Legal Knowledge and Information Systems
S. Villata et al. (Eds.)

doi:10.3233/FAIA200855

Topic Modelling Brazilian Supreme Court Lawsuits

Pedro Henrique LUZ DE ARAUJO [a,1] and Teófilo DE CAMPOS [a,2]
[a] *Department of Computer Science, University of Brasília, Brasília, DF, Brazil*

Abstract. The present work proposes the use of Latent Dirichlet Allocation to model Extraordinary Appeals received by Brazil's Supreme Court. The data consist of a *corpus* of 45,532 lawsuits manually annotated by the Court's experts with theme labels, a multi-class and multi-label classification task. We initially train models with 10 and 30 topics and analyze their semantics by examining each topic's most relevant words and their most representative texts, aiming to evaluate model interpretability and quality. We also train models with 30, 100, 300 and 1,000 topics, and quantitatively evaluate their potential using the topics to generate feature vectors for each appeal. These vectors are then used to train a lawsuit theme classifier. We compare traditional bag-of-words approaches (word counts and tf-idf values) with the topic-based text representation to assess topic relevancy. Our topics semantic analysis demonstrate that our models with 10 and 30 topics were capable of capturing some of the legal matters discussed by the Court. In addition, our experiments show that the model with 300 topics was the best text vectoriser and that the interpretable, low dimensional representations it generates achieve good classification results.

Keywords. topic models, legal domain, document analysis, Latent Dirichlet Allocation

1. Introduction

Brazil's court system suffers from an excessive amount of lawsuits [1]. About 80 million suits awaited judgement in 2017, which amounts to almost one for every three Brazilians. There was an increase of 19.4 million suits between 2009 and 2017. Furthermore, the average processing time reaches more than seven years in some cases. Such long waiting times negatively impact Brazil's legal certainty and brings about greater budgetary needs—Brazil spent R$ 90.7 billions in 2017 to maintain the judiciary, corresponding to about 28 billion[3] dollars [2].

Natural Language Processing (NLP) and Machine Leaning techniques can contribute to a quicker, cheaper and more efficient analysis of legal proceedings and as a result help promote greater effectiveness and democratization of justice. Some works already explore the use of artificial intelligence in the context of Brazil's courts [3,4,5].

[1]Corresponding Author: Pedro Henrique Luz de Araujo, UnB - Brasília, DF, Brazil; E-mail: pedro.luz@aluno.unb.br.
[2]Corresponding Author: Teófio Emidio de Campos, UnB - Brasília, DF, Brazil; E-mail: t.decampos@oxfordalumni.org.
[3]Considering average exchange rate of 2017: 3.19 reais to 1 dollar.

That being said, we are not aware of publications regarding the topic modelling of Brazilian lawsuits.

Topic models are a family of statistical models used to discover in an automatic and unsupervised manner themes (topics) present in a collection of documents [6]. The topics are obtained from the statistical analysis of the words that comprise the documents. Since annotations and labelling of documents are not needed, topic models enable the organisation, exploration and indexing of massive amounts of data in a scale that could be prohibitively expensive if human made. The trained models may also be used for downstream tasks such as sentiment analysis [7] and document classification [8]. In addition, the approach is not restricted to text data and may be used to model genomic data, images and social networks [6].

In this paper, we employ Latent Dirichlet Analysis (LDA) to model Extraordinary Appeals (*Recursos Extraordinários*—RE) received by Brazil's Supreme Court (*Supremo Tribunal Federal*—STF). Each suit has been manually annotated by the Court's employees to include information on its general repercussion (*repercussão geral*) themes. This is a multi-label classification task, which we will further discuss in Section 3. Our contributions are:

1. The qualitative analysis of the semantics of each topic from models with 10 and 30 topics trained on the STF data.
2. The quantitative analysis of topic relevance by using topic distribution vectors as input for general repercussion theme classification. We experiment with models of 10, 30, 100, 300 and 1,000 topics.

The rest of the paper is organized as follows. Section 2 briefly review Topic Model literature and NLP applied to the legal domain approaches. Sections 3 and 4 describe the dataset and the model employed, respectively. Section 5 reports our experiments and Section 6 presents and discusses the results. Section 7 concludes the paper.

2. Related Work

2.1. Topic Models

Topic models have been an area of research since 1990, when Deerwester et al. [9] proposed Latent Semantic Indexing (LSI). The method uses Singular Value Decomposition (SVD) to factorize a matrix of term-document co-occurrence values to construct a "semantic" space where terms and documents closely associated are near one another. The method is further explored by Hofmann [10], who introduced probabilistic LSI (PLSI). Like LSI, PLSI decomposes a co-occurrence matrix, but while the former uses a linear algebra approach, the latter method is statistical, modelling the document-word co-occurrence probability as a mixture of conditionally independent multinomial distributions. On the other hand, PLSI has some weaknesses, such as the linear growth of the parameters with the size of the corpus, which causes overfitting issues, and the lack of procedure to assign probability to a document not seen in the training set.

To overcome PLSI weaknesses, Blei et al. [11] proposed Latent Dirichlet Allocation (LDA). The authors show that LDA can be used for a range of tasks, such as document modelling, text classification and collaborative filtering, outperforming approaches based on unigrams and PLSI.

Since then, the study of extensions of LDA by relaxing some of its assumptions has been an active area of research [6]. For example, by relaxing the assumption that the order of the documents can be neglected, Blei and Lafferty [12] propose Dynamic Topic Models, capable of modelling the time evolution of topics in a corpus.

2.2. *Natural Language Processing and Topic Models in Legal Text*

Efforts have been made to apply Natural Language Processing and Machine Learning techniques to legal text. NLP has been used to automatically extract and classify relevant entities in court documents [13,14,4]. Other works [15,16,17,18] focus on using automatic summarization to reduce the amount of information legal professionals have to process. Document classification has been explored for decision prediction [19,20], area of legal practice attribution [21] and fine-grained legal-issue classification [22].

LDA has been employed to model legal corpora. Carter et al. [23] model documents from the Australian High Court; Remmits [24] models decisions from the Supreme Court of the Netherlands; O'Neill et al. [25] used LDA to explore British legislative texts.

Some works explore the processing of Brazilian legal documents. Correia da Silva et al. [3] use a CNN to classify STF's documents. De Vargas Feijó and Moreira [5] introduce a dataset for decision summarization. Luz de Araujo et al. [4] built a manually annotated corpus for named entity recognition and classification with legislation and legal decision classes. On the other hand, we are not aware of publications examining topic modelling of Brazilian legal corpora.

3. Data

We use the VICTOR dataset [26], a corpus containing 45,532 Extraordinary Appeals. Each instance is a legal proceeding as it is received by the STF, that is, before it is processed and judged. Each lawsuit is represented as an ordered sequence of pages containing text.

The dataset contains manual annotation that assigns to each lawsuit one or more general repercussion[4] themes. More specifically, the options are the 28 most important themes according to the STF, each one identified by a unique integer[5]; e.g., theme 6 deals with the State's duty to supply costly medications to citizens who suffer from serious diseases and are not able to buy them. The integer 0 identifies the instances that contain at least one theme that does not belong to any of those 28 classes. It follows that theme assignment is a multi-label classification task.

The data is divided into train/validation/test splits containing 70%/15%/15% of all suits, respectively. The theme distribution is the same in all splits as figure 1 shows.

The following preprocessing steps were applied to the raw text: lower-casing, removal of stop words and alphanumeric tokens, email and URL tokenization, and identification of simple law citations; e.g., we change *Lei* (law) 11.419 to LEI_11419.

[4] An appeal must have general repercussion to be judged by the STF. This means that lawsuit must relate to relevant economic, political, social or legal issues that exceed the interests of the parties.

[5] A list of all themes is available at http://www.stf.jus.br/portal/jurisprudenciaRepercussao/abrirTemasComRG.asp.

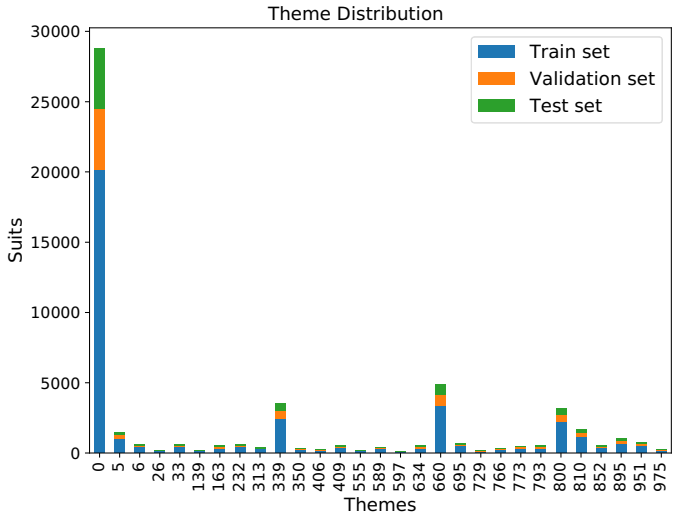

Figure 1. Theme counts.

4. Model

Inspired by previous attempts to model different kinds of legal text [23,24,25], we choose Latent Dirichlet Allocation [11] as the method for topic generation. LDA is a probabilistic generative model of a corpus, where each document is represented as a random mixture over latent topics. Each topic is in turn a distribution over words. That is, LDA assumes the following generative process for a corpus D of m documents of length n_i, $i \in [1,\ldots,m]$, assuming a fixed set of k topics:

1. $\boldsymbol{\theta}_i$, $i \in \{1,\ldots,m\}$, the topic distribution of document i, is chosen from a Dirichlet distribution Dir($\boldsymbol{\alpha}$)
2. $\boldsymbol{\phi}_j$, $j \in \{1,\ldots,k\}$, the word distribution of topic j, is chosen from a Dirichlet distribution Dir($\boldsymbol{\beta}$).
3. For each word position (i,j), $i \in \{1,\ldots,m\}$, $j \in \{1,\ldots,n_i\}$:
   (a) A topic $\mathbf{z}_{i,j} \sim$ Multinomial($\boldsymbol{\theta}_i$) is chosen.
   (b) A word $\mathbf{w}_{i,j} \sim$ Multinomial($\boldsymbol{\phi}_{\mathbf{z}_{i,j}}$) is chosen.

Given this generative assumption, the LDA procedure assigns: a topic distribution for each document, a topic for each word in each document and a word distribution for each topic.

5. Experiments

5.1. Model Training for Exploratory Analysis

We perform an exploratory analysis of the data aiming to understand its most relevant topics by training LDA models. We train two models on the training split of the data,

one with 10 topics and the other with 30. Since the whole data does not fit into memory, we use the algorithm proposed by [27] for the online training of LDA models, based on stochastic optimisation with gradient steps.

To select the most informative words, we restrict our vocabulary to the words that appear in at least 50 lawsuits of the training set and in no more than 50% of them. In addition, we filter words with only one letter, with the intuition that they probably do not help with topic interpretability. The obtained vocabulary contains 81,418 entries.

We use mini-batches of 4,096 suits, with a maximum number of 400 iterations per mini-batch, and train for 4 epochs. The hyper-parameters were chosen empirically and were sufficient for the convergence of most lawsuits in the training set.

5.2. Topic Distribution as Text Representation

In order to have a quantitative analysis of the detected topics, we use LDA as a lawsuit feature extractor; that is, the topic distribution of each lawsuit is used as its vector representation and fed to a classifier to predict general repercussion themes. We run experiments with models of 10, 30, 100, 300 and 1,000 topics, using eXtreme Gradient Boosting [28] (XGBoost) as the classifier.

We compare the topic representation with two traditional bag-of-words representations: i) Tf-idf values and ii) word counts. To establish a fair comparison, all models use the same vocabulary. Since we have a multi-label task, we employ a One-vs-All approach where we train a binary classifier for each theme and the final classification is the aggregation of all predictions. Formally, let C be the set of all themes, t a threshold value, $f_c(\cdot)$ the decision function of the classifier for class c, and l a lawsuit:

$$\forall c \in C, \text{assign } c \text{ to } l \text{ if } f_c(l) \geq t\,. \tag{1}$$

We set 0.5 as the threshold value.

Finally, we use the validation set to tune the following XGBoost hyperparameters through random search: number of trees, maximum depth and shrinkage factor.

All results are reported on the test set unless otherwise stated. As a baseline method we choose a classifier that assigns all themes to any input, which achieves a F1 score weighted by class frequency of 41.17% and an average F1 score of 5.48%.

6. Results

6.1. Topic Analysis

In order to evaluate the topic quality of the models with 10 and 30 topics we examine the most relevant words and lawsuits from each topic and assign it a label [29]. Table 1 presents the results of the labelling process. For each topic we show its four most relevant words, where relevance is defined [30] as

$$r(\mathbf{w},\mathbf{z}|\lambda) = \lambda \log P(\mathbf{w}|\mathbf{z}) + (1-\lambda)\log\frac{P(\mathbf{w}|\mathbf{z})}{P(\mathbf{w})}\,, \tag{2}$$

and the parameter λ $(0 \leq \lambda \leq 1)$ determines weight given to the probability of term $\mathbf{w}$ given topic $\mathbf{z}$ relative to the ratio between that probability and the marginal probability

of **w** on the whole corpus. For each topic, through manual inspection, we select the value with the most descriptive top words, which have been translated to English, except in the case of acronyms and names, which are shown in italic.

Table 1. Topic labels and their respective four most relevant words (10 topics).

Topic	λ	Assigned label	Words
1	0.6	Public servant remuneration	servants, servant, limitation, remuneration
2	0	Criminal Law	narcotic, hydrometer, clandestine, interrogation
3	0.6	Pension Law	benefit, event, retirement, pension
4	0.6	Civil Law	bank, contract, consumer, *projudi*
5	0.6	Right to health	health, city, municipal, medication
6	0.4	OCR errors	*ento*, no, *ro*, *co*
7	0.6	Tax Law	*icms*, *ipi*, tax, income
8	0	Entities	*econorte*, *rcte*, *pieter*
9	0.4	Labor Law	*fgts*, *pss*, hours, payroll
10	0.6	Document access	original, site, access, report

Regarding the model with 10 topics, the results show that most topics are identified with legal matters routinely discussed by the STF. That being said, topics 6 and 8 were challenging to label. The lawsuits with the highest proportion of these topics were useful in that enterprise.

In the first case, the most representative lawsuits were found to contain a great amount of OCR noise. The most relevant suit, with 99.99957% topic 6 content, contains the following passage: "r cm emoi oit incm m t i o i m cofl inoioem oulfl tofl cmcmh co ffl ffl ffl a z a z ffl o t a o u ffl otoidtoaz d to a i o tn ffl em cmcocoulococm eo cocm [...]", which is pure gibberish.

While examining topic 8, we discovered that its most representative lawsuits contained a lot of named entities; e.g., from the 15 most frequent words in the suit with most topic 8 content, 8 referred to people or organisations.

The model with 30 topics, as shown in Table 2, was also able to identify interpretable topics, many of them directly related to legal matters discussed by the Court. To label each topic, we once again analyze its most relevant words from each topic while varying the value of λ. To label the most challenging topics we also examine their most representative lawsuits. Due to the greater number of topics, some of them deal with much more specific matters than in the case of the model with 10 topics. For example, while the model with fewer topics has only one generic topic for Tax Law, the one with 30 topics has four different topics related to different facets of that legal area (topics 3, 25, 27 and 28).

That said, some of the topics have relevant words that do not belong to related matters. Topic 19, for example, assign high probabilities to words related to both Consumer Law and the Brazilian state of Bahia, with mentions to cities such as Bahia's capital city Salvador. On the other hand, there are topics with very specific relevant words, such as topic 20, that groups names of people. These results can be explained by the nature of the data, which combines various types of documents; e. g. petitions, judgments, orders, proxy statements, certificates, and other supporting documents. We expect that by training only on the Court's rulings the topics would be even more related to specific legal matters discusses by the Justices.

Table 2. Topic labels and their respective four most relevant words (30 topics).

Topic	λ	Assigned label	Words
1	0.6	Civil liability	damage, damages, compensation, non-material
2	0.22	Expiration of social security benefit	benefit, expiration, limit, social security (*previdenciário*)
3	0.6	Tax Law	treasury, tax, revenue, taxation
4	0.1	Miscellaneous - Legal vocabulary, enttities and laws	serial number, *pet*, stamp, *itaperuna*
5	0.4	Public servant bonus	bonus, performance, inactive, evaluation
6	0.4	Rural social security	rural, contribution, LEI_8212, pension
7	0.6	Public servant remuneration readjustment	readjustment, servants, remuneration, *urv*
8	0.4	OCR errors	*ento*, no, *ro*, *ffl*
9	0.6	Members of the military	military, servant, servicemen, servants
10	0	Criminal Law	clandestine, *sepetiba*, semi-open, narcotic
11	0.4	Contract law	contract, contracts, fee, accounts
12	0.05	Technical Councils	*confea*, *crea*, agronomy, LEI_6496
13	0.2	Public tender	tender, candidate, notice, openings
14	0.4	Anticipation of remuneration readjustment	*upag*, *pccs*, labor, LEI_8460
15	0.6	Right to health	health, medication (plural), treatment, medication (singular)
16	0.9	Savings account, interest and monetary correction	correction, monetary, savings account, delay
17	0.6	Document access	original, site, acesse, report
18	0.6	labor complaints	*estran*, *tst*, entity, claimant
19	0.4	Miscellaneous - Consumer Law and Bahia (Brazilian state)	consumer, *salvador*, *bahia*, *pdf*
20	0	Entities - names	*lauxen*, *tainá*, *heloise*, *soeli*
21	0.7	Qualification	*num*, normal, internment, *foz*
22	0.5	insurance	insurance, *previd*, institute, *dpu*
23	0.4	Payroll	hours, *fgts*, payroll, overtime
24	0	Miscellaneous - Organisations, charters and non-Portuguese words	*andaterra*, *peixer*, funds, market
25	0.5	Fiscal documents	*ltda*, *ipi*, *nfe*, *icms*
26	0.4	Rio Grande do Sul (Brazilian state)	*sul* , *grande*, *alegre*, *paese*
27	0.4	Income tax	updated, months, *rra*, *irpf*
28	0.2	Tax Law - circulation of goods	compatible, *issqn*, exit, *eireli*
29	0.2	Miscellaneous - Procedure and Paraná (Brazilian state)	*paraná*, *arq*, *curitiba*, *mov*
30	0.4	Payments	*jam*, *vlr*, received, credit

6.2. Quantitative Analysis

Figure 2 compares the performance on the validation set of classifiers trained on text features obtained from models with 10, 30, 100, 300 and 1,000 topics. All models greatly outperformed a baseline that simply assigns all themes to each instance. Increasing the dimensionality of the representation up to 300 topics improves performance. The model with 1,000 topics, on the other hand, is comparable to the one with 300.

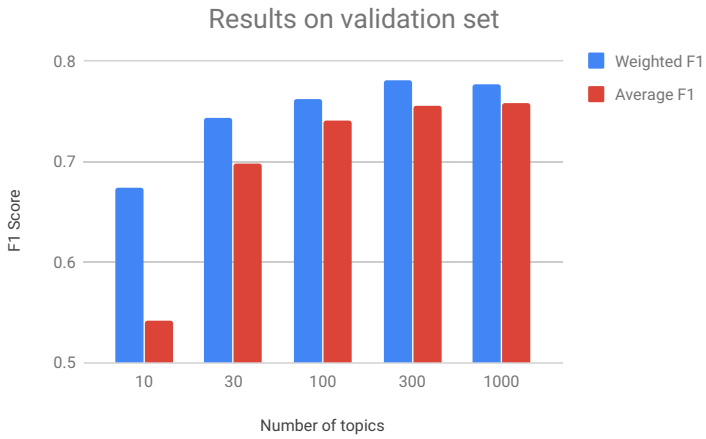

Figure 2. Validation set performance of classifiers trained with different numbers of topics.

Table 3 compares the 300-dimensional lawsuit representation with the word counts and tf-idf values bag-of-words representations on the test set. The topic distribution representation did not outperform the traditional methods, but achieved good performance—much better than the baseline that assigns all themes. These results suggest that the detected topics are related to the themes relevant to the Court and have the potential to aid the judiciary with the management of cases.

Furthermore, it has an advantage over the traditional approaches with respect to the dimensionality of the representation—it describes a lawsuit using 300 dimensions instead of 81,418, a relative reduction of 99.63%. As a result, the training and inference is much faster.

Table 3. F1 scores (in %) on the test set of each text representation method. Assigning all themes to all samples yield a weighted (by class frequency) F1 score of 41.17 and an average F1 score of 5.48.

	Word counts	Tf-idf	300 topics
Weighted	**89.29**	89.22	78.07
Average	87.54	**88.37**	75.81

7. Conclusion

We proposed the use of Latent Dirichlet allocation to build topic models of Extraordinary Appeals from Brazil's Supreme Court (STF). We labelled and analysed the models with 10 and 30 topics, showing the correspondence between them and legal matters that reach the Court. We used the obtained topic distribution vectors as input for a supervised multi-label classification task in order to establish a quantitative analysis of topic relevance. The topic distribution representation, with an optimal value of 300 topics, achieved good results using much lower dimensionality than the traditional methods. The technique can be leveraged to help organize, explore and extract information of the massive amounts of data that reach the Court.

8. Acknowledgements

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. TdC received support from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), grant PQ 314154/2018-3. We acknowledge the support of "Projeto de Pesquisa & Desenvolvimento de aprendizado de máquina (machine learning) sobre dados judiciais das repercussões gerais do Supremo Tribunal Federal - STF". We are also grateful for the support from Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF, project KnEDLe, convênio 07/2019) and Fundação de Empreendimentos Científicos e Tecnológicos (Finatec).

References

[1] de Cássia Carvalho Lopes R. migalhas com, editor. Eventual Influences of Common Law on the Brazilian Legal System; 2017. Available at https://www.migalhas.com/HotTopics/63,MI255372,51045-Eventual+Influences+of+Common+Law+on+the+Brazilian+Legal+System Available from: https://www.migalhas.com/HotTopics/63,MI255372,51045-Eventual+Influences+of+Common+Law+on+the+Brazilian+Legal+System.

[2] Secretaria de Comunicação Social do Conselho Nacional de Justiça. CNJ, editor. Sumário Executivo do Relatório Justiça em Números; 2018. Available at http://www.cnj.jus.br/files/conteudo/arquivo/2018/09/da64a36ddee693ddf735b9ec03319e84.pdf.

[3] da Silva NC, Braz FA, de Campos TE, Gusmao DB, Chaves FB, Mendes DB, et al. Document type classification for Brazil's supreme court using a Convolutional Neural Network. In: 10th International Conference on Forensic Computer Science and Cyber Law (ICoFCS). Sao Paulo, Brazil; 2018. Winner of the best paper award.

[4] Luz de Araujo PH, de Campos TE, de Oliveira RRR, Stauffer M, Couto S, Bermejo P. LeNER-Br: a Dataset for Named Entity Recognition in Brazilian Legal Text. In: International Conference on the Computational Processing of Portuguese (PROPOR). Lecture Notes on Computer Science (LNCS). Canela, RS, Brazil: Springer; 2018. p. 313–323. Available from: https://cic.unb.br/~teodecampos/LeNER-Br/.

[5] de Vargas Feijó D, Moreira VP. RulingBR: A Summarization Dataset for Legal Texts. In: Villavicencio A, Moreira V, Abad A, Caseli H, Gamallo P, Ramisch C, et al., editors. Computational Processing of the Portuguese Language. Cham: Springer International Publishing; 2018. p. 255–264.

[6] Blei DM. Probabilistic Topic Models. Commun ACM. 2012 Apr;55(4):77–84. Available from: http://doi.acm.org/10.1145/2133806.2133826.

[7] Mauá DD. Modelos de tópicos na classificação automática de resenhas de usuários. [Master thesis]. Escola Politécnica da Universidade de São Paulo; 2009.

[8] Rubin TN, Chambers A, Smyth P, Steyvers M. Statistical topic models for multi-label document classification. Machine Learning. 2012 Jul;88(1):157–208. Available from: https://doi.org/10.1007/s10994-011-5272-5.

[9] Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R. Indexing by latent semantic analysis. Journal of The American Society for Information Science. 1990;41(6):391–407.

[10] Hofmann T. Probabilistic Latent Semantic Indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR. New York, NY, USA: ACM; 1999. p. 50–57. Available from: http://doi.acm.org/10.1145/312624.312649.

[11] Blei DM, Ng AY, Jordan MI. Latent Dirichlet Allocation. Journal of Machine Learning Research. 2003 Mar;3:993–1022. Available from: http://dl.acm.org/citation.cfm?id=944919.944937.

[12] Blei DM, Lafferty JD. Dynamic Topic Models. In: Proceedings of the 23rd International Conference on Machine Learning. ICML. New York, NY, USA: ACM; 2006. p. 113–120. Available from: http://doi.acm.org/10.1145/1143844.1143859.

[13] Dozier C, Kondadadi R, Light M, Vachher A, Veeramachaneni S, Wudali R. Named entity recognition and resolution in legal text. In: Semantic Processing of Legal Texts. Springer; 2010. p. 27–43.

[14] Cardellino C, Teruel M, Alonso Alemany L, Villata S. A Low-cost, High-coverage Legal Named Entity Recognizer, Classifier and Linker. In: Proceedints of the 16th International Conference on Artificial Intelligence and Law (ICAIL). London, United Kingdom; 2017. Preprint available from https://hal.archives-ouvertes.fr/hal-01541446.

[15] Kanapala A, Pal S, Pamula R. Text summarization from legal documents: a survey. Artificial Intelligence Review. 2017 Jun;Available from: https://doi.org/10.1007/s10462-017-9566-2.

[16] Galgani F, Compton P, Hoffmann A. Combining Different Summarization Techniques for Legal Text. In: Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data. HYBRID. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012. p. 115–123. Available from: http://dl.acm.org/citation.cfm?id=2388632.2388647.

[17] Kumar R, Raghuveer K. Legal document summarization using latent dirichlet allocation. International Journal of Computer Science and Telecommunications. 2012;3:114–117.

[18] Kim MY, Xu Y, Goebel R. Summarization of Legal Texts with High Cohesion and Automatic Compression Rate. In: New frontiers in artificial intelligence. Springer; 2013. .

[19] Aletras N, Tsarapatsanis D, Preotiuc-Pietro D, Lampos V. Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective. PeerJ in Computer Science. 2016 10;.

[20] Katz DM, Bommarito I Michael J, Blackman J. Predicting the Behavior of the Supreme Court of the United States: A General Approach. arXiv e-prints. 2014 Jul;p. arXiv:1407.6333.

[21] Şulea OM, Zampieri M, Vela M, van Genabith J. Predicting the Law Area and Decisions of French Supreme Court Cases. In: Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP. INCOMA Ltd.; 2017. p. 716–722. Available from: https://doi.org/10.26615/978-954-452-049-6_092.

[22] Undavia S, Meyers A, Ortega JE. A Comparative Study of Classifying Legal Documents with Neural Networks. In: Federated Conference on Computer Science and Information Systems (FedCSIS); 2018. p. 515–522.

[23] Carter DJ, Brown J, Rahmani A. Reading the high court at a distance: Topic modelling the legal subject matter and judicial activity of the high court of Australia, 1903-2015. UNSWLJ. 2016;39:1300.

[24] Remmits Y. Finding the Topics of Case Law: Latent Dirichlet Allocation on Supreme Court Decisions [Bachelor's Thesis]; 2017. Bachelor's thesis, Radboud University, July 2017.

[25] O'Neill J, Robin C, O'Brien L, Buitelaar P. An analysis of topic modelling for legislative texts. In: Proceedings of the Second Workshop on Automated Semantic Analysis of Information in Legal Texts; 2016. .

[26] Luz de Araujo PH, de Campos TE, Braz FA, Correia da Silva N. Victor: a dataset for Brazilian legal documents classification. In: Proceedings of The 12th Language Resources and Evaluation Conference (LREC). Marseille, France: European Language Resources Association; 2020. p. 1449–1458. Source code, dataset and further information available from https://cic.unb.br/~teodecampos/ViP/lrec/. Available from: https://www.aclweb.org/anthology/2020.lrec-1.181.

[27] Hoffman MD, Blei DM, Bach F. Online Learning for Latent Dirichlet Allocation. In: Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1. NIPS. USA: Curran Associates Inc.; 2010. p. 856–864. Available from: http://dl.acm.org/citation.cfm?id=2997189.2997285.

[28] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD. New York, NY, USA: ACM; 2016. p. 785–794. Available from: http://doi.acm.org/10.1145/2939672.2939785.

[29] Grimmer J, Stewart BM. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. Political Analysis. 2013;21(3):267–297.

[30] Sievert C, Shirley K. LDAvis: A method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces. Baltimore, Maryland, USA: Association for Computational Linguistics; 2014. p. 63–70. Available from: https://www.aclweb.org/anthology/W14-3110.