec performing more consistently across a variety of parameters.

## 5 CONCLUSION

In this work we address the challenge of efficiently organizing Brazilian Portuguese legal documents. To this end, we evaluated three different approaches for topic modeling and topic discovery, CombinedTM (CTM), Top2Vec, and BERTopic. Given the complexity and uniqueness of the legal language, we concentrated on using text representation models trained using Brazilian Portuguese texts, with some of them adjusted using texts from the legal domain. The evaluations were conducted using a collection of 2864 Brazilian court decisions. From those 2864 documents, we do not have previous knowledge about the themes that compose 2439 of those documents. The remaining 425 documents are distributed between six different reference collections, our golden sets, with topics defined in advanced by legal specialists. During our study, we considered a broad range of parameters for data pre-processing, with Top2Vec being the model that seems to be less impacted by changes in pre-processing parameters. Most models seems to work better with a list of stopwords tailored for the legal domain. Overal, Top2Vec and BERTopic performed well for the six reference collections, with the BERTopic flexibility of building topic representations using meaningful ngrams, a plus from the point of view of legal specialists. Currently, at Jusbrasil, we are applying those solutions to organize documents from The Brazilian General Data Protection Law (Lei Geral de Proteção de Dados Pessoais or LGPD, in Portuguese) collection, supporting legal specialists in their daily work. One future goal is to be able to organize Jusbrasil impressive collection of legal documents supporting more efficient solutions for tasks such as legal case retrieval and legal judgment prediction.

## 6 COMPANY PORTRAIT

Jusbrasil, the largest legal tech company in Brazil, combines law and technology approximating Brazilians to the justice. With millions of users accessing the company's platform daily and a very large and complex dataset composed of billions of documents related to the Brazilian judiciary system, Jusbrasil is constantly seeking the development of state-of-the-art intelligent methods geared towards legal data, with the goal of nurturing a fast and fair judiciary system. The company offers a variety of legal related products, including an effective and efficient search system.

## 7 PRESENTER

Daniela Vianna is a postdoc researcher at Universidade Federal do Amazonas (UFAM) in a partnership project with Jusbrasil. Daniela has a Ph.D. in Computer Science from Rutgers University, USA, and a Master's and Bachelor's degree in Computer Science from Universidade Federal Fluminense (UFF), Brazil. Her interests are in the areas of Information Retrieval and Natural Language Processing.

## ACKNOWLEDGMENTS

This research is partially supported by Jusbrasil Postdoctoral Fellowship Program, Brazilian funding agency FAPEAM-POSGRAD 2020 (Resolution 002/2020), the Coordination for the Improvement of Higher Education Personnel-Brazil (CAPES) financial code 001, and individual grant from CNPq (309765/2018-8) to Edleno Silva de Moura.

## REFERENCES

[1] Dimo Angelov. 2020. Top2Vec: Distributed Representations of Topics. *CoRR* abs/2008.09470 (2020). arXiv:2008.09470 https://arxiv.org/abs/2008.09470

[2] Carlos Badenes-Olmedo, José Luis Redondo-García, and Oscar Corcho. 2019. Scalable Cross-Lingual Document Similarity through Language-Specific Concept Hierarchies. In *Proceedings of the 10th International Conference on Knowledge Capture* (Marina Del Rey, CA, USA) *(K-CAP '19)*. Association for Computing Machinery, New York, NY, USA, 147–153. https://doi.org/10.1145/3360901.3364444

[3] Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Online, 759–766. https://doi.org/10.18653/v1/2021.acl-short.96

[4] Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural Legal Judgment Prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4317–4323. https://doi.org/10.18653/v1/P19-1424

[5] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-Scale Multi-Label Text Classification on EU Legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6314–6322. https://doi.org/10.18653/v1/P19-1636

[6] Yin Fang, Xin Tian, Hao Wu, Songyuan Gu, Zhu Wang, Feng Wang, Junliang Li, and Yang Weng. 2020. Few-Shot Learning for Chinese Legal Controversial Issues Classification. *IEEE Access* 8 (2020), 75022–75034.

[7] Maarten Grootendorst. 2020. BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. https://doi.org/10.5281/zenodo.4381785

[8] Arpan Mandal, Raktim Chaki, Sarbajit Saha, Kripabandhu Ghosh, Arindam Pal, and Saptarshi Ghosh. 2017. Measuring Similarity among Legal Court Case Documents. In *Proceedings of the 10th Annual ACM India Compute Conference* (Bhopal, India) *(Compute '17)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3140107.3140119

[9] Rohan Nanda, Adebayo Kolawole John, Luigi Di Caro, Guido Boella, and Livio Robaldo. 2017. Legal Information Retrieval Using Topic Clustering and Neural Networks. In *COLIEE 2017. 4th Competition on Legal Information Extraction and Entailment (EPiC Series in Computing, Vol. 47)*, Ken Satoh, Mi-Young Kim, Yoshinobu Kano, Randy Goebel, and Tiago Oliveira (Eds.). EasyChair, 68–78. https://doi.org/10.29007/psgx

[10] Felipe Maia Polo, Gabriel Caiaffa Floriano Mendonça, Kauê Capellato J Parreira, Lucka Gianvechio, Peterson Cordeiro, Jonathan Batista Ferreira, Leticia Maria Paz de Lima, Antônio Carlos do Amaral Maia, and Renato Vicente. 2021. LegalNLP-Natural Language Processing methods for the Brazilian Legal Language. In *Anais do XVIII Encontro Nacional de Inteligência Artificial e Comput.* SBC, 763–774.

[11] Guilherme Moraes Rosa, Ruan Chaves Rodrigues, Roberto de Alencar Lotufo, and Rodrigo Nogueira. 2021. Yes, BM25 is a Strong Baseline for Legal Case Retrieval. *CoRR* abs/2105.05686 (2021). arXiv:2105.05686 https://arxiv.org/abs/2105.05686

[12] Yunqiu Shao, Jiaxin Mao, Yiqun Liu, Weizhi Ma, Ken Satoh, Min Zhang, and Shaoping Ma. 2020. BERT-PLI: Modeling Paragraph-Level Interactions for Legal Case Retrieval. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3501–3507. https://doi.org/10.24963/ijcai.2020/484 Main track.

[13] Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

[14] Akash Srivastava and Charles Sutton. 2017. Autoencoding Variational Inference For Topic Models. *arXiv e-prints*, Article arXiv:1703.01488 (March 2017), arXiv:1703.01488 pages. arXiv:1703.01488 [stat.ML]

[15] Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021. Word Embedding-Based Topic Similarity Measures. In *Natural Language Processing and Information Systems*, Elisabeth Métais, Farid Meziane, Helmut Horacek, and Epaminondas Kapetanios (Eds.). Springer International Publishing, Cham, 33–45.

[16] Vu Tran, Minh Le Nguyen, and Ken Satoh. 2019. Building Legal Case Retrieval Systems with Lexical Matching and Summarization Using A Pre-Trained Phrase Scoring Model. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law* (Montreal, QC, Canada) *(ICAIL '19)*. Association for Computing Machinery, New York, NY, USA, 275–282.

[17] Ivan Vulić, Wim De Smet, Jie Tang, and Marie-Francine Moens. 2015. Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications. *Information Processing & Management* 51, 1 (2015), 111–147. https://doi.org/10.1016/j.ipm.2014.08.003

[18] Jorge A. Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC Corpus: A New Open Resource for Brazilian Portuguese. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan. https://aclanthology.org/L18-1686

[19] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. *ACM Trans. Inf. Syst.* 28, 4, Article 20 (nov 2010), 38 pages.

[20] Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Chaojun Xiao, Zhiyuan Liu, and Maosong Sun. 2018. Legal Judgment Prediction via Topological Learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3540–3549. https://doi.org/10.18653/v1/D18-1390