





RESEARCH PAPER

Interpreting Lawsuits Contexts through Probabilistic Topic Modeling

Étore Braga e Santos   [Universidade de São Paulo | etorebraga@usp.br]
Ildeberto Aparecido Rodello   [Universidade de São Paulo | rodello@usp.br]

Abstract. The increasing volume and complexity of digital legal records have underscored the need for scalable analytical tools capable of extracting meaningful insights from unstructured text. This study investigates the application of topic modeling techniques Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Probabilistic Latent Semantic Analysis (pLSA) to the classification and interpretation of legal documents, with a focus on lawsuits related to special education policies in Brazil. A corpus of 4,259 judicial cases was collected from the São Paulo Court of Justice, preprocessed, and analyzed to identify latent thematic structures. Model performance was evaluated using the coherence score and further validated through human interpretability assessments. The results indicate that pLSA performs well with fewer topics, capturing broader legal themes, while LDA excels at higher topic counts, effectively distinguishing nuanced legal issues such as access to education and contractual disputes. The findings highlight the potential of probabilistic topic modeling as a decision-support tool, reinforcing the role of artificial intelligence in improving legal transparency, accessibility, and analytical depth, without compromising the autonomy of legal interpretation.

Keywords: Topic Modeling, Jurimetry, LDA, LSA, pLSA, Natural Language Processing, Text Mining, Legal Analytics

Received: 25 April 2025 • Accepted: 07 June 2025 • Published: 13 June 2025

1 Introduction

The exponential growth of digital legal records has created vast repositories of Legal Big Data, encompassing millions of judicial decisions and statutes. This proliferation makes manual analysis impractical, necessitating computational approaches to extract meaningful insights [Devins *et al.*, 2017; Zödi, 2017; Garg and Ma, 2025]

Artificial Intelligence (AI) has emerged as a crucial tool for legal professionals navigating complex regulatory frameworks. Topic modeling plays a key role in uncovering hidden themes within unstructured legal texts, revealing patterns that enhance legal research and support judicial decision-making [Zödi, 2017; Garg and Ma, 2025]. Unlike traditional keyword searches, topic modeling employs probabilistic frameworks to uncover semantic structures, enabling nuanced interpretation of legal discourse [Devins *et al.*, 2017; Chen *et al.*, 2023].

Several studies have applied AI techniques to judicial cases, demonstrating the potential to improve legal analytics and decision support. For example, recent research has leveraged Machine Learning to classify legal documents, predict case outcomes, and extract thematic patterns from court rulings [Garg and Ma, 2025; Ma *et al.*, 2024]. These works highlight both the opportunities and challenges in integrating AI into legal workflows, such as data privacy concerns and the complexity of legal language. However, many existing studies focus on Deep Learning or Large Language Models (LLMs), which often require extensive computational resources and large datasets.

In contrast, this study employs conventional machine learning techniques—Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Probabilistic Latent Semantic Analysis (pLSA)—to analyze 4,259 Brazilian legal documents related to special education lawsuits. These meth-

ods offer several advantages: they generate explicit, interpretable topic representations; require less computational power; and provide greater control over model outputs. This approach aligns with the need for transparent and domain-relevant AI tools in legal contexts, complementing recent advances in generative AI while addressing practical constraints [Garg and Ma, 2025; Ma *et al.*, 2024].

The objective of this research is to evaluate and compare these topic modeling techniques in terms of their coherence and interpretability, aiming to support legal transparency and accessibility through scalable computational methods. By focusing on specialized legal domains and combining quantitative metrics with human analysis, this work contributes to the growing field of AI-assisted legal analytics, addressing both technical and domain-specific challenges identified in recent literature [Garg and Ma, 2025; Ma *et al.*, 2024].

2 Topic Modeling Fundamentals

Topic modeling is a statistical technique that identifies the pattern of words and word sequences within a collection of documents containing unstructured text data [Blei, 2012]. The technique analyzes natural language to infer topic clusters within a collection of documents. This inference is based on the frequency of words throughout the collection and within individual documents, grouping semantically similar words based on their embedding vectors [Ramage *et al.*, 2009]. These clusters are represented by the words they contain, enabling the rapid extraction of meaningful insights from large collections of data.

Topic modeling has broad applications across diverse fields. As an unsupervised machine learning technique, it eliminates the need for training data, significantly reducing processing time. The applications of topic modeling are broad in diverse domains, including information re-

trieval [Mehrotra *et al.*, 2013], text summarization [Griffiths *et al.*, 2004], document segmentation [Mimno *et al.*, 2011], sentiment analysis [Naskar *et al.*, 2021], and document classification [Wang, 2008]. This methodology provides a robust framework for systematically analyzing and organizing vast amounts of unstructured text data.

Therefore, documents consist of unstructured text data that encompass a mixture of various topics. Topics are represented as distributions over words, while words or sequences of words serve as the fundamental observable units of text.

2.1 Term Frequency - Inverse Term Frequency (TF-IDF)

The TF-IDF score, or Term Frequency-Inverse Document Frequency, quantifies the importance of a word within a document relative to a larger collection of documents [Jones, 1972]. This collection of documents and terms is structured as a Document-Term Matrix (DTM), which organizes documents (rows) and their terms (columns) into a matrix, enabling quantitative text analysis. As a foundational technique for advanced topic modeling, TF-IDF assigns weighted significance scores by penalizing frequently occurring terms across the corpus while amplifying the weight of rarer terms. Unlike simple word counting, TF-IDF enhances significance assessment by considering a term's distribution throughout the corpus.

The mathematical definition is as follows:

$$W(i, j) = tf(i, j) \times \log \left(\frac{N}{df(i)} \right) \quad (1)$$

where $W(i, j)$ represents the weight of term i in document j , $tf(i, j)$ denotes the term frequency, N is the total number of documents, and $df(i)$ signifies the document frequency of term i .

Equation (1) forms the basis for extracting insights from the Document-Term Matrix (DTM), facilitating advanced techniques such as Latent Semantic Analysis (LSA) and Probabilistic Latent Semantic Analysis (pLSA).

2.2 Latent Semantic Analysis (LSA)

The LSA utilizes Singular Value Decomposition (SVD) to reveal semantic patterns within textual data by decomposing the Document-Term Matrix (DTM) [Steiyvers and Griffiths, 2006]. This decomposition is represented as Equation (2):

$$\max_{\text{DTM}[m \cdot n]} \{U[m \cdot t] \cdot \Sigma[t \cdot t] \cdot V^T[t \cdot n]\} \quad (2)$$

where U and V are orthogonal matrices representing the similarity between documents (U) and terms (V), and Σ is a diagonal matrix containing the singular values of the DTM. These singular values emphasize the importance of latent semantic structures, helping to capture meaningful patterns in textual data.

The TF-IDF-weighted DTM serves as the input for SVD, decomposing the matrix into three components: U , Σ , and V^T . In this factorization, $U[m \cdot t]$ encodes the relationship between documents and topics, $V^T[t \cdot n]$ captures the relationship between terms and topics, and $\Sigma[t \cdot t]$ contains the singular values, representing the importance of each topic.

To emphasize the most significant latent topics, LSA retains only the first t singular values, where t is less than or equal to the smaller dimension of the DTM. By discarding the remaining $m - t$ rows of U and $n - t$ columns of V , the reduced matrices effectively capture the most relevant semantic structures while minimizing noise.

$$U[m \cdot t], \quad \Sigma[t \cdot t], \quad V^T[t \cdot n] \quad (3)$$

The resulting decomposition provides two critical matrices:

- **Document-topic matrix:** $U[m \cdot t]$, which associates documents with topics.
- **Term-topic matrix:** $V^T[t \cdot n]$, which links terms to topics.

Together, these matrices enable the analysis of semantic relationships within the text, with topic importance encoded in Σ serving as a guiding factor [Steiyvers and Griffiths, 2006].

LSA effectively reduces the dimensionality of text data while preserving its semantic structure. However, it requires a predefined number of topics and involves computationally intensive preprocessing, which can affect scalability and efficiency [Steiyvers and Griffiths, 2006].

2.3 Probabilistic Latent Semantic Analysis (pLSA)

The pLSA distinguishes itself from LSA by adopting a probabilistic framework based on a latent class model. Unlike LSA, which employs Singular Value Decomposition (SVD) to approximate the DTM, pLSA utilizes a mixture decomposition, offering a principled statistical foundation [Hofmann, 1999].

The pLSA model expresses the joint probability of a document d and a word w as follows in the Equation (4):

$$P(d, w) = \sum_z P(z)P(d|z)P(w|z), \quad (4)$$

where z represents the set of latent topic, $P(z)$ denotes the probability of a topic, $P(d|z)$ represents the probability of a document given a topic, and $P(w|z)$ is the probability of a word given a topic. This formulation assumes conditional independence between documents and words given the latent variable z .

To estimate the parameters of the model, pLSA uses the Expectation-Maximization (EM) algorithm [Meng and van Dyk, 1997]. The algorithm alternates between two steps. In the expectation step, the posterior probabilities of the latent topics are calculated as (5):

$$P(z|d, w) = \frac{P(z)P(d|z)P(w|z)}{\sum_{z'} P(z')P(d|z')P(w|z')}. \quad (5)$$

In the maximization step, the model parameters are updated to maximize the likelihood of the observed data. The updates are performed as follows (6) (7) (8):

$$P(w|z) \propto \sum_d n(d, w)P(z|d, w), \quad (6)$$

$$P(d|z) \propto \sum_w n(d, w) P(z|d, w), \quad (7)$$

$$P(z) \propto \sum_{d, w} n(d, w) P(z|d, w). \quad (8)$$

Here, $n(d, w)$ represents the frequency of word w in document d . By maximizing the likelihood function, pLSA minimizes the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951] between the empirical data distribution and the model's predictions. This process creates a probabilistic latent semantic space that effectively captures key co-occurrence patterns, as discussed by [Hofmann, 1999].

Compared to LSA, pLSA offers a probabilistic framework that results in well-normalized distributions, where latent topics correspond to meaningful word distributions and document-topic relationships. While LSA relies on linear algebra, which can produce negative values or non-interpretable directions in the latent space, pLSA provides a statistical interpretation with properly normalized probability distributions. This ensures a more interpretable and coherent mapping of topics, enhancing the understanding of latent semantic structures.

The iterative nature of the EM algorithm makes pLSA computationally more intensive than LSA. However, it often demonstrates superior predictive performance, as evidenced by lower perplexity and improved retrieval tasks in experimental evaluations [Hofmann, 1999]. This increased computational cost is typically offset by the enhanced accuracy and interpretability of the resulting topic models.

2.4 Latent Dirichlet Analysis (LDA)

The LDA builds on the probabilistic framework of pLSA but differs by incorporating Bayesian principles for term-topic distributions. Instead of using the joint probability approach of pLSA, LDA employs Dirichlet distributions for modeling document-topic and topic-word distributions [Blei *et al.*, 2003]. This Bayesian approach introduces priors that allow for more robust handling of sparse data and improves the generalization of topic models.

Performing as a distribution of distributions, the Dirichlet distribution provides a framework for understanding the probability distributions. The main function of LDA can be divided into the following components:

- α : The Dirichlet hyperparameter that controls the probability distribution of topics within each document. Determines how likely it is for a document to exhibit a mixture of topics.
- η : The Dirichlet hyperparameter that governs the probability distribution of words across topics. Describes the likelihood that each word will be assigned to a specific topic.
- θ : The multinomial distribution that represents the probability of topics assigned to each document. It describes the topic distribution for a given document.
- β_k : The multinomial distribution that represents the probability of words assigned to the topic k . It captures the likelihood of words occurring on a particular topic.

From these foundational components, LDA derives:

- z , representing the topic assignment for each word within a document, forming the distribution of topics throughout the document.
- w , indicating the distribution of words in topics.

As these distributions collectively form W (the set of observed words), iterating over multiple documents constructs M , the entire corpus. The objective of LDA is to estimate the parameters:

- θ , derived from α , representing the topic distribution per document.
- β_k , derived from η , representing the word distribution per topic.

To achieve this, LDA maximizes the likelihood function:

$$P(W | \alpha, \eta). \quad (9)$$

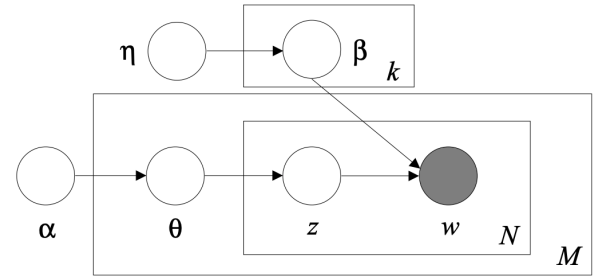


Figure 1. Graphical representation of the Latent Dirichlet Allocation model, illustrating the relationships between parameters and variables in the probabilistic framework Blei *et al.* [2003].

Notation:

- α : Hyperparameter of the Dirichlet prior on the per-document topic distributions.
- β_k : The topic-word distribution contained in the universe of topic k , a multinomial distribution over words for each topic, influenced by η .
- η : Hyperparameter of the Dirichlet prior on the per-topic word distributions.
- θ : The document-topic distribution, a multinomial distribution over topics for each document, influenced by α .
- z : The topic for each word in a document, representing the topic assignment in the topic distribution θ .
- w : The observed word in the document, sampled from the word distribution β_k for the corresponding topic z .
- N : The number of words in a document, corresponding to the inner plate in the graphical model.
- M : The number of documents in the corpus, corresponding to the outer plate in the graphical model.
- k : The number of topics in the model.

The core formula of LDA is given by Equation (10), which defines the generative process:

$$P(W, Z, \theta, \beta; \alpha, \eta) = \prod_{j=1}^M P(\theta_j; \alpha) \prod_{k=1}^K P(\beta_k; \eta) \prod_{t=1}^N P(Z_{j,t}|\theta_j) P(W_{j,t}|\beta_{Z_{j,t}}). \quad (10)$$

where $P(W, Z, \theta, \beta; \alpha, \eta)$ is the joint probability distribution over the observed words (W), the latent topic assignments (Z), and the parameters (θ, β). Specifically:

- $P(\theta_j; \alpha)$ is the Dirichlet prior for the document-topic distribution θ_j , controlled by hyperparameter α .
- $P(\beta_k; \eta)$ is the Dirichlet prior for the topic-word distribution β_k , controlled by hyperparameter η .
- $P(Z_{j,t}|\theta_j)$ is the probability of topic assignment $Z_{j,t}$ for word t in document j , drawn from the multinomial distribution θ_j .
- $P(W_{j,t}|\beta_{Z_{j,t}})$ is the probability of the observed word $W_{j,t}$ given the topic-word distribution $\beta_{Z_{j,t}}$.

This generative process models a corpus of M documents, each containing N words, with K latent topics [Blei *et al.*, 2003].

3 Materials and methods

This study follows a structured approach to analyze legal texts using topic modeling techniques. It consists of the following key steps: data collection, preprocessing, topic modeling, coherence score evaluation, sampling and human analysis. Figure 2 provides an overview of the workflow.

First, judicial documents (lawsuits) were collected through web scraping from the São Paulo Court of Justice (TJSP) platform, ensuring a comprehensive dataset of legal cases. This dataset includes a diverse range of rulings, covering various legal domains.

The preprocessing phase was essential to refine the textual data and improve the effectiveness of topic modeling. It involved multiple steps. First, text normalization was applied to standardize the data by converting them to lowercase and ensuring consistent character encoding. Then, tokenization was performed, splitting the text into individual words or tokens for further analysis.

To reduce noise, stopwords removal was conducted, eliminating common words that do not contribute meaningful information. Given the complexity of legal language, a custom stopwords list was developed to exclude domain-specific but non-informative terms. Finally, lemmatization was used to reduce words to their base forms, enhancing linguistic coherence and improving topic extraction.

To identify latent structures within the dataset, topic modeling techniques were employed, including LDA [Blei *et al.*, 2003], LSA [Steyvers and Griffiths, 2006], and pLSA [Hofmann, 1999]. Each method analyzed word distributions to cluster legal cases into meaningful themes.

To assess the interpretability of the generated topics, the coherence score [Röder *et al.*, 2015] was computed for each model. This metric provided a quantitative measure of topic quality by evaluating the semantic consistency of the words within each topic. The coherence score also guided the selection of the optimal number of topics, ensuring a balance between granularity and interpretability.

Sampling was conducted to ensure a balanced representation of topics, maintaining diversity in legal case selection.

This step was essential to prevent overrepresentation of specific themes and to capture the full spectrum of legal discourse.

Finally, a human evaluation was performed to verify the interpretability of the extracted topics. A human compared model outputs with case details to assess classification accuracy and relevance. This qualitative assessment complemented the computational analysis, ensuring that the identified topics aligned with real-world legal reasoning.

By combining computational modeling with human validation, the methodology provided robust and meaningful insights into legal text classification, reinforcing the reliability of topic modeling techniques in the legal domain.

3.1 Data Collection

The data gathering process was designed to capture lawsuits related to special education laws and policies. After preprocessing, a total of 4,259 unique cases remained following deduplication.

To further refine the dataset, topic modeling techniques, specifically Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003], were applied to identify and filter cases aligned with the scope of the study. This step ensured that only the most relevant legal texts were retained, providing a solid foundation for subsequent legal and linguistic analyses.

3.1.1 Exploratory Data Analysis (EDA)

Each record of the dataset contains 27 attributes that capture essential legal elements such as case type (*classe processual*), jurisdiction (*comarca*), judicial body (*órgão julgador*), and ruling (*sentença*).

Most cases belong to the first instance (*1ª instância*), with 44.1% classified as *Mandado de Segurança Cível* (Writ of Mandamus – Civil) and 35.1% as *Procedimento Comum Cível* (Ordinary Civil Procedure). The most frequent legal topic is *Estabelecimentos de Ensino* (Educational Institutions), accounting for 30.5% of the cases, underscoring the judiciary's role in educational matters.

Regarding case outcomes, the majority of rulings (*sentenças*) were favorable to plaintiffs, with 50.7% classified as *procedente* (granted), indicating strong legal grounds for the claims. Geospatial analysis revealed that São Paulo is the jurisdiction with the highest litigation volume (20.5%), followed by other key cities.

3.2 Preprocessing

The preprocessing stage is essential for preparing legal documents for topic modeling by standardizing textual content and reducing noise. The process begins with text normalization, which includes converting all text to lowercase, removing punctuation, and eliminating extra whitespace. In addition, numerical values are discarded because they do not contribute to the semantic meaning in topic modeling.

To enhance linguistic coherence, tokenization is performed, segmenting text into individual words while preserving their structural integrity. The lemmatization is then applied to reduce words to their base forms, ensuring that vari-

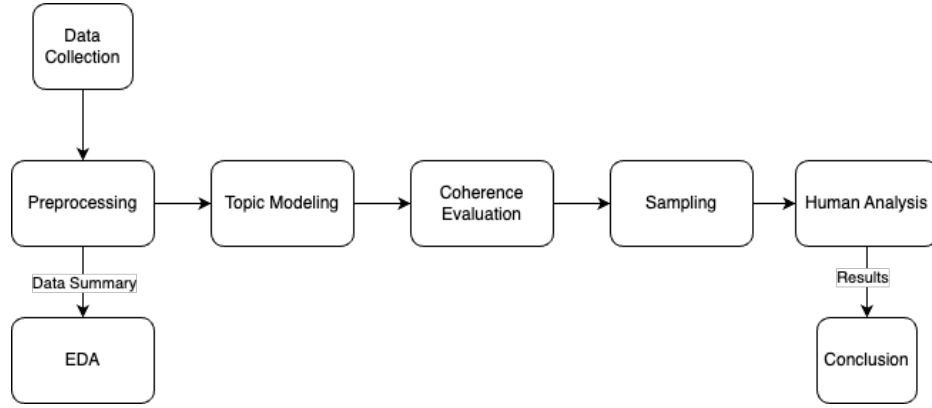


Figure 2. Workflow diagram of the methodology

ations of the same term are treated consistently. In addition, a stopword removal step is performed to remove frequent but semantically insignificant words, such as articles, prepositions, and auxiliary verbs.

Given the specificity of legal language, a custom list of stopwords was curated to exclude common legal procedural terms that do not contribute to topic differentiation. This step ensures that the analysis focuses on the substantive content of legal texts rather than procedural artifacts. The final preprocessed text is structured as a sequence of meaningful terms, forming the foundation for efficient and interpretable topic modeling.

3.3 Topic Modeling

In this study, three widely recognized topic modeling methods were employed to analyze the dataset: LDA [Blei *et al.*, 2003], LSA [Steyvers and Griffiths, 2006], and pLSA [Hofmann, 1999]. Each of these techniques provides a unique approach to modeling topics, balancing interpretability and computational efficiency.

To ensure robustness, topic models were evaluated over a range of topic numbers, and their coherence scores were compared. The dataset was first transformed into a vector representation, utilizing a term-frequency and inverse document frequency (TF-IDF) approach for LSA and pLSA models. LDA, in contrast, directly modeled the word distributions from a bag-of-words representation. After training the models, topic-word distributions were extracted to identify the dominant themes in the corpus, and documents were assigned to topics based on their highest probability scores. The best-performing models were selected based on their coherence scores, which measure how semantically interpretable the topics are.

Each model's performance depends critically on the tuning of key hyperparameters that influence topic granularity and coherence. For LDA, the primary hyperparameters include the number of topics, the Dirichlet priors for document-topic distributions (alpha), and topic-word distributions (beta), which control sparsity and topic overlap. LSA relies on the number of singular values retained during singular value decomposition, affecting the dimensionality reduction and semantic space representation. pLSA's main hyperparameter is the number of topics, which governs the probabilistic mixture components modeling the corpus. Careful selection and optimization of these parameters are essen-

tial to balance model interpretability and computational efficiency, as highlighted in recent systematic reviews [Chen *et al.*, 2023].

3.4 Coherence Score

The coherence score is a key metric for evaluating the quality of topics generated in topic modeling. Among various coherence measures, the c_v coherence metric, introduced by Röder, Both, and Hinneburg [Röder *et al.*, 2015], has demonstrated strong alignment with human interpretability.

This metric leverages Normalized Pointwise Mutual Information (NPMI) and cosine similarity to assess the semantic consistency among the top words within each topic. The c_v score is computed as (11):

$$C_v = \frac{1}{|P|} \sum_{(w_i, w_j) \in P} \text{cosine}(\text{NPMI}(w_i, w_j)) \quad (11)$$

where P represents word pairs, and NPMI captures the co-occurrence relationships. A higher coherence score indicates better semantic consistency, ensuring that extracted topics are meaningful and well-clustered.

To evaluate model performance, coherence scores were calculated for LDA, LSA, and pLSA across multiple topic numbers. The results were analyzed to determine the optimal number of topics for each model.

The best-performing models were selected based on the highest coherence score and minimal spread, optimizing both accuracy and consistency in legal text classification.

3.5 Sampling Strategy

To ensure a representative dataset for analysis, we applied a sampling process based on topic distribution. The dataset was first filtered to remove irrelevant or duplicate cases. The documents were then stratified by their assigned topic to maintain a proportional representation. This process allowed us to include a balanced selection of legal cases across multiple themes, ensuring a fair evaluation of the performance of the topic modeling. The final sample was structured to include examples from different legal domains, maximizing the reliability of the insights derived from the model.

3.6 Human Analysis of Topics

A human interpretability analysis was performed to validate the coherence and relevance of topics assigned by the mod-

els. For each selected model (LDA, LSA, pLSA), the most representative words per topic were extracted and examined, with five (5) samples analyzed per model across a total of 45 samples from the corpus of 4,259 documents.

The human analysis was conducted by the authors themselves, chosen for reasons of convenience and as a proof of concept. Given the exploratory scope of the study, we consider expert validation unnecessary at this stage, as the primary objective is to demonstrate the methodological feasibility and interpretability of topic modeling in the legal domain.

Additionally, the distribution of topics across documents was analyzed to ensure that the dominant themes aligned with the legal contexts. This qualitative assessment complemented the quantitative coherence scores, reinforcing the reliability of the topic modeling results.

To assess classification accuracy, randomly selected documents were manually reviewed along with their assigned topics. The evaluation compared the extracted keywords with the original text to determine whether topic assignments were semantically appropriate.

Misclassified cases were identified, which guided refinements in preprocessing and model selection. In addition, the best models were determined based on coherence scores and topic stability, ensuring that the selected approach provided meaningful information on legal document classification.

3.6.1 Document Analysis Process

Human analysis followed a systematic approach to evaluate topic coherence, as illustrated in the following example.

Document Snippet:

"...ingressou com o presente incidente de cumprimento de sentença, visando que o requerido restabeleça o fornecimento de professor auxiliar, conforme fixado em sentença, sob pena de arbitramento de multa... os serviços de apoio especializado, na escola regular, para atender às peculiaridades da clientela de educação especial, estão previstos legalmente ao menos desde 1996 (Lei nº 9.394/96, art. 58)..."

Topic Keywords: (Result topics from the method) deficiência, especializar, transporte, pessoa, professor, criança, auxiliar, portador, educação, social

Analysis Steps:

Keyword Identification: Key terms in the document matching topic keywords were identified: "professor auxiliar" (matches "professor" and "auxiliar"), "apoio especializado" (relates to "especializar"), "educação especial" (matches "educação"). **Thematic Alignment:** The document's main subject—specialized educational support services—aligns with Topic's focus on special education and support for people with disabilities. **Legal Framework Assessment:** The document refers to relevant legislation (Lei nº 9.394/96) establishing the right to specialized educational support, reinforcing the connection to the Topic. **Contextual Analysis:** The broader context involves a legal proceeding seeking enforcement of a court order to provide an auxiliary teacher, directly relating to educational support for students with special needs. **Conclusion:** Based on the strong presence of multiple topic keywords and thematic alignment, the document was correctly classified under a Topic, with a determination of "Adequate: Yes."

4 Results

The results demonstrate the ability of topic modeling to reveal meaningful patterns in legal texts, aiding in the classification and retrieval of judicial information.

4.1 Coherence Score Analysis

According to Figure 3, in the lower range of two to five topics, the probabilistic framework of pLSA consistently achieves higher coherence, indicating that it effectively captures broader thematic differences with fewer clusters. Approximately five or six topics, the "minimum spread" scenario emerges, in which LDA, LSA, and pLSA converge toward similar coherence levels, suggesting that most methods begin to stabilize and produce coherent topic structures.

As shown in Table 2, this configuration results in different document distributions, with LSA concentrating 2,778 documents on a single topic (Topic 1) while pLSA distributes documents more evenly across topics related to educational rights (Topic 2) and contractual disputes (Topic 4). As the number of topics climbs beyond ten, pLSA tends to maintain a slight lead, although LDA narrows the gap and remains highly competitive in distinguishing nuanced legal themes. Meanwhile, LSA shows moderate coherence throughout and presents fewer improvements from additional topics.

By approximately fourteen or fifteen topics, the "best average" zone appears, in which the overall coherence reaches its peak, reflecting finer-grained semantic clarity in the topics. Table 5 illustrates this increased granularity, with LDA effectively separating educational access topics (Topics 0, 1, 7) from procedural matters (Topics 10, 11) and contractual disputes (Topics 6, 13). Overall, while pLSA excels at smaller dimensionalities and LDA holds strong at higher ones, each approach can yield coherent results once the chosen number of topics strikes a balance between thematic granularity and the complexity of the legal documents.

4.2 Human Analysis

4.2.1 Individual Results

Across the individual analyses of LDA, LSA, and pLSA, most documents were coherently assigned to thematically consistent topics, although with varying degrees of accuracy. LDA with 15 topics distinguished clearer clusters related to children's access to education (with keywords like "vaga em creche" [daycare vacancy], "ensino fundamental" [elementary education], and "atendimento especializado" [specialized care]) and contractual or indemnification disputes (with prominent terms including "danos morais" [moral damages], "documentos" [documents], and "pagamento" [payment]).

Human analysis of sample documents revealed that LDA effectively sampled cases involving access to public education from those centered on contractual breaches in private educational institutions. However, occasional misclassifications appeared in certain documents, for instance, some discussing inadimplência (non-payment) in post-graduate courses were grouped with "dano moral" topics, even though indemnification claims were not the central issue.

These misclassifications typically occurred when legal documents contained multiple interrelated claims with overlapping terminology. As shown in Table 1, LDA achieved 80% manual coherence with a 20% error rate across 741 doc-

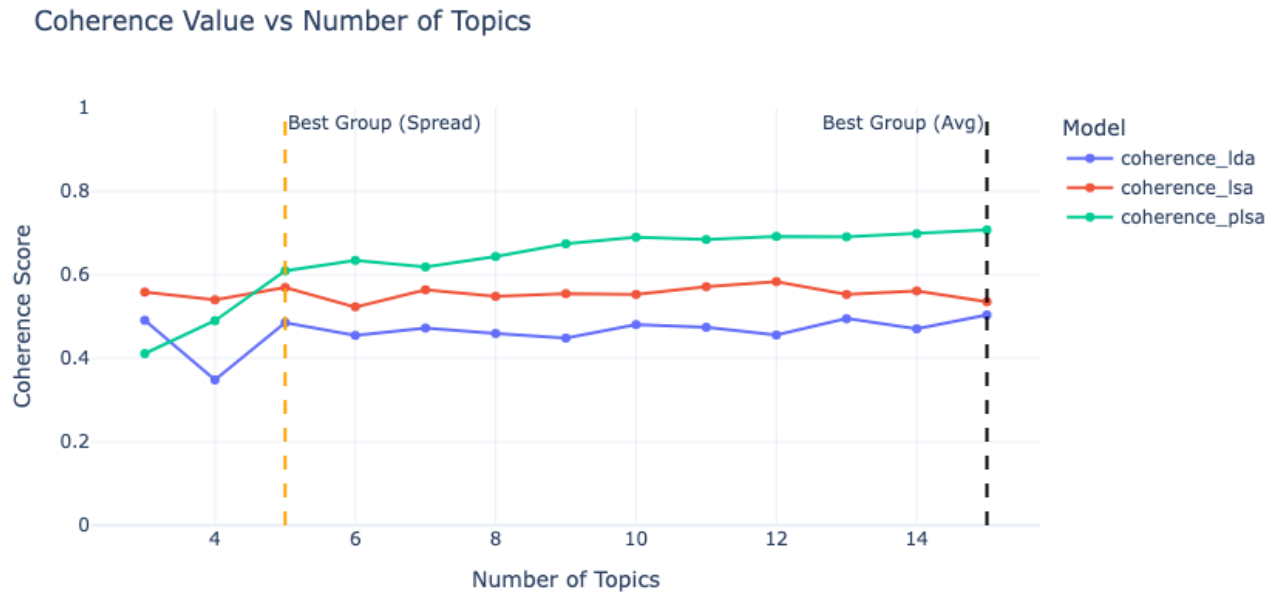


Figure 3. Comparison of coherence scores across topic modeling algorithms (LDA, LSA, and pLSA) with varying numbers of topics. The vertical dashed lines highlight two critical thresholds: the left line at 5-6 topics indicates the "minimum spread" point where all algorithms converge to similar coherence values, while the right line at 14-15 topics marks the "best average" zone where maximum coherence is achieved across all methods.

uments in its main topic, reflecting its generally strong but imperfect discrimination between related legal concepts.

In pLSA, documents focusing on specialized educational support (involving cases with "Síndrome de Down" [Down Syndrome] and requests for "professor auxiliar" [auxiliary teacher]) typically clustered coherently, reflecting appropriate matching with keywords such as "deficiência" [disability], "atendimento" [service], and "necessidade" [necessity].

Human interpretation of these topics revealed that pLSA excelled at identifying specialized education needs cases, capturing the nuances between different types of accommodation requests, such as physical adaptations versus specialized teaching support. This model demonstrated the highest performance with a perfect 100% coherence and 0% error rate in its 590 main topic documents, suggesting particular strength in handling educational rights litigation.

LSA, while uncovering consistent patterns, showed the weakest performance (60% coherence, 40% error rate) of the three models, tending to merge different forms of contract disputes into a single dimension. The manual review identified that LSA frequently combined conceptually distinct legal issues, such as tuition disputes and educational negligence claims, capturing broad semantic relationships without fine-tuning them to the specific legal context. This limitation became particularly evident in cases with complex factual backgrounds involving multiple legal theories, where LSA struggled to differentiate between primary and secondary legal issues presented in the documents.

4.2.2 Grouped Model Analysis

When examining the two optimized topic configurations, the models showed distinct performance patterns that highlight their algorithmic strengths and weaknesses. In the "best

average" configuration (15 topics), LDA excelled at separating well-defined themes such as "educational access," "specialized educational support," "civil procedure/financial disputes," and "document-related indemnifications," achieving perfect 100% coherence as shown in Table 3. This superior performance can be attributed to LDA's probabilistic framework that effectively leverages Dirichlet priors to model word distributions within topics and topic distributions within documents.

Human analysis revealed that the LDA topics were not only statistically coherent but also legally meaningful, with a clear separation between procedural matters (featuring keywords like "prazo" [deadline], "intimação" [notification], and "recurso" [appeal]) and substantive legal disputes (with terms like "contrato" [contract], "mensalidade" [monthly fee], and "prestação de serviço" [service provision]). This granular differentiation proved particularly valuable for distinguishing between cases with similar fact patterns but different legal theories.

In contrast, LSA struggled significantly with this granular topic structure, showing only 20% coherence and an 80% error rate in the 15-topic configuration. This poor performance stemmed from LSA's tendency to combine unrelated legal proceedings, such as mixing purely procedural appeals with cases centered on special education. Manual inspection of the misclassified documents revealed that LSA's singular value decomposition approach, while mathematically elegant, failed to capture the contextual relationships critical to legal document classification.

For example, LSA frequently grouped documents containing terms such as "escola" [school] together, regardless of whether they concerned administrative proceedings, contract disputes, or educational rights litigation. pLSA maintained strong performance (80% coherence) at 15 topics but

Table 1. Error Rates Based on Human Analysis - Individual Models

Model	Topics	Main Topic	Manual Coherence	Error Rate
LDA	15	Topic 1 (741 docs)	80%	20%
LSA	12	Topic 1 (1984 docs)	60%	40%
pLSA	15	Topics 0 & 13 (590 docs)	100%	0%

Table 2. Distribution of Documents by Topic (5 Topics)

Model	Topic	Top 3 Keywords	Documents
LDA	Topic 0	curso, dano, diploma	983
	Topic 1	ensino, educação, professor	565
	Topic 2	criança, ensino, fundamental	1100
	Topic 3	fls, transporte, deficiência	799
	Topic 4	segurança, criança, educação	812
LSA	Topic 0	aposentadoria, Grande, guarulho	590
	Topic 1	criança, residência, cabreúva	2778
	Topic 2	cabreúva, residência, lesão	0
	Topic 3	morato, Francisco, creche	189
	Topic 4	diploma, morato, dano	702
pLSA	Topic 0	judicio, prôtês, propugnar	590
	Topic 1	criança, cabreúva, residência	433
	Topic 2	deficiência, especializar, transporte	1706
	Topic 3	morato, Francisco, creche	224
	Topic 4	diploma, dano, curso	1306

Table 3. Error Rates Based on Human Analysis - Best Average Coherence (15 Topics)

Model	Main Topic	Manual Coherence	Error Rate
LDA	Topic 1 (741 docs)	100%	0%
LSA	Topic 1 (1931 docs)	20%	80%
pLSA	Topics 0 & 13 (590 docs)	80%	20%

Table 4. Error Rates Based on Human Analysis - Minimum Spread Configuration (5 Topics)

Model	Main Topic	Manual Coherence	Error Rate
LDA	Topic 2 (1100 docs)	80%	20%
LSA	Topic 1 (2778 docs)	0%	100%
pLSA	Topic 2 (1706 docs)	100%	0%

fell short of its individual model results, occasionally merging distinct subcategories of educational disputes that human analysts would typically separate.

In the "minimum spread" configuration (5 topics) shown in Table 4, pLSA demonstrated perfect coherence with broader topic categories, successfully clustering 1,706 documents with 100% accuracy. This exceptional performance suggests that pLSA's probabilistic approach to document-topic and topic-word distributions works particularly well when modeling broader legal themes rather than fine-grained distinctions.

The manual review confirmed that pLSA accurately grouped documents into fundamental categories such as "educational access rights" and "breach of educational services contracts" without conflating these distinct legal domains. LSA's performance deteriorated completely in this configuration, failing to produce any coherent topics (0% coherence) even with broader categories, reflecting fundamental limitations in its ability to model complex legal language. LSA's term-frequency matrix approach appeared to be confounded by the legal corpus's specialized vocabulary and formulaic language patterns, creating topical clusters that lacked legal coherence.

LDA maintained consistent performance (80% coherence) in both configurations, demonstrating the stability of its probabilistic structure and Dirichlet priors in producing interpretable clusters, particularly when capturing nuanced distinctions among documents involving varied educational and contractual disputes. The consistency of LDA across dif-

ferent topic numbers suggests a robust modeling approach that adapts well to different levels of thematic granularity in legal corpora, making it particularly suitable for jurimetric applications where flexibility and reliability are essential.

5 Conclusion

In conclusion, the study demonstrates the feasibility and value of applying topic modeling to Brazilian legal documents in order to organize and extract thematic patterns. Among the three techniques examined—LDA, LSA, and pLSA—each exhibited distinct advantages and limitations, with notable differences between automated coherence metrics and human evaluation. As shown in Figure 3, pLSA consistently maintained the highest mathematical coherence scores in the most topic configurations, including 15 topics. However, our human analysis revealed a more nuanced picture: LDA achieved perfect human-judged coherence (100%) in the 15-topic configuration, while pLSA reached 80% coherence at this granularity despite its superior automated scores.

This discrepancy highlights an important consideration in topic modeling evaluation—mathematical coherence measures don't always align perfectly with human judgments of topic quality in specialized domains like legal text. pLSA demonstrated remarkable versatility, achieving perfect human-evaluated coherence (100%) in the minimum spread configuration with 5 topics and maintaining strong performance with finer-grained topics. LDA showed excel-

Table 5. Distribution of Documents by Topic (15 Topics)

Model	Topic	Top 3 Keywords	Documents
LDA	Topic 0	educação, fundamental, município	692
	Topic 1	deficiência, pessoa, educação	741
	Topic 2	ensino, segurança, Paulo	466
	Topic 3	curso, ensino, fls	421
	Topic 4	Porto, feliz, fundeb	0
	Topic 5	magistério, aposentadoria, Francisco	115
	Topic 6	dano, diploma, moral	418
	Topic 7	criança, ensino, residência	414
	Topic 8	criança, ano, idade	313
	Topic 9	educação, vaga, município	221
	Topic 10	fls, valor, Paulo	169
	Topic 11	valor, prazo, dever	78
	Topic 12	transporte, guarulho, fls	145
	Topic 13	fies, financiamento, instituição	26
	Topic 14	educação, licenciatura, aula	40
LSA	Topic 0	fies, uniesp, financiamento	590
	Topic 1	criança, residência, cabreúva	1931
	Topic 2	cabreúva, residência, lesão	0
	Topic 3	morato, Francisco, creche	120
	Topic 4	diploma, morato, dano	605
	Topic 5	lençóis, paulista, Creche	88
	Topic 6	lençóis, paulista, vaga	195
	Topic 7	diploma, transporte, guarulho	4
	Topic 8	morato, Francisco, transporte	71
	Topic 9	transporte, guarulho, curricular	286
	Topic 10	serra, Grande, Rio	66
	Topic 11	cargo, aposentadoria, magistério	179
	Topic 12	aposentadoria, magistério, guarulho	38
	Topic 13	aposentadoria, magistério, ribeirão	44
	Topic 14	guarulho, ribeirão, pir	42
pLSA	Topic 0	judicio, prôtês, propugnar	590
	Topic 1	cabreúva, criança, residência	412
	Topic 2	deficiência, especializar, auxiliar	508
	Topic 3	morato, Francisco, creche	119
	Topic 4	curricular, curso, disciplina	483
	Topic 5	lençóis, paulista, vaga	206
	Topic 6	paulínia, programático, eficácia	99
	Topic 7	diploma, dano, moral	402
	Topic 8	morato, Francisco, impetrada	75
	Topic 9	transporte, guarulho, saúde	280
	Topic 10	serra, Grande, Rio	67
	Topic 11	cargo, concurso, edital	176
	Topic 12	aposentadoria, magistério, assessoramento	55
	Topic 13	médio, curso, vestibular	590
	Topic 14	ribeirão, criança, pir	197

lent human-interpretable boundaries between legal concepts at higher dimensionalities, successfully separating procedural matters from substantive disputes despite its relatively modest automated coherence scores. LSA consistently underperformed in human evaluation (reaching only 60% coherence in individual analysis and completely failing with 0% coherence in the minimum spread configuration) despite its moderate mathematical coherence.

The coherence score analyses affirmed the importance of choosing an optimal number of topics: around five or six to minimize inter-model variability (minimum spread) and fifteen to maximize average coherence. Our human analysis validated these configurations while demonstrating that topic models should be evaluated using both quantitative metrics and qualitative legal expertise. By revealing hidden structures in legal texts—ranging from children’s educational access to specialized support cases and civil procedure disputes—our findings highlight how data-driven techniques can bring clarity and efficiency to the legal domain while maintaining contextual relevance.

For practical jurimetric applications, our results suggest that while pLSA may show superior mathematical coherence, LDA offers more legally meaningful distinctions at higher topic granularity—a critical consideration for legal analytics where both broad classification and fine-grained distinction may be required. Future work might explore hybrid or neural-based topic modeling approaches and incorporate domain-specific legal knowledge bases for enhanced interpretability. Ultimately, the results underscore the promise of AI-assisted analytics in supporting legal practitioners, courts, and policy makers by enabling scalable, transparent, and context-aware exploration of complex judicial records, with both quantitative and qualitative evaluation being essential to selecting the most appropriate topic modeling approach.

Future research will benefit from the participation of a multidisciplinary expert team, particularly legal professionals, to conduct a more rigorous human interpretability analysis. While this study served as a proof of concept, expert evaluation will provide deeper validation of topic coherence and legal relevance. Additionally, we propose the integration

of Explainable Artificial Intelligence (XAI) techniques, such as Topic Embeddings, to enhance transparency and semantic traceability in topic modeling outputs.

Declarations

Acknowledgements

This research was supported by the University of São Paulo undergraduate grant.

Authors' Contributions

EBS is the main contributor at this paper, and contributed to the scripts development and manuscript writing. IAR contributed to the conception, research supervision, manuscript writing and revision. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests

Availability of data and materials

The datasets and/or scripts generated and/or analyzed during the current study are available upon request and can also be accessed at the GitHub repository: <https://github.com/Etore-BeS/interpreting-lawsuits-topic-modeling>.

References

- Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77–84. DOI: 10.1145/2133806.2133826.
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022. Available at: <https://dl.acm.org/doi/10.5555/944919.944937>.
- Chen, Y., Peng, Z., Kim, S.-H., and Choi, C. W. (2023). What we can do and cannot do with topic modeling: A systematic review. *Communication Methods and Measures*, 17(2):1–20. DOI: 10.1080/19312458.2023.2167965.
- Devins, N., Levine, R., Liptak, A., and Bhatia, K. S. (2017). The law and big data. *Cornell Journal of Law and Public Policy*, 27(2):357–401. Available at: <https://scholarship.law.cornell.edu/cjlp/vol27/iss2/3/>.
- Garg, A. and Ma, M. (2025). Opportunities and challenges in legal ai. Technical report, Stanford Law School. White Paper, CodeX. Available at: <https://law.stanford.edu/publications/opportunities-and-challenges-in-legal-ai/>.
- Griffiths, T. L., Steyvers, M., Blei, D. M., and Tenenbaum, J. B. (2004). Integrating topics and syntax. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'04, page 537–544, Cambridge, MA, USA. MIT Press. Available at: <https://dl.acm.org/doi/10.5555/2976040.2976108>.
- Hofmann, T. (1999). Probabilistic latent semantic analysis. UAI'99, page 289–296, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc. Available at: <https://dl.acm.org/doi/10.5555/2073796.2073829>.
- Jones, K. S. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21. DOI: 10.1108/eb026526.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86. DOI: 10.1214/aoms/1177729694.
- Ma, M., Sinha, A., Tandon, A., and Richards, J. (2024). Generative ai legal landscape 2024. Technical report, Stanford Law School. White Paper, CodeX. Available at: <https://law.stanford.edu/publications/generative-ai-legal-landscape-2024-2/>.
- Mehrotra, R., Sanner, S., Buntine, W., and Xie, L. (2013). Improving lda topic models for microblogs via tweet pooling and automatic labeling. SIGIR '13, page 889–892, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/2484028.2484166.
- Meng, X.-L. and van Dyk, D. (1997). The em algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):511–567. Read before The Royal Statistical Society at a meeting organized by the Research Section on December 11, 1996, Professor P. J. Green in the Chair. Available at: <http://www.jstor.org/stable/2346009>.
- Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, page 262–272, USA. Association for Computational Linguistics. Available at: <https://dl.acm.org/doi/10.5555/2145432.2145462>.
- Naskar, D., Mokaddem, S., Rebollo, M., and Onaindia, E. (2021). Sentiment analysis in social networks through topic modeling. *Proceedings of the the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Available at: <http://www.lrec-conf.org/proceedings/lrec2016/summaries/682.html>.
- Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, page 248–256, USA. Association for Computational Linguistics. Available at: <https://dl.acm.org/doi/10.5555/1699510.1699543>.
- Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 399–408. ACM. DOI: 10.1145/2684822.2685324.
- Steyvers, M. and Griffiths, T. (2006). Probabilistic topic models. In Landauer, T., McNamara, D., Dennis, S., and Kintsch, W., editors, *Latent Semantic Analysis: A Road to Meaning*, pages 427–448. Erlbaum. Available at: <https://scholar.google.com/scholar?cluster=8269087903800927658>.
- Wang, Y. (2008). Distributed gibbs sampling of latent dirichlet allocation: The gritty details. Accessed: January 16, 2025. Available at: <https://www.scribd.com/document/383010076/Lda>.
- Zödi, Z. (2017). Law and legal science in the age of big data. *Intersections. East European Journal of Society and Politics*, 3(2). DOI: 10.17356/ieejsp.v3i2.324.