Balancing Coherence and Granularity in Legal Topic Modeling: A BERTopic-based Study on Portuguese Appellate Court Decisions

Marjory S. S. Lima[1], Felipe R. Ahad[1], Vih A. S. Silva[1], Thiago M. Ventura[1], Josiel M. Figueiredo[1], Allan G. de Oliveira[1]

[1]Instituto de Computação, Universidade Federal de Mato Grosso, Brazil

{marjory.lima, felipe.ahad, vitoria.silva7}@sou.ufmt.br,
{thiago, josiel, allan}@ic.ufmt.br

Abstract. *The growing backlog of the Brazilian Judiciary demands scalable semantic organization beyond keyword-based retrieval. This study investigates the trade-off between topic coherence and thematic granularity in neural topic modeling for legal texts. We cluster 1,831 Brazilian labor summaries using BERTopic with LegalBERT-pt embeddings and evaluate the impact of the min_topic_size parameter on coherence (C_v), granularity, and diversity. We propose a balanced evaluation score integrating coherence and granularity. Results show higher coherence ($C_v \approx 0.78$) than Latent Dirichlet Allocation ($C_v \approx 0.49$), while enabling controlled topic resolution. These findings provide a practical criterion for tuning topic models in large-scale legal systems.*

1. Introduction

The Brazilian Judiciary has been facing a significant increase in procedural demand. Over 39,4 million new lawsuits were filed in 2024, raising the national caseload to more than 80 million pending cases [Conselho Nacional de Justiça 2025]. This scenario poses substantial challenges to the management of legal knowledge. The high complexity and volume of judicial documents require increasingly efficient methods for the interpretation, retrieval, and reuse of legal information.

Among the various types of procedural documents, appellate court decisions (*acórdãos*) stand out, as they are collective rulings issued by courts that consolidate legal interpretations on recurring or complex cases. The search for case law, previous decisions that guide the interpretation of legal norms, often depends on the ability to locate thematically similar *acórdãos*, which demands considerable effort from legal professionals in terms of reading, interpreting, and comparing texts. Depending on the context, legal professionals may need to review dozens or even hundreds of *acórdãos* to identify semantic patterns. Although keyword-based search systems are available on court websites, these tools are generally limited to literal term queries and do not account for the semantics or context of the words used. In this context, the need for automated methods to assist in the exploratory analysis of large amounts of legal texts becomes evident, particularly for pattern extraction, information synthesis, and thematic organization of court decisions.

In recent years, there has been significant progress in the application of Natural Language Processing (NLP) techniques in the legal domain, with particular emphasis on topic modeling, contextual embeddings, and semantic clustering mechanisms [da Silva et al. 2024, Silva et al. 2024]. Studies have shown that these techniques can be successfully applied to tasks such as archive organization, case retrieval, and support for drafting legal documents [Amorim et al. 2022, Silveira et al. 2023, Novaes et al. 2023].

While recent research has already explored topic modeling in the legal domain, especially for organizing and retrieving case law, there is still limited emphasis on systematically evaluating the thematic coherence of the generated topics, particularly in unsupervised settings.

By focusing on Brazilian court decisions written in Portuguese, this study investigates how modern NLP techniques can support the semantic organization of legal archives, paying special attention to the impact of modeling choices on interpretability and thematic granularity. However, most evaluation approaches emphasize semantic coherence, neglecting aspects such as diversity and distribution of topics across the corpus. Models optimized exclusively for coherence tend to generate few broad and dominant topics, while increasing granularity can produce fragmented clusters. To address this gap, this work proposes a composite evaluation approach that integrates semantic coherence and topic distribution (as a proxy for granularity), providing a more comprehensive criterion for validating the quality of topics in legal corpora.

2. Related Works

The application of NLP techniques to the legal domain has emerged as a growing research area, driven by the need to organize large volumes of court decisions and assist in the retrieval of similar cases [Martins et al. 2021]. Among the approaches adopted, notable ones include topic modeling techniques, semantic representation using embeddings, document clustering, and textual similarity measurement.

Initially, probabilistic topic models such as Latent Dirichlet Allocation (LDA) [Blei et al. 2003] dominated the literature by enabling the automatic identification of latent themes in large text collections. However, these models have limitations in scenarios involving short documents or specialized vocabulary, as is often the case with court decisions and legal headnotes. To address these limitations, several approaches have been proposed, comparing LDA variants with more recent embedding-based methods. Amorim et al. [2022], for instance, conducted an experimental evaluation using short Twitter texts, showing that models like BERTopic and Pseudo-document based Topic Model (PTM) produce more coherent and interpretable topics than traditional LDA.

Although topic coherence is widely used to evaluate topic models, it does not capture other important aspects of topic quality, such as diversity and document coverage [Röder et al. 2015, Mimno et al. 2009]. Empirical studies show that coherence-based evaluation may favor models with fewer and more dominant topics, limiting thematic granularity [Campagnolo et al. 2022]. In the Brazilian legal context, advances in domain-adapted embeddings have improved semantic modeling, yet evaluation remains largely coherence-driven [Silveira et al. 2023]. This limitation motivates evaluation strategies that jointly consider coherence and diversity. Accordingly, this work proposes a composite score that balances these dimensions to better assess topic quality in legal corpora.

In the Brazilian legal context, the semantic representation of legal documents has also advanced with the development of specialized embeddings. Silveira et al. [2023] introduced LegalBERT-pt, a model based on the BERT architecture, trained on a large national legal corpus, and demonstrated its superiority over generic models in tasks such as sentence classification and entity extraction. This research line is crucial

for the effectiveness of semantic clustering and topic modeling techniques that depend on domain-adapted vector representations.

In parallel, research has emerged focusing on the organization and retrieval of case law through thematic clustering. Borges et al. [2025] explored different clustering algorithms using BERTimbau embeddings applied to Portuguese texts, validating the feasibility of unsupervised methods for thematic segmentation of documents. Complementarily, Novaes et al. [2023] proposed a two-stage architecture for legal case retrieval: first, filtering documents based on topics extracted with BERTopic; then, re-ranking the results using semantic ranking functions. This strategy was evaluated in the Competition on Legal Information Extraction and Entailment (COLIEE 2023) international competition, showing promising results for documents from the Brazilian Superior Court of Justice (STJ).

Another important research line is the measurement of semantic similarity between legal documents, which has gained momentum with the use of large language models (LLMs). Hussain et al. [2023] proposed the CoT-STS method, which decomposes the task of textual comparison into interpretable dimensions (topic, agents, actions, context), enabling more accurate semantic evaluations even in zero-shot scenarios. Although not specific to the legal domain, the method provides an interpretable approach that can be applied to comparisons between *acórdãos* and their summaries.

Overall, the literature points to a convergence around three methodological pillars: (i) the use of specialized semantic representations; (ii) the adoption of unsupervised techniques for thematic clustering; and (iii) the use of human or structural resources, such as case summaries or procedural classes, as validation references. The present work fits within this context by employing legal-specific embeddings and BERTopic for clustering judicial summaries. In addition, it contributes by proposing a composite evaluation approach that integrates semantic coherence and topic distribution, addressing limitations of traditional coherence-based evaluation, which often overlooks the balance and coverage of topics across the corpus.

3. Materials and Methods

3.1. Data Collection and Corpus Description

For the development of this work, 1,831 *acórdãos* made publicly available on the website of the Regional Labor Court of the 23rd Region (TRT-23) were used. The data collection was done through web scraping, a technique widely adopted when data cannot be accessed through an API [Mitchell 2024]. This approach consists of using a tool that systematically and automatically simulates browsing and extracting information from web pages [Glez-Peñaa et al. 2014].

At this stage, to ensure records with standardized structures, the following terms were applied to the text filter fields of the TRT-23 search interface[1]: "*isto posto*" (therefore), for the field "*Contendo as palavras (e)*"; and "*ementa*" (summary), for the field "*Palavras na ementa (e)*". In addition, the document types were limited to court decisions, and the judging bodies were restricted to the "*1ª Turma*", "*2ª Turma*" and

[1] https://pje.trt23.jus.br/jurisprudencia

"*Tribunal Pleno*". These criteria aimed to collect only finalized appellate decisions with clearly delimited summaries.

The decision to rely exclusively on case summaries (*ementas*) was guided not only by computational constraints, but also by their functional role within Brazilian legal practice. *Ementas* are designed to synthesize the core legal issues and outcomes of judicial decisions in a standardized and comparable format, making them particularly suitable for large-scale thematic analysis and jurisprudential mapping. While full decisions (*acórdãos*) provide richer argumentative and doctrinal detail, they often exceed the maximum token limits supported by transformer-based models such as BERT and LegalBERT-pt, requiring truncation or segmentation strategies that may disrupt contextual coherence. By focusing on *ementas*, this study prioritizes consistency, comparability, and practical applicability, while acknowledging that certain nuances of judicial reasoning may not be fully captured.

3.2. Topic Modeling Pipeline

3.2.1. Embedding Generation and Topic Modeling

Semantic representations of the summaries were obtained using pre-trained transformer-based embeddings derived from the BERT (Bidirectional Encoder Representations from Transformers) architecture. For this purpose, LegalBERT-pt [Silveira et al. 2023] was employed, a model specialized in the legal domain and based on BERTimbau [Souza et al. 2020], which stands out as the state of the art among models specialized in Brazilian Portuguese [Silva et al. 2024].

The embeddings were then used as input for the topic modeling phase, an approach that enables the synthesis and organization of large document corpora, allowing the discovery of themes and their relationships. This process began with a dimensionality reduction step using UMAP [McInnes et al. 2018], which projects high-dimensional embeddings into a dense representation optimized for neighborhood preservation. The resulting reduced embeddings were then modeled using BERTopic [Grootendorst 2022], a method that employs the HDBSCAN [Campello et al. 2013] algorithm to identify clusters, which are subsequently represented as topics by comparing the importance of words using a class-based variant of the TF-IDF measure, called c-TF-IDF (1).

$$Wt,\, c = TF_{t,\,c} \cdot log\left(1 + \frac{A}{F_t}\right) \tag{1}$$

As shown in Equation 1, $TF_{t,\,c}$ refers to the frequency of term t in class c, F_t refers to the frequency of term t across all classes, and A represents the average number of words per class.

3.2.2. Experimental Configurations with BERTopic

To systematically evaluate the behavior of neural topic modeling under different representational and granularity settings, multiple experimental configurations were executed using the BERTopic framework. In all configurations, the same document embeddings, UMAP dimensionality reduction model, tokenized documents, and dictionary were employed, ensuring that performance differences could be attributed solely to modeling choices rather than data variation.

The default configuration served as a reference point and employed unigram-based representations only, with a fixed minimum topic size of 10 documents. This setup reflects the standard usage of BERTopic without explicit modeling of multiword expressions. To assess the impact of compound legal expressions on topic quality, two additional configurations were evaluated using expanded n-gram representations. The bigram configuration incorporated n-grams in the range (1,2), while the trigram configuration extended this range to (1,3). In both cases, the minimum topic size was kept constant at 10 documents to isolate the effect of n-gram expansion on topic coherence and granularity.

Beyond fixed configurations, a parametric exploration of topic granularity was conducted by systematically varying the *min_topic_size* parameter from 2 to 52. This interval was empirically defined based on preliminary experiments, which indicated that values outside this range led to either excessively fragmented topics or overly dominant clusters. Parameter selection followed empirical tuning guided by prior studies using BERTopic in short-text corpora [Amorim et al. 2022; Novaes et al. 2023]. For each value in this range, a separate BERTopic model was trained and evaluated, enabling an empirical analysis of how topic size constraints influence coherence, diversity, and document distribution across topics.

From this parametric sweep, two reference configurations were identified. The coherence-optimal configuration corresponds to the model achieving the highest C_v, while the balanced configuration was selected based on the maximum value of the balanced score, which jointly considers semantic coherence and topic distribution equity. These configurations represent, respectively, the upper bound of semantic consistency and the best trade-off between coherence and thematic granularity.

3.3. Evaluation Approach

3.3.1. Quantitative Metrics

Finally, to evaluate the quality of the generated topics, the Coherence Score (C_v) [Mimno et al. 2011] was applied. This is a metric that uses a variation of the Normalized Pointwise Mutual Information (NPMI) [Bouma 2009] measure, in which the co-occurrence of a word within a given topic is calculated by comparing that word with all the other words in the same topic. This metric stands out for its high sensitivity to the introduction of noise in topics, making it particularly effective in identifying semantically consistent groupings [Campagnolo et al. 2022]. This is an essential feature in the legal domain, where topics must reflect precise and well-defined normative concepts, such as "*horas extras*" (overtime) or "*assédio moral*" (moral harassment) even in an unsupervised setting. If we consider the words "*horas*" (hours) and "*extras*" the C_v calculation begins with the creation of a vector for both words, given the Equation 2:

$$\nu_{horas,\,extras} = NPMI(horas,\ extras)^{\gamma} = \left(\frac{PMI(horas, extras)}{-log(P(horas, extras) + \epsilon)}\right)^{\gamma} \quad (2)$$

Then, for each word in the topic, a vector is created containing its NPMI with all other words in the same group. For example, the vector for "*horas*" includes its co-occurrence with "*extras*" and "*trabalhadas*" (3).

$$\vec{v}_{horas} = \{\, NPMI(horas,\ horas)^{\gamma}\ ,\ NPMI(horas,\ extras)^{\gamma}\ ,\ NPMI(horas,\ trabalhadas)^{\gamma}\ \} \quad (3)$$

Finally, the distance between the vectors are calculated using the cosine similarity measure (4).

$$C_v = \frac{1}{3} \cdot (cos(\vec{v}_{horas}, \vec{v}_c), cos(\vec{v}_{extras}, \vec{v}_c), cos(\vec{v}_{trabalhadas}, \vec{v}_c)) \quad (4)$$

In addition to the C_v metric, which evaluates the semantic consistency of the most representative terms of each topic, a novel metric is proposed to also consider the balance of document distribution across the topics. The diversity of the topics was defined by Equation 5.

$$Diversity = 1 - \left(\frac{Qty.\ documents\ on\ biggest\ topic}{Total\ no\ of\ documents}\right) \quad (5)$$

Subsequently, the balanced score was defined as:

$$Balanced\ Score = C_v \cdot Diversity \quad (6)$$

While metrics like Topic Diversity [Dieng et al. 2020], the Gini coefficient, and entropy-based measures [Röder et al. 2015] evaluate structural properties independently, the proposed Balanced Score provides a simple, interpretable selection criterion. Rather than replacing these measures, it explicitly couples semantic coherence with topic dominance to penalize overly concentrated clusters without requiring additional parameterization, simultaneously assessing semantic quality and distributional equity to identify the configuration considered the most balanced.

3.3.2. Baseline Model: Latent Dirichlet Allocation

To contextualize the results obtained with neural topic modeling, LDA was adopted as a classical probabilistic baseline [Blei et al. 2003, Mimno et al. 2009, Röder et al. 2015]. The model was trained on the same corpus of tokenized summaries using the same dictionary employed in the BERTopic experiments. Preprocessing additionally included the removal of standard Portuguese stopwords to reduce the influence of non-informative tokens on word co-occurrence statistics.

The number of topics was set to 82, matching the balanced BERTopic configuration to enable a fair comparison. Training was performed with a fixed random seed and 10 passes to ensure convergence stability. Topic quality was evaluated using the same C_v coherence metric adopted for BERTopic. The resulting coherence score (0.4883) serves as a reference for comparing probabilistic and neural topic modeling approaches on Portuguese legal texts.

3.3.3. Qualitative Evaluation by Domain Experts

In addition to the quantitative evaluation metrics, a qualitative assessment was conducted to examine the interpretability and practical relevance of the topics generated by the proposed approach. This evaluation aimed to complement automatic measures by incorporating expert judgment, which is particularly important in the legal domain, where semantic adequacy and doctrinal coherence are essential.

The qualitative analysis focused on the balanced BERTopic configuration, selected for presenting the best trade-off between semantic coherence and equitable document distribution, as indicated by the proposed balanced score. From the total set of 113 topics produced by this configuration, a subset of five topics was selected for expert evaluation. The selected topics were chosen to reflect thematic diversity and to

cover different types of legal issues commonly addressed in labor court decisions. For each topic, evaluators were provided with: (i) a list of the most representative terms automatically extracted by the model, and (ii) a sample of *ementas* associated with the topic, allowing for an informed assessment of the relationship between the topic descriptors and the underlying judicial content.

The evaluation was conducted through an online survey, in which three legal domain specialists were asked to assess each topic based on a single evaluative statement encompassing three dimensions simultaneously: legal thematic coherence, clarity and adequacy of the representative terms, and practical usefulness for legal research or jurisprudential analysis. Responses were recorded using a five-point Likert Scale ranging from "Strongly disagree" to "Strongly agree" [Joshi et al. 2015]. Each evaluator assessed all selected topics independently. The results of this qualitative evaluation were subsequently analyzed to identify patterns of agreement and to contrast expert perceptions with the quantitative findings derived from coherence and distribution-based metrics.

3.4. Reproducibility Considerations

All experimental runs, including BERTopic configurations and the LDA baseline, were executed with fixed random seeds to ensure deterministic behavior. Intermediate results, evaluation metrics, and configuration parameters were systematically logged and stored. The complete experimental pipeline, including preprocessing routines, modeling configurations, and evaluation scripts, is publicly available at https://github.com/msoehn3/legal-topic-modeling, enabling full replication of the results.

4. Results

The experiments were conducted with the objective of evaluating the performance of the BERTopic algorithm to identify latent topics in *ementas* under different parameter configurations and compare them with a LDA baseline. The consolidated results are presented in Table 1, which shows, for each configuration, the number of topics generated, the average thematic coherence (measured by the C_v metric), the proportions of topics with the lowest and highest document concentration, as well as the deviation between these extremes, used as a proxy for thematic balancing.

Table 1 - Results of the Experiments

Experiment	Min. Topic Size	Topics	C_v	Balanced Score	Min. Proportion	Max. Proportion	Deviation
Balanced	4	113	0.75	**0.63**	0.22%	16.82%	**16.60%**
Optimal	7	71	**0.78**	0.62	0.44%	20.32%	19.88%
Default	10	52	0.75	0.57	0.55%	23.48%	22.94%
Trigrams	10	52	0.74	0.57	0.55%	23.48%	22.94%
Bigrams	10	52	0.74	0.57	0.55%	23.48%	22.94%
LDA Baseline	-	113	0.49	-	-	-	-

Among the BERTopic configurations, coherence values ranged from 0.7409 to 0.7782, substantially outperforming the LDA baseline, which achieved a C_v score of 0.4883 with the same number of topics (113). This confirms the advantage of neural

topic modeling with contextualized legal embeddings over probabilistic bag-of-words approaches in the legal domain.

The configuration Optimal exclusively for coherence achieved the highest coherence score (C_v = 0.78), generating 71 topics. While this setting slightly increased semantic consistency, it also exhibited a higher concentration of documents in its largest topic (max proportion = 20.32%), resulting in a larger deviation between the smallest and largest topics (19.88%). This behavior suggests a moderate tendency toward topic concentration when coherence is prioritized, with potential implications for thematic granularity.

The Balanced configuration produced a finer-grained topic structure with 113 topics, while preserving a high level of coherence (C_v = 0.75). This configuration achieved the highest balanced score (0.63) and reduced document concentration across topics, yielding the smallest deviation (16.60%) among the BERTopic variants. As a result indicating a slightly more even distribution of documents with only a marginal reduction in semantic coherence.

The Default configuration generated 52 topics with a coherence of 0.75 and moderate balanced score (0.57). Expanding the representation to include bigrams and trigrams did not improve performance. Both Bigrams and Trigrams configurations resulted in slightly lower coherence values (0.7409 and 0.7443, respectively), identical topic counts, and unchanged distributional characteristics. These results suggest that, for this corpus, increasing the n-gram range did not significantly enhance topic quality.

Figure 1 illustrates the trade-off between coherence and diversity as a function of the *min_topic_size* parameter. The curve shows that coherence increases as topics become larger and more semantically stable, reaching a peak at intermediate values. Beyond this point, diversity declines sharply, reflecting the concentration of documents into fewer dominant topics. This behavior highlights the structural tension between semantic coherence and thematic granularity in legal topic modeling and supports the use of a combined evaluation criterion.

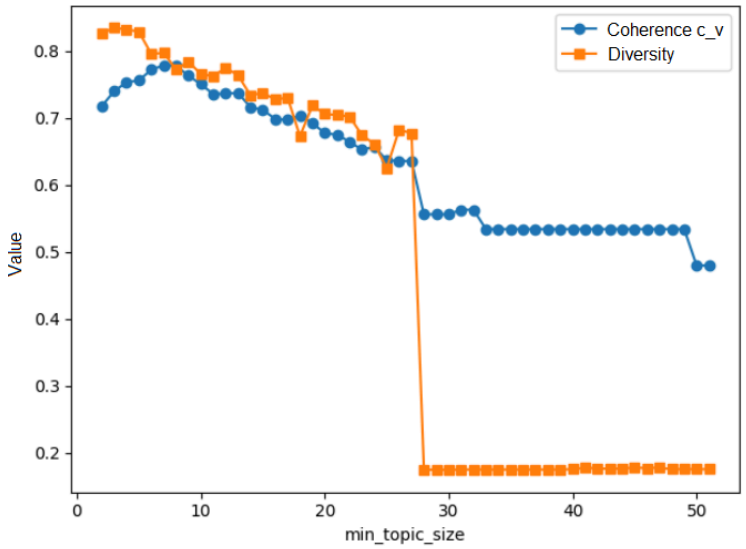

Figure 1. ***Sensitivity to the min_topic_size parameter.***

The qualitative inspection of the LDA topics further reinforces the quantitative findings. As illustrated by the extracted topic-word lists, many LDA topics are

dominated by function words and high-frequency legal terms (e.g., “de”, “da”, “o”, “que”), with limited semantic specificity. Even when legally relevant terms appear (e.g., “horas extras”, “insalubridade”, “dano moral”), they are often mixed with syntactic fillers, reducing interpretability. This lack of semantic focus is consistent with the low coherence score obtained by LDA (0.49) and contrasts sharply with the more thematically cohesive topics produced by BERTopic.

With C_v coherence of up to 0.74, using BERTopic combined with LegalBERT-pt, this study presents robust performance in generating semantically consistent topics, overcoming challenges frequently observed in the literature. When compared to the results reported by Amorim et al. [2022], who obtained C_v coherence values in the range of 0.492 to 0.623 in short texts using different topic modeling techniques, the results of this work are significantly superior. The difference can be attributed mainly to the use of specific embeddings, which capture more robust semantic relationships.

The results of the qualitative evaluation are presented in Table 2. The evaluation conducted with legal experts corroborates the quantitative findings obtained through coherence and distribution-based metrics. Topics associated with well-established legal concepts, such as working hours, discriminatory dismissal, and procedural appeals, achieved the highest levels of expert agreement, indicating strong semantic coherence and practical relevance. Lower variance across evaluators further suggests consistency in expert perception, reinforcing the interpretability and robustness of the proposed topic modeling approach.

Table 2 - Human Evaluation of BERTopic Topics (Balanced Configuration)

ID	Representative Terms	Mean Score	Std. Dev.	Agreement Level
81	*dispensa, discriminatória, discriminação, hiv, portador, preconceito, estigma, grave, justa, discriminatório*	4.33	0.58	High
14	*agravo, petição, decisão, impugnação, conhecido, cálculos, oferecida, ultrapassa, barreira, aprecia*	4.67	0.58	Very High
3	*jornada, ponto, cartões, ônus, controles, extras, horários, 338, horas, prova*	5.00	0.00	Full
94	*71, intervalo, 4º, térmico, suprimido, indenizatória, 384, 2017, redação, 11*	3.67	0.58	Moderate
35	*audiência, comparecimento, arquivamento, acordo, extrajudicial, importa, 844, homologação, partes, inaugural*	4.33	0.58	High

Overall, the results indicate that it is possible to obtain semantically consistent and distributively balanced models. The C_v metric proved useful as an initial evaluation criterion, but its combination with distribution measures provides a more robust analysis for real-world applications in the legal context.

5. Discussion

The results obtained indicate the potential of the proposed approach for identifying thematic structures within judicial documents. The experimental results evidenced a substantial advantage of BERTopic over the LDA baseline in the context of Brazilian labor court decisions. While LDA achieved a low coherence score, BERTopic consistently produced more semantically coherent and interpretable topics across all

evaluated configurations. This gap highlights the limitations of probabilistic topic models based on bag-of-words assumptions when applied to complex legal texts, which are characterized by long documents, formulaic language, and high lexical redundancy. This finding is consistent with prior studies that report reduced effectiveness of LDA in short or domain-specific texts, where limited word co-occurrence hinders topic formation [Amorim et al. 2022].

In addition, the dataset is composed of labor court decisions, which typically involve recurrent thematic patterns across documents. This characteristic provides important context for interpreting the results, especially given the consistent performance differences observed between the evaluated models.

An important finding of this study concerns the trade-off between semantic coherence and topic diversity controlled by the *min_topic_size* parameter. Configurations optimized exclusively for coherence tended to generate fewer, more dominant topics, increasing document concentration and reducing thematic granularity. Conversely, configurations favoring diversity produced a larger number of more evenly distributed topics, at the cost of a modest reduction in coherence.

Furthermore, prior work has emphasized that coherence metrics alone may not fully capture topic quality. Campagnolo et al. 2022 demonstrate that coherence measures are sensitive to noise but do not account for distributional aspects of topics, while Röder et al. 2015 highlight that different coherence formulations may favor different topic structures. In line with these findings, the results of this study provide empirical evidence of this effect. The coherence-optimal configuration, despite achieving the highest C_v score (0.78), also exhibits a higher maximum topic proportion (20.32%) compared to the balanced configuration (16.82%), indicating stronger dominance of specific topics.

Notably, the difference between the Optimal and Balanced configurations was quantitatively small, suggesting a plateau effect in which further gains in coherence no longer translate into meaningful improvements in topic quality. This observation reinforces the limitation of relying solely on coherence metrics, as also discussed in the literature, where coherence improvements may not correspond to better semantic coverage or usability in downstream tasks. Importantly, a high C_v score indicates lexical coherence within the corpus, but does not guarantee identical legal outcomes across documents, reinforcing the necessity of complementary expert analysis.

The alignment between human evaluations and automatic coherence metrics provides empirical evidence that the proposed balanced score effectively captures dimensions of topic quality that are meaningful to legal experts. While configurations optimized solely for coherence may inflate semantic consistency at the expense of thematic granularity, the balanced configuration indicates that it is possible to achieve high interpretability without sacrificing topic diversity. This alignment strengthens the argument that the proposed evaluation framework is suitable for real-world legal information retrieval and jurisprudential analysis tasks.

6. Limitations

This study presents some limitations that should be considered when interpreting the results. First, while standard Portuguese stopwords and digits were removed to prevent non-informative function words from dominating LDA topics, the model was evaluated

using default tokenization parameters without task-specific optimization. Since probabilistic topic models rely on word co-occurrence patterns, this setup may not reflect the best achievable performance of LDA, particularly in short or highly standardized texts. Additionally, while the human evaluation provided key expert validation, it was restricted to a representative subset of five topics due to manual annotation costs, serving as a focused qualitative sample rather than an exhaustive assessment of all generated clusters. Finally, the experiments were conducted on a specific corpus of Brazilian labor court decisions, which may limit the generalizability of the findings to other legal domains or languages, especially those with different linguistic or structural characteristics, which may influence topic distribution patterns.

7. Conclusion and Future Work

This study addressed the challenge of organizing large volumes of judicial documents by investigating how topic modeling can generate semantically coherent and well-balanced thematic structures in Brazilian labor court summaries. The objective was to analyze the trade-off between topic coherence and granularity, and to propose an evaluation framework capable of capturing both aspects.

The results show that BERTopic consistently outperforms the LDA baseline in terms of semantic coherence. However, configurations optimized exclusively for coherence tend to produce dominant topics and reduced thematic coverage. In contrast, the balanced configuration achieved a more equitable distribution of documents while maintaining high coherence, resulting in the best overall performance. These findings indicate that coherence alone is insufficient to assess topic quality and that incorporating distributional aspects leads to more informative and usable topic structures.

For future work, it is recommended to expand the scope of the analysis beyond abstract summaries by incorporating full judicial decisions, which may enable the identification of more diverse and detailed topics. It is also important to validate the proposed approach on datasets from other courts, in order to assess its generalizability across different legal contexts. Additionally, comparing the performance of different domain-specific embedding models may help clarify their impact on topic quality. Finally, exploring Neural Topic Models (NTMs) constitutes a promising direction to further extend the results presented in this study.

Ethics Statement

This research uses publicly available legal documents and does not involve personal or sensitive data. In addition, a qualitative evaluation was conducted with human participants. The data collection was performed anonymously, without the collection of personal or sensitive information. All procedures followed applicable ethical standards.

Use of Generative AI

Generative AI tools were used to assist in language revision and text refinement. All scientific content, experimental design, and analysis were conducted by the authors.

References

Amorim, A., Murrugarra-Llerena, N., Silva, V., de Oliveira, D. and Paes, A. (2022) "Modelagem de Tópicos em Textos Curtos: uma Avaliação Experimental", In: Proceedings of the 37th Simpósio Brasileiro de Banco de Dados (SBBD), Búzios, Brazil. Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/sbbd.2022.224314.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003) "Latent Dirichlet Allocation", Journal of Machine Learning Research, vol. 3, pp. 993–1022. Available at: https://dl.acm.org/doi/10.5555/944919.944937.

Borges, B. R. (2025) "Comparison of Clustering Techniques in Text Documents in Portuguese", ISys - Brazilian Journal of Information Systems, vol. 18, no. 1, pp. 4:1–4:17. DOI: https://doi.org/10.5753/isys.2025.5029.

Bouma, G. (2009) "Normalized (Pointwise) Mutual Information in Collocation Extraction". Available at: https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf.

Campagnolo, J. M., Duarte, D. and Dal Bianco, G. (2022) "Topic coherence metrics: How sensitive are they?", Journal of Information and Data Management, vol. 13, no. 4, pp. 476–491. DOI: https://doi.org/10.5753/jidm.2022.2181.

Campello, R., Moulavi, D. and Sander, J. (2013) "Density-Based Clustering Based on Hierarchical Density Estimates", In: Advances in Knowledge Discovery and Data Mining, Springer. DOI: https://doi.org/10.1007/978-3-642-37456-2_14.

Conselho Nacional de Justiça (CNJ) (2025) Justiça em Números 2025: Ano-base 2024. Departamento de Pesquisas Judiciárias, Brasília, DF.

da Silva, L., Rodrigues, M., Archanjo, A., Pessoa, L., Silva, M., de Almeida, T., & Silveira, L. (2024). Segmentação Textual Baseada em Tópicos em Português Utilizando BERTimbau. In Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana, (pp. 32-36). Porto Alegre: SBC. doi:10.5753/stil.2024.245080.

Dieng, A. B., Ruiz, F. J. R. and Blei, D. M. (2020) "Topic Modeling in Embedding Spaces", Transactions of the Association for Computational Linguistics, vol. 8, pp. 439–453. DOI: https://doi.org/10.1162/tacl_a_00325.

Dragnut, E. et al. (2009) "Stop word and related problems in web interface integration", Proceedings of the VLDB Endowment. DOI: http://dx.doi.org/10.14778/1687627.1687667.

Glez-Peña, D. et al. (2014) "Web scraping technologies in an API world". Briefings in Bioinformatics. DOI: https://doi.org/10.1093/bib/bbt026.

Grootendorst, M. (2022) "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure", arXiv preprint. DOI: https://doi.org/10.48550/arXiv.2203.05794.

Hussain, M., Rehman, U. U., Nguyen, T. D. T. and Lee, S. (2023) "CoT-STS: A Zero-shot Chain-of-thought Prompting for Semantic Textual Similarity", In:

Proceedings of the 6th Artificial Intelligence and Cloud Computing Conference (AICCC), Kyoto, Japan. ACM. DOI: https://doi.org/10.1145/3639592.3639611.

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert Scale: Explored and Explained. Current Journal of Applied Science and Technology, 7(4), 396–403. https://doi.org/10.9734/BJAST/2015/14975.

Martins, V. S. and Silva, C. D. (2021) "Text Classification in Law Area: A Systematic Review", In: Proceedings of the Symposium on Knowledge Discovery, Mining and Learning (KDMiLe). DOI: http://dx.doi.org/10.5753/kdmile.2021.17458.

Mimno, D. et al. (2009) "Optimizing Semantic Coherence in Topic Models", In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). DOI: https://dl.acm.org/doi/10.5555/2145432.2145462.

Mitchell, R. (2024) Web Scraping com Python (3rd ed.). São Paulo: Novatec Editora.

Novaes, L. P., Vianna, D. and da Silva, A. S. (2023) "Modelagem de Tópicos para a Tarefa de Recuperação de Casos Legais", In: Proceedings of the 38th Simpósio Brasileiro de Banco de Dados (SBBD), Belo Horizonte, Brazil. Sociedade Brasileira de Computação. DOI: https://doi.org/10.5753/sbbd.2023.232576.

Röder, M., Both, A. and Hinneburg, A. (2015) "Exploring the Space of Topic Coherence Measures", In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining (WSDM), Shanghai, China. ACM, pp. 399–408. Available at: http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf.

Sarica, S. and Luo, J. (2021) "Stopwords in Technical Language Processing", PLOS ONE, vol. 16, no. 7. DOI: https://doi.org/10.1371/journal.pone.0254937.

Silva, M. et al. (2024) "Evaluating Domain-adapted Language Models for Governmental Text Classification Tasks in Portuguese", In: Proceedings of the 39th Brazilian Symposium on Databases (SBBD), Florianópolis, Brazil. DOI: https://doi.org/10.5753/sbbd.2024.240508.

Souza, F., Nogueira, R., and Lotufo, R. (2020). Bertimbau: pretrained bert models for brazilian portuguese. In Proceedings of the Brazilian Conference on Intelligent Systems, pages 403–417. Springer. DOI: https://doi.org/10.1007/978-3-030-61377-8_28.

Silveira, R., Ponte, C., Almeida, V., Pinheiro, V. and Furtado, V. (2023) "LegalBERT-pt: A Pretrained Language Model for the Brazilian Portuguese Legal Domain", In: Proceedings of the 12th Brazilian Conference on Intelligent Systems (BRACIS), Belo Horizonte, Brazil. Sociedade Brasileira de Computação. Available at: https://sol.sbc.org.br/index.php/bracis/article/view/28420.