Legal Knowledge and Information Systems
E. Francesconi et al. (Eds.)

doi:10.3233/FAIA220454

Why Do Tenants Sue Their Landlords? Answers from a Topic Model

Olivier SALAÜN [a,1], Fabrizio GOTTI [a], Philippe LANGLAIS [a] and Karim BENYEKHLEF [b]
[a] *RALI, DIRO, Université de Montréal*
[b] *Cyberjustice Laboratory, Faculty of Law, Université de Montréal*

Abstract. Topic modeling is widely used in various domains for extracting latent topics underlying large corpora, including judicial texts. In the latter, topics tend to be made by and for domain experts, but remain unintelligible for laymen. In the framework of housing law court decisions in French which mixes abstract legal terminology with real-life situations described in common language, similarly to [1], we aim at identifying different situations that can cause a tenant to prosecute their landlord in court with the application of topic models. Upon quantitative evaluation, LDA and BERTopic deliver the best results, but a closer manual analysis reveals that the second embedding-based approach is much better at producing and even uncovering topics that describe a tenant's real-life issues and situations.

Keywords. topic modeling, court decisions, French language, housing law, knowledge extraction

1. Context

Topic modeling is an application of natural language processing (NLP) widely used for summarizing large corpora into different clusters of terms. These terms describe latent topics present in the text but not immediately visible to the reader. In this work, we aim at applying and evaluating topic models to a corpus of court decisions in French from the Tribunal administratif du logement (TAL, Housing law tribunal) in Canada. This court deals exclusively with all disputes involving landlords and tenants bound by an accommodation lease contract. These litigations are mostly motivated by rent arrears or substandard housing. An analysis was performed by [1] over cases in which a tenant was claiming damages from their landlord before the TAL. The objective was to find which were the concrete factors (e.g. water/electricity access issue, bedbugs, lack of maintenance) that caused the judge to accept the tenant's claims and award them damages. The facts and evidence of tenant claims are judged in the light of articles 1854, 1864 and 1910 from Civil Code of Québec[2] that define landlord contractual obligations to provide an accommodation *in a good state of repair* and to ensure both *peaceable enjoyment* and *good state of habitability*. These articles provide general abstract legal concepts without making an exhaustive list of concrete criteria. This is left to case-by-case interpretation

[1] Corresponding Author: Olivier Salaün, salaunol@iro.umontreal.ca
[2] Articles 1854, 1864 and 1910 can be read at https://canlii.ca/t/55g4j

by the judge, hence [1]'s initiative of manually annotating 149 cases which ultimately yielded 44 *factors* (some are shown in the top left blue box of Figure 2). We shall call these factors **reference topics** (RTs) here. [1]'s manual topic extraction is costly (a legal expertise is required, on long documents) and only covers a tiny portion of the several hundred thousand cases. In this work, we intend to apply topic models as an attempt to automate such examination of cases and to design methods able to isolate relevant topics with respect to the housing law domain.

2. Related Work

Topic modeling is used in a wide variety of domains, such as social networks analysis [2], scientific papers [3] or medical data [4] for instance. Some may even call it *distant reading* [5] as it consists in applying approaches, such as LDA (explained later), to extract thematic representations of large corpora.

2.1. Topic Modeling for the Legal Domain

For legal practitioners, topic modeling also provides useful unsupervised methods for soft clustering/categorization of legal documents, without having a prior classification scheme [6]. Such methods extract **topics** that can be described as collections of words clustered together based on their distribution across the documents. In the legal field, topic models were applied for instance to UK legislative documents [7], Latvian legal acts [8] and court decisions from Australia [9], the Netherlands [10], Brazil [11,12] and the United States [13]. We must emphasize that these topic modeling experiments usually yielded topics meant for experts in the field. However, a certain language gap exists between the specialized legal terminology used by judges, and laymen's generic language as shown by [14]: They describe the same reality in different terms. This discrepancy exists for housing law, and we posit that extracted topics can help bridge this gap by illuminating a *taxonomy of practice* [9] i.e. real-world situations cues discussed in the scope of legal abstract concepts. Such a resource would be useful for laymen seeking legal information, for instance when trying to connect concrete problems (e.g. *water leakage*) to relevant legal concepts (e.g. *good state of habitability*).

2.2. Challenges about Topic Evaluation

Despite readily available toolkits facilitating topic modeling, evaluating topics is still an open question, in part because the assessment differs depending on the data and the legal area. A possible strategy is an extrinsic evaluation of the topics, for instance measuring the improvement topics bring when carrying out text classification [11,12]. Such a protocol obviously requires the cases to be manually classified beforehand (this is not our case). Intrinsic evaluation of topics can also be manually assessed by comparing topics automatically assigned to a document with its original ground truth category [13]. When no such categories are available, typical manual methods include ordinal three-point Likert-scales and intrusion tests [15,10]. Finally, a common automated metric for topic evaluation is topic coherence [16,8] based on word co-occurrences from an external reference corpus.

3. Data Description and Preprocessing

The dataset manually examined by [1] consisted only of 149 cases from 2017. We extended this to 12,102 cases spanning 2001 to 2018 that included clearly separated sections described by [17]: facts (rich with real-life situations described by laymen) and legal analysis (rich with legal terminology). Clear section boundaries are necessary because, unlike works from Section 2 and following [10], we do not carry out topic modeling on the entire document but only on the facts section so that our topics contain as little legal terminology as possible and more real-life situations. As in [1]'s work, we retained cases citing articles 1854, 1864 or 1910 from the Civil Code of Québec in which the tenant is the applicant, thus amounting to 1,381 cases. Since each case can contain several litigation factors, topic modeling is done at the paragraph level. Our resulting dataset of 34,685 paragraphs is processed in a standardized fashion with a SpaCy tokenizer: We remove dates, monies, digits, symbols, French- and law-specific stopwords, then filter tokens with specific part-of-speech tags, lemmatize and lowercase them before merging bigram collocations (e.g. *hot_water*). We remove paragraphs with less than 5 terms, yielding 26,815 paragraphs.

4. Models

We selected two traditional and one neural topic models, respectively: Latent Semantic Indexing (LSI, also known as Latent Semantic Analysis) [18], Latent Dirichlet Allocation (LDA) [19] and BERTopic [20]. When conducting training for each of these models, the number of topics (a hyperparameter) is set beforehand to 50, 100, and 200.

4.1. Latent Semantic Indexing (LSI)

LSI relies on singular value decomposition (SVD) of a sparse $paragraphs \times words$ ($P \times W$) matrix M in which each value $v_{p,w}$ represents the term frequency-inverse paragraph frequency (TF-I*D*F) weight of word w for paragraph p.$v_{p,w}$ increases with the frequency of w in p but decreases if w is widespread among paragraphs. The SVD of M produces 3 matrices M', M'' and M''' of respective shapes : $paragraphs \times topics$ ($P \times T$), $T \times T$ and $T \times W$. The last matrix provides word distribution to each topic. We use Gensim library [21] for training LSI models and set the number of power iteration steps to 100 for improving the accuracy of the SVD approximation with large sparse matrices.

4.2. Latent Dirichlet Allocation (LDA)

LDA is a generative stochastic model that aims at recreating the original corpus through a pseudo-corpus generation. For a preset number of topics, it generates a collection of pseudo-paragraphs whose word and topic distributions approximate as closely as possible those of the real dataset. Paragraphs are considered as random mixtures over latent topics, and topics as distributions over words. $Dir(\alpha)$ is the Dirichlet distribution of topics over paragraphs while $Dir(\beta)$ is the Dirichlet one of words over topics. $Dir(\alpha)$ is the prior for multinomial topic distribution θ_p for paragraph p while $Dir(\beta)$ is the prior for multinomial word distribution ϕ_k for topic k. As shown on Figure 1, at position i of pseudo-paragraph p of pseudo-corpus C, word $w_{p,i}$ is defined by both θ_p that defines the

Table 1. Top 10 terms from a randomly chosen topic for each model translated from French (predefined number of topics: 100)

LSI : door window room last place repair landlord day problem floor
LDA : party owner infiltration estimate dwelling concern heat finish list receive_notification
BERTopic : building manager management company caretaker witness son occupies responsible viner

topic $z_{p,i}$ at i in p, and by ϕ_k. After random initialization of these distributions and several passes over the documents, LDA is able to identify a steady collection of salient words for each topic k. We again use the Gensim library and set the number of passes at 100.

4.3. BERTopic

BERTopic relies on context-based representation derived from transformer [22,23] embeddings, thus allowing the model to access semantic information. First, the paragraphs are encoded with sentence-BERT embeddings [24] that are suitable for paraphrase detection and clustering. In our case, we chose a multilingual (over 50 languages) model[3] [25] for embedding French paragraphs. Next, these representations are dimensionally reduced with UMAP [26] before being passed to HDBSCAN [27] for soft and hierarchical clustering. A cluster contains paragraphs that are assumed to relate to the same topic. Said topic is then represented by a collection of the most salient words contained in its paragraphs through TF-IDF measures. After observing topics such as those in Table 1, we decided to retain only the top 5 words for each candidate topic during evaluation as the remaining terms are less topic-representative and noisier.

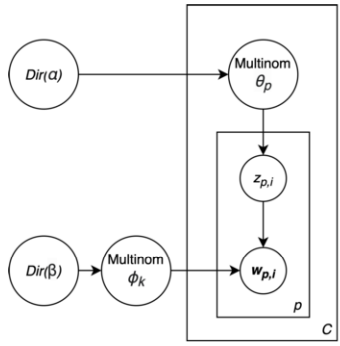

Figure 1. Graphical model representation of LDA generative process of word $w_{p,i}$ at position i of pseudo-paragraph p in pseudo-corpus C.

5. Quantitative Automated Evaluation

As discussed in Section 2.2, topic evaluation is challenging. In our case, despite the lack of text classification labels that would allow an extrinsic evaluation, we focus on two possible automatic approaches. The first one consists in comparing the **candidate topics** (CT) with the 44 **reference topics** (RT) manually identified by [1]. The second one relies on commonly used automatic topic coherence metrics.

5.1. Automated Evaluation with Respect to Domain-Specific Reference Topics

Comparing candidate topics (CTs) with [1]'s reference topics (RTs) addresses the question of whether a topic model can identify these RTs. An automated pairwise comparison

[3]The pretrained sentence-transformer model we used is available at: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

approach is illustrated in Figure 2. For each pair (CT_i, RT_j), a score S is computed yielding a $|CT| \times |RT|$ score matrix from which we retain the top $|CT|$ scores and the corresponding (CT_i, RT_j) pairs. From these pairs, we count the number d of **d**istinct matched RTs. The recall and precision are obtained by dividing d by $|RT|$ and $|CT|$, respectively. Although these metrics are not perfect, they are a necessary compromise, since the RTs identified by [1] only cover a tiny portion of all decisions.

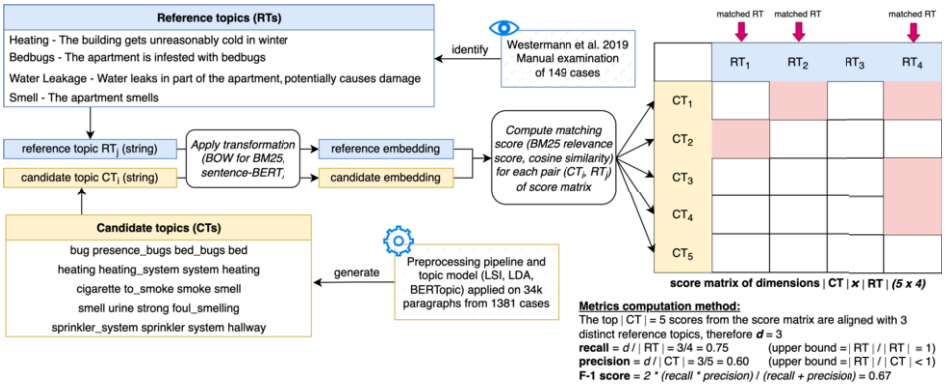

Figure 2. A toy example where 5 CTs are compared with 4 RTs. All topics are taken from actual [1]'s reference topics and candidate topics translated from French for illustration purposes.

Computing the similarity score S is delicate because RTs are short sentence-like descriptions while a CT is a sequence containing the 5 terms most representative of the topic. Two approaches are used for computing similarity: **1)** For each pair (CT_i, RT_j), each string is encoded with a multilingual sentence-transformer embedder[4] and the cosine similarity between the two is used as a matching score. **2)** We use an Okapi BM25 [28] approach in which RT and CT are queries and documents, respectively. Scores correspond to BM25 bag-of-words-based proximity scores assigned to CT w.r.t. to RT.

5.2. *Topic Coherence: Evaluation with Respect to an External Reference Corpus*

Topic quality is commonly measured with topic **c**oherence metrics, in particular normalized pointwise information (NPMI) shown by [29] to be positively correlated with human judgment. A topic CT_t gets a high **c**_NPMI score, shown in Eq. 1, if its N top terms have high pairwise joint probabilities.

$$c\_NPMI(CT_t) = \sum_{j=2}^{N} \sum_{i=1}^{j-1} \frac{log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-logP(w_i,w_j)} \tag{1}$$

Joint probabilities $P(w_i, w_j)$ are computed on the basis of a large corpus, usually Wikipedia [16], with a 10-word co-occurrence window. We extracted and preprocessed (see Section 3) a French Wikipedia snapshot dated 1 September 2022 (2.4 million articles) and a corpus of housing law decisions (531k cases from Tribunal administratif du logement shown in Table 2 as TAL) as generic and domain-specific reference corpora,

[4] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

Table 2. Scores per model and per number of topics in terms of CTs-RTs similarity and topic coherence NPMI. Highest and second highest values are in bold and underlined respectively for each metric.

Model and number of topics		BM25 Proximity score			SBERT Cosine Similarity			c_NPMI	
		R	P	F1	R	P	F1	Wiki	TAL
LSI	50	0.273	0.240	0.255	0.409	0.360	0.383	0.0090	0.0504
	100	0.409	0.180	0.250	0.591	0.260	0.361	-0.0008	0.0512
	200	0.545	0.120	0.197	0.659	0.145	0.238	-0.0058	0.0361
LDA	50	0.568	**0.500**	**0.532**	0.523	0.460	0.489	0.0215	0.0780
	100	0.614	0.270	0.375	0.705	0.310	0.431	0.0035	0.0607
	200	**0.636**	0.140	0.230	0.818	0.180	0.295	-0.0181	0.0412
BERTopic	50	0.545	0.480	0.511	0.614	**0.540**	**0.574**	**0.1462**	**0.3087**
	100	0.591	0.260	0.361	0.818	0.360	0.500	0.1266	0.2368
	200	0.614	0.135	0.221	**0.909**	0.200	0.328	0.0830	0.1852

respectively. The resulting c_NPMI scores range from -1 to 1. -1 implies the complete absence of co-occurrence of a topic pair of words within the reference corpus while 1 means complete co-occurrence.

5.3. Results

As shown in Table 2, LDA and BERTopic overall outperform LSI across all metrics. For a given model, raising the preset number of generated topics improves BM25 and SBERT recalls, which is expected. Overall, for a fixed number of topics, considering precisions and F1 measures, LDA outperforms BERTopic in terms of BM25 proximity scores while BERTopic achieves the highest performance with SBERT Cosine Similarities. This could be explained by the fact that LDA and BERTopic are respectively bag-of-words and embedding-based models. Concerning c_NPMI scores and regardless of the reference corpus, LSI and LDA get scores close to 0, the latter slightly outperforming the former, while BERTopic achieves the highest scores. This suggests that terms in LSI and LDA topics co-occur by chance under independent distributions while terms clustered by BERTopic co-occur beyond chance [30]. Given that both BM25 and SBERT recall scores for LDA and BERTopic increase with the number of topics, and since we want to assess to what extent topic models can help in identifying housing law use cases, we decided to manually evaluate the set of 200 topics generated by each of these two models.

6. Qualitative Human Evaluation

6.1. Intrinsic Evaluation of Candidate Topics Relevance

In order to assess the quality and intelligibility of CTs for laymen, two non-legal experts (co-authors of this work) were asked to evaluate the top 5 terms of a total of 400 topics from LDA and BERTopic models (each yielded 200 topics). These topics were shown in random order to evaluators who had no information on the models that produced them. For each topic, evaluators were asked whether they were able to identify an issue or a situation that would concern a tenant. When this was the case, the annotators were further asked to succinctly describe the theme they detected (e.g. *mold*, *disagreement on rent*).

Table 3. Selected candidate topics qualified as relevant by evaluators and that match reference topics from [1] (translated from French). CQ2s come from BERTopic except when in italics (generated by LDA).

Reference topic	Number of matching CQ2s	CQ1s	Examples of CQ2s
Water Leakage	9	10	water hot_water water_damage water_pressure occur water_infiltration infiltration occur roof roofing *water_infiltration garage occur complete use* *let water_damage believe place material*
Noise	8	10	noise jackhammer excessive_noise saw noise_arise music loud_music excessive_neighbourhood party play noise child child_run disturb play subject_soundproofing complaint_unbelievable unit_verify
Bedbugs	6	6	exterminator treatment extermination bedbug proceed_treatment insomnia stress sleep bug_bite phobia *bug mattress infestation deliver kitchen_floor*
Heating	5	8	heating temperature thermostat degree furnace cold temperature winter october heating dismantling_heater close_start outdated_problem heater heating_system
Exterior Issues	5	9	balcony rear_balcony antenna rot banister staircase staircase_lead hand_rail stair_step solidly access_terrace access rear give_access lock level_elevator refreshment_put sprinkler_system corridor_need

Relevant CTs are reported in Tables 3 and 4. Overall, Cohen's kappa score[5] for inter-annotator agreement amounts to 0.562 for all topics, 0.386 for the 200 LDA topics and 0.649 for the 200 BERTtopic ones. The main difference between evaluators resides in the fact that one only considered material issues as relevant topics but dismissed several issues that could be induced by people (e.g. harassment, violence, intrusion), though these actually account for housing law litigation factors. If we put aside the 14 CTs related to interpersonal issues, the aforementioned kappa scores increase to respectively 0.619, 0.440 and 0.706. Such a difference in inter-annotator agreement between LDA and BERTopic can be explained by the fact that LDA bag-of-words topics were harder to interpret as they gather terms that make less sense together in comparison to BERTopic. Such a low score for LDA is consistent with its c_NPMI scores from Table 2 and with the fact that CTs defined by both annotators as relevant topics amount to 10.5% and 33.0% for LDA and BERTopic, respectively.

6.2. Qualitative Analysis of Candidate Topics Evaluated as Relevant

After the identification of actually relevant CTs by non-experts, a domain expert (co-author of this work) manually paired these CTs to [1]'s RTs. For convenience, CTs **q**ualified as relevant by at least one and by exactly both evaluators are named C**Q**1s and C**Q**2s, respectively. The number of distinct RTs that could be paired to CQ1s and CQ2s amount respectively to 28 and 22 out of 44. The top 5 RTs with the more matching CQ1s

[5] According to [31], Cohen suggested kappa scores to be interpreted as fair, moderate and substantial agreement for values in the respective ranges $0.21 - 0.40$, $0.41 - 0.60$, and $0.61 - 0.80$

Table 4. Examples of candidate topics qualified as relevant by non-expert evaluators that correspond to topics not included in [1]'s reference (translated from French). CQ2s come from BERTopic except when in italics (generated by LDA).

Uncovered topic	Number of matching CQ2s	CQ1s	Examples of CQ2s
Plumbing	5	7	*affect plumbing hot lacking finishing*
			plumber plumbing drain valve batur
			kitchen repair_faucet washbasin_room water meal
			faucet noise water trickle_water adjust_definitely
Air quality	4	4	asthma symptom suffer doctor nose
			ventilation ventilation_system air exhaust duct
			allergy allergic mélabo test respiratory_problem
Internet access	3	3	telephone phone_number internet_service call telephone_line
			cable_origin optic_upgrade get_fibre bell_videotron hole_made
			videotron cable panel technician cable_television
Lighting	1	1	light lighting break_height burnt_pole lighting_deficient
Disabled accessibility	1	1	person_with_disability redo_june intercom_ramp hall_entrance autumn

and CQ2s are shown in Table 3. When pairing RTs and CQ2s, we noticed that RTs could be abstract and vague while CQ2s helped in bringing more nuance and precision by pinpointing precise themes. For instance, topic modeling allows extracting different noise-related issues such as construction (*jackhammer*), *loud_music*, *child*[ren]_*run*[ning] and *soundproofing*. The benefits of topic modeling are even more noticeable for *Exterior issues* by naming precise elements: *rear_balcony*, *staircase*, *hand_rail*, *access_terrace*, *level_elevator*. For relevant topics that could not be paired with RTs, the domain expert created new labels: 33 for CQ1s and 12 for CQ2s. This allowed uncovering new *litigation factors* that were not included in [1]'s RTs and that are shown in Table 4. For instance, several CQ2s relate to plumbing issues without involving water leakage. Other CQ2s, despite their small number, pinpoint sensitive issues such as air quality and accessibility for the disabled.

7. Discussion

Overall, relevant CTs are more likely to be obtained with BERTopic rather than with LDA. One explanation is that unlike most works described in Section 2 that dealt with documents from different legal areas [11,12,13], our corpus of paragraphs is much more homogeneous as it is only related to housing law, hence making topic modeling more difficult. Consequently, the bag-of-words approach of LDA gives less relevant topics compared to BERTopic, which has access to word semantic information. We also noticed that, when increasing the number of output topics, LDA was more likely to produce repeated noisy meaningless topics such as *berat blood applicances best pilule* (sic) reported by evaluators. A tentative explanation is that setting a very high number of topics can cause the LDA model to manufacture topics from noisy words. Such an issue was not observed with topics obtained from BERTopic.

Furthermore, the ratio of relevant CQ2s and CQ1s only covers a minority of all CTs, respectively 21.7% and 41.2%. CQ2s cover 10.5% and 33.0% of CTs by LDA and

BERTopic. These figures amount to 32.0% and 50.5% for CQ1s. A tentative explanation is that although input paragraphs are extracted from the facts section of court decisions, some legal jargon phrases still persist in them, yielding topics that do not refer to real-life situations but rather to formal legal procedures. The lack of domain knowledge and familiarity with housing law may also hinder evaluators from identifying relevant topics. Despite this issue, we must also emphasize that our topic modeling approach revealed new topics not included in [1]'s RTs. On the basis of CQ1s, 11% and 18% of LDA and BERTopic CTs referred to such uncovered situations.

8. Conclusion

In this paper, we applied topic modeling methods to a corpus of housing law decisions with the goal of automatically extracting topics similar to [1]'s factors. A quantitative analysis showed that LDA and BERTopic seemed to provide the best results, although a further manual analysis revealed that the latter method yielded more relevant topics thanks to its access to semantic information while the former was limited by its bag-of-words approach. As a guideline, we recommend using embedding-based rather than bag-of-words-based topic modeling approaches when dealing with a corpus focused on a single legal area. As future work, we intend to repeat the experiment with a larger corpus by adding claims from landlords, to include experts and non-experts for a broader manual evaluation of topics, and to improve the robustness of automatic metrics for filtering out relevant topics from noisy ones. So far, our results show that we are on a promising track to assist laymen in navigating through technical legal documents by connecting abstract legal concepts with concrete, real-life situations described in everyday language.

Acknowledgements We would like to thank the Cyberjustice Laboratory at the Université de Montréal, the LexUM Chair on Legal Information and the Autonomy through Cyberjustice Technologies project for supporting this research.

References

[1] Westermann H, Walker VR, Ashley KD, Benyekhlef K. Using Factors to Predict and Analyze Landlord-Tenant Decisions to Increase Access to Justice. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law; 2019. p. 133-42.

[2] Tan S, Li Y, Sun H, Guan Z, Yan X, Bu J, et al. Interpreting the public sentiment variations on twitter. IEEE transactions on knowledge and data engineering. 2013;26(5):1158-70.

[3] Griffiths TL, Steyvers M. Finding scientific topics. Proceedings of the National Academy of Sciences. 2004;101(suppl_1):5228-35.

[4] Song CW, Jung H, Chung K. Development of a medical big-data mining process using topic modeling. Cluster Computing. 2019;22(1):1949-58.

[5] Moretti F. Distant reading. Verso Books; 2013.

[6] Dyevre A. Text-mining for lawyers: how machine learning techniques can advance our understanding of legal discourse. Erasmus L Rev. 2021;14:7.

[7] O'Neill J, Robin C, O'Brien L, Buitelaar P. An Analysis of Topic Modelling for Legislative Texts. In: ASAIL@ICAIL; 2017. .

[8] Vīksna R, Kirikova M, Kiopa D. Exploring the Use of Topic Analysis in Latvian Legal Documents. In: Tagarelli A, Zumpano E, Latific AK, Calì A, editors. Proceedings of the First International Workshop

"CAiSE for Legal Documents" (COUrT 2020) co-located with the 32nd International Conference on Advanced Information Systems Engineering (CAiSE 2020), Grenoble, France, June 9, 2020. vol. 2690 of CEUR Workshop Proceedings. CEUR-WS.org; 2020. p. 39-47.

[9] 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. University of New South Wales Law Journal. 2016;39(4):1300-54.

[10] Remmits Y. Finding the topics of case law: Latent dirichlet allocation on supreme court decisions. 2017.

[11] Luz De Araujo PH, De Campos T. Topic Modelling Brazilian Supreme Court Lawsuits. In: Legal Knowledge and Information Systems. IOS Press; 2020. p. 113-22.

[12] Aguiar A, Silveira R, Furtado V, Pinheiro V, Neto JAM. Using Topic Modeling in Classification of Brazilian Lawsuits. In: International Conference on Computational Processing of the Portuguese Language. Springer; 2022. p. 233-42.

[13] Silveira R, Fernandes C, Neto JAM, Furtado V, Pimentel Filho JE. Topic Modelling of Legal Documents via LEGAL-BERT. Proceedings http://ceur-ws org ISSN. 2021;1613:0073.

[14] Branting K, Balhana C, Pfeifer C, Aberdeen JS, Brown B. Judges Are from Mars, Pro Se Litigants Are from Venus: Predicting Decisions from Lay Text. In: JURIX; 2020. p. 215-8.

[15] Chang J, Gerrish S, Wang C, Boyd-Graber J, Blei D. Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems. 2009;22.

[16] Röder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining; 2015. p. 399-408.

[17] Lou A, Salaün O, Westermann H, Kosseim L. Extracting Facts from Case Rulings Through Paragraph Segmentation of Judicial Decisions. In: International Conference on Applications of Natural Language to Information Systems. Springer; 2021. p. 187-98.

[18] Dumais ST, Furnas GW, Landauer TK, Deerwester S, Harshman R. Using latent semantic analysis to improve access to textual information. In: Proceedings of the SIGCHI conference on Human factors in computing systems; 1988. p. 281-5.

[19] Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. Journal of machine Learning research. 2003;3(Jan):993-1022.

[20] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv preprint arXiv:220305794. 2022.

[21] Řehůřek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA; 2010. p. 45-50.

[22] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems; 2017. p. 5998-6008.

[23] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT. Minneapolis, Minnesota: Association for Computational Linguistics; 2019. p. 4171-86.

[24] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019. p. 3982-92.

[25] Reimers N, Gurevych I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 4512-25.

[26] McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018;3(29):861.

[27] McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. The Journal of Open Source Software. 2017;2(11):205.

[28] Trotman A, Puurula A, Burgess B. Improvements to BM25 and language models examined. In: Proceedings of the 2014 Australasian Document Computing Symposium; 2014. p. 58-65.

[29] Lau JH, Newman D, Baldwin T. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics; 2014. p. 530-9.

[30] Bouma G. Normalized (pointwise) mutual information in collocation extraction. Proceedings of GSCL. 2009;30:31-40.

[31] McHugh ML. Interrater reliability: the kappa statistic. Biochemia medica. 2012;22(3):276-82.