Discovery of Legal Patterns in Civil Petitions via LLM-Based Fact Extraction and Density Clustering

Rhedson Esashika
PPGEEL – Postgraduate Program in Electrical Engineering
State University of Amazonas (UEA)
Manaus, AM, Brazil
rffe.mee25@uea.edu.br

Carlos M. S. Figueiredo
LSI – Intelligent Systems Laboratory
State University of Amazonas (UEA)
Manaus, AM, Brazil
cfigueiredo@uea.edu.br

Tiago de Melo
PPGEEL – Postgraduate Program in Electrical Engineering
State University of Amazonas (UEA)
Manaus, AM, Brazil
tmelo@uea.edu.br

Abstract

The analysis of unstructured civil petitions is often hindered by procedural noise and verbose argumentation. To address this, we propose a pipeline composed of LLM-based fact extraction followed by legal-domain embeddings of texts for unsupervised density clustering. We employ Large Language Models to isolate factual narratives from raw texts, which are then encoded using domain-specific representations (Legal-BERT) and grouped via UMAP dimensionality reduction and the HDBSCAN algorithm. Comparative experiments on a Brazilian judicial corpus reveal that clustering based solely on extracted *Factual Segments* yields significantly more cohesive and semantically well-defined groups than *Full-text petitions*, which suffer from fragmentation due to content variability. Results indicate that the proposed method is a promising approach for thematic organization, procedural triage support, and large-scale discovery of legal patterns.

1 Introduction

The rapid expansion of judicial digitization has produced vast amounts of text, especially in initial petitions, decisions, and judgments. However, the sheer volume of documents, along with the complexity of legal language, makes traditional analytical approaches difficult to scale. In response, Natural Language Processing (NLP) techniques, particularly unsupervised learning, have become key tools for organizing legal texts semantically and uncovering recurring patterns.

Text clustering enables the grouping of documents according to semantic similarities, revealing recurrent themes and argumentative structures. This technique can support tasks such as procedural triage and large-scale discovery of legal patterns. The problem is that the technical and heterogeneous nature of legal language requires models capable of accurately representing the linguistic context. Consequently, there has been a shift from classical representations, such as TF-IDF, to contextual embeddings based on Transformer architectures. Specifically, domain-specific models such as Legal-BERT (Chalkidis et al., 2020) have demonstrated superior performance by capturing legal semantic nuances and contextual relationships with greater fidelity than generic models. Additionally, legal documents tend to be verbose, which introduces noise in text representations, making the clustering task more difficult.

To effectively navigate the high-dimensional space of these embeddings, modern pipelines often employ dimensionality reduction coupled with density-based clustering. The combination of Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) has proven effective for complex datasets (Blanco-Portals et al., 2022). HDBSCAN is particularly distinguished by its ability to identify dense regions without requiring a predefined number of clusters and its robustness in handling noise and arbitrarily shaped distributions (Moulavi et al., 2014), highly desirable characteristics for the irregular topology of legal corpora.

The work by Oliveira and Sperandio Nascimento (2021) confirmed the feasibility of clustering legal documents, establishing a baseline with traditional methods. Recent advances, such as Esashika et al. (2025), explore the use of Large Language Models (LLMs) to structure petitions, highlighting a movement toward integrating classical NLP with generative artificial intelligence. Works such as Marulli et al. (2025) have applied the UMAP and HDBSCAN pipelines to judicial decisions, showing the potential of these approaches.

However, applying clustering directly to raw legal texts often introduces noise due to procedural scaffolding and verbose argumentation. To address this, this study proposes a hybrid approach for dis-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 416–424

covering legal patterns in initial petitions. We introduce a pipeline that first leverages LLMs to extract *Factual Narratives*, isolating the core events of the dispute, before applying density-based clustering. This research evaluates the impact of this extraction step by comparing the semantic organization of *Full-text* versus *Factual-only* embeddings, aiming to unveil the natural thematic structure of petitions without prior labels. Results show that factual extraction and representation with Legal-BERT leads to more homogeneous and complete clusters. As a further contribution to the field of Legal NLP, we release the curated corpus of anonymized petitions used in these experiments.

The paper is organized as follows: section 2 presents related work, section 3 describes the proposed methodology, section 4 presents and discuss the results, and section 5 concludes the work and point to future work.

2 Related Work

Recent literature emphasizes that applying NLP directly to full-text legal documents often yields suboptimal results due to noise and excessive context length. Held and Habernal (2025) demonstrated that providing full judgment documents to Large Language Models (LLMs) significantly degrades performance compared to paragraph-level processing, validating the need for targeted content selection. To address this, hybrid architectures have gained traction. For example, Nan et al. (2024) proposed a "select-then-extract" framework for enforcement decisions, using rule-based filters to identify candidate segments before invoking LLMs. This approach mirrors our strategy of isolating *Factual Narratives*, ensuring that downstream tasks rely on dense, hallucination-free representations. In the Brazilian portuguese context, Esashika et al. (2025) advanced this direction by employing generative AI to structure initial petitions, highlighting the potential of integrating LLMs into traditional legal analysis workflows.

Regarding the unsupervised organization of these documents, the combination of semantic embeddings, Uniform Manifold Approximation and Projection (UMAP), and Hierarchical Density-Based Spatial Clustering (HDBSCAN) (McInnes et al., 2017) has emerged as a robust framework. Bastola and Choi (2025) introduced a pipeline integrating Top2Vec and graph embeddings with UMAP-HDBSCAN, achieving superior cluster coherence compared to LDA or NMF. Similarly, Marulli et al. (2025) applied a BERTopic-based pipeline (which internally leverages UMAP and HDBSCAN) to Italian Supreme Court rulings, reinforcing the method's interpretability in judicial contexts. Finally, at the national level, Oliveira and Sperandio Nascimento (2021) established a baseline for clustering Brazilian legal documents using TF-IDF and K-Means. While their study predates current density-based approaches, it provides a critical benchmark for evaluating the semantic gains proposed in our research.

In summary, while the aforementioned studies address isolated components of the legal NLP pipeline—focusing either on structured extraction (Nan et al., 2024; Esashika et al., 2025) or unsupervised clustering mechanisms (Bastola and Choi, 2025; Marulli et al., 2025)—our work proposes a unified framework that bridges these domains. Unlike Oliveira and Sperandio Nascimento (2021), who rely on centroid-based methods (K-Means), we adopt density-based clustering to capture the non-linear semantic topology of legal disputes. Furthermore, we advance beyond the scope of general topic modeling by explicitly investigating the impact of *factual isolation* on clustering quality. By benchmarking *Factual Narratives* against *Full-text* representations, this study fills a critical gap, empirically demonstrating that mitigating procedural noise via LLMs is a prerequisite for effective pattern discovery in unstructured Brazilian civil petitions.

3 Methodology

The methodological framework adopted in this study follows a sequential pipeline designed to uncover latent semantic patterns in legal documents. As illustrated in Figure 1, the workflow is composed of five distinct stages: (1) Dataset Acquisition, involving the collection and anonymization of initial civil petitions from the Court of Justice of Amazonas (TJAM); (2) Fact Extraction, where Large Language Models (LLMs) are employed to isolate factual narratives from the full text; (3) Embedding Generation, which utilizes distinct architectures—Legal-BERT, OpenAI, and Gemini—to encode semantic content; (4) the Clustering Pipeline, which integrates UMAP dimensionality reduction with HDBSCAN density-based grouping; and (5) Evaluation, comprising internal quality metrics (e.g., V-Measure) and external vali-

dation by legal experts.

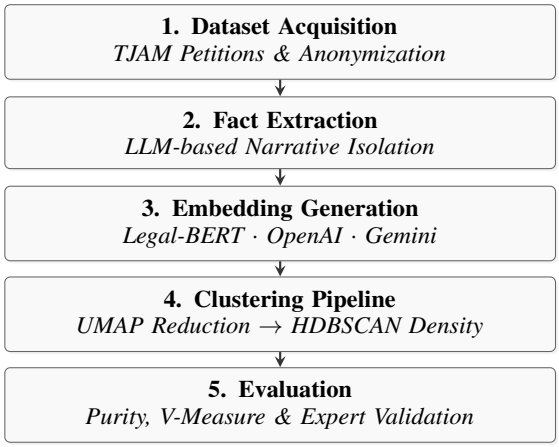


Figure 1: Methodological pipeline.

3.1 Dataset Acquisition and Preprocessing

The dataset used in this study consists of initial procedural petitions obtained from the Court of Justice of the State of Amazonas (TJAM). The collection focused on cases filed in January 2025, resulting in a total of 1,860 documents. The dataset has 930 petitions from the Small Claims Civil Courts and 930 from the Ordinary Civil Courts.

This division ensures the representation of different levels of legal formalism and textual complexity. Ordinary Courts typically handle complex litigation requiring extensive argumentation and strict adherence to procedural codes. In contrast, Small Claims Courts operate under the principles of simplicity and celerity (Law No. 9.099/95), often resulting in more concise, objective, and less dogmatic petitions. This stylistic heterogeneity within the same substantive domain (Civil Law) poses a significant challenge for clustering models, as they must distinguish semantic patterns despite the variation in verbosity and structural scaffolding.

After collection, all petitions underwent textual normalization and anonymization procedures, in accordance with Law No. 13.709/2018 (Brazilian General Data Protection Law — LGPD) (Brasil, 2018) and Resolution CNJ No. 615/2025 (de Justiça , CNJ), which regulates the use of AI in the Judiciary. The anonymization was performed as an integral component of the extraction pipeline developed by Esashika et al. (2025). Through specific prompt engineering techniques validated in their framework, the LLMs were instructed to act as filters, identifying and replacing sensitive personal data (such as names of parties and attorneys) with generic placeholders during the generation of the factual narratives, thus ensuring privacy by design.

The resulting corpus provides a sample of full-text petitions, comprising the complete raw petition including procedural headers and legal argumentation. Petitions presented a global average length of 2,576 words (max: 6,070). This aggregate metric reflects the structural heterogeneity of the dataset, encompassing both the typically concise petitions from Small Claims Courts and the more verbose filings from Ordinary Courts.

To establish a ground truth for quantitative evaluation, the legal cases were manually categorized by judicial experts into fifteen thematic classes representing the nature of the dispute. As detailed in Table 1, the dataset exhibits a long-tail distribution, characteristic of real-world judicial archives. Categories such as Unauthorized Discounts and Service Failures (Labels 0 and 1) comprise a significant portion of the corpus, while specific issues like Defective Products appear as rare events. This imbalance poses a realistic challenge for density-based clustering algorithms, testing their ability to distinguish sparse clusters from noise.

3.2 LLM-based Fact Extraction

The procedure for extracting facts from initial petitions was developed based on the methodological workflow validated by Esashika et al.'s (2025), in which Large Language Models (LLMs) were applied to structure legal information contained in Brazilian legal documents. The core process involved guiding the LLM with a carefully designed prompt to identify and isolate factual elements within the text, distinguishing them from legal arguments or claims. The prompt incorporated semantic instructions inspired by persona prompting strategies, directing the model to act as a legal analyst capable of recognizing the argumentative structure typical of Brazilian petitions. This configuration aimed to capture the "facts" in their procedural sense, that is, the narrative of events that substantiate the claim, while preserving textual coherence and minimizing redundancy.

3.3 Semantic Embedding Models

For each document in the dataset, a 768-dimensional embedding vector was generated to represent its semantic content. Three different models were employed for this task to enable comparative analysis of general-purpose and domain-

Thematic Class (Category)	Label ID	Count (N)
Unauthorized Bank Deductions (Checking Account)	0	499
Moral Damages (Banking Service Failure)	1	383
PASEP Balance Revision	2	8
Judicial Order for Fund Withdrawal (Alvará)	3	1
Unsolicited Consignment Credit Card	4	52
Others	5	771
Undue Credit Card Billing	6	38
Utility Billing Error (Amazonas Energia)	7	49
Undue Negative Credit Report (Banking)	8	26
Consignment Loan (Non-Contracted Allegation)	9	9
Undue Consignment Loan Refinancing	10	7
Undue Negative Credit Report (Telephony)	11	14
Defective Product (Exchange Refusal)	12	1
Undue Billing (Non-Contracted Phone Plan)	13	1
Banking Contract Revision	14	1
Total	-	**1.860**

Table 1: Distribution of ground truth categories established by judicial analysts, mapped to English labels.

specific representations. The first model, Legal-BERT-base-uncased (Chalkidis et al., 2020), was selected for its pretraining on legal corpora, enabling the capture of domain-specific terminology and argumentation patterns. The second, OpenAI text-embedding-3-large [1], is a popular general embedding model designed to encode diverse textual contexts with high precision. Finally, Google gemini-embedding-001 [2] was used to obtain text embeddings from a large language model, allowing the evaluation of its capacity to generalize across legal factual narratives.

To ensure consistency and comparability across the different embedding architectures, all generated vectors were subjected to L_2 normalization. This process projects the embeddings onto a unit hypersphere, ensuring that the dot product between any two vectors corresponds directly to their cosine similarity—the metric subsequently used in the dimensionality reduction and clustering stages. Formally, each embedding vector v was transformed into $\hat{v}$ as follows:

$$\hat{v} = \frac{v}{\|v\|_2} = \frac{v}{\sqrt{\sum_{i=1}^{d} v_i^2}} \quad (1)$$

where d represents the dimensionality of the vector space ($d = 768$). This standardization mitigates the impact of vector magnitude, focusing the analysis strictly on the semantic orientation of the legal narratives, which provides a more robust foundation for evaluating grouping quality.

[1] https://platform.openai.com/docs/models/text-embedding-3-large

[2] https://ai.google.dev/gemini-api/docs/embeddings?hl=pt-br

3.4 Dimensionality Reduction Setup

To enhance clustering performance, high-dimensional embeddings were projected into a lower-dimensional space using UMAP (McInnes et al., 2018). UMAP was selected because, unlike linear methods (e.g., PCA) or non-linear alternatives like t-SNE, it better preserves both local and global intrinsic manifold structures of dense representations while ensuring computational scalability. The algorithm was configured with cosine distance as the similarity metric, aligning with the semantic nature of the embedding representations.

A grid search was conducted on key hyperparameters — the number of neighbors (n_neighbors), the number of components (n_components), and the minimum distance (min_dist) — was conducted to identify configurations that improved cluster compactness and separation. The values of the selected parameters for each embedding model and the configuration of the dataset are summarized in Table 2. The quality of each projection was evaluated through the trustworthiness metric, ensuring that the reduced representations maintained the essential topological relationships of the original space. These optimized low-dimensional embeddings served as input for the subsequent clustering analysis of legal factual narratives.

3.5 Density-Based Clustering

For semantic grouping, we employed HDBSCAN. Unlike centroid-based methods such as K-Means that force all points into spherical clusters, HDBSCAN is robust to the irregular semantic shapes and long-tail distribution typical of our legal corpus. Furthermore, it improves upon standard DBSCAN

Model	Dataset	UMAP			HDBSCAN	
		n_neighbors	*n_components*	*min_dist*	*min_cluster_size*	*min_samples*
legal-BERT	Facts	100	10	0.1	25	20
legal-BERT	Full	50	100	0.2	30	25
gemini001	Facts	15	50	0.4	20	20
gemini001	Full	15	100	0.2	20	20
text-3-large	Facts	100	10	0.1	30	20

Table 2: Hyperparameters for dimensionality reduction (UMAP) and clustering (HDBSCAN).

by dynamically adapting to varying densities and automatically isolating noise (atypical cases) without requiring a predefined number of clusters.

A systematic exploration of the hyperparameters *min_cluster_size* and *min_samples* was conducted, as summarized in Table 2. These parameters were adjusted for each embedding model and dataset configuration to balance cluster granularity and noise sensitivity. Data points labeled as noise (-1) were excluded from quantitative evaluation but analyzed qualitatively to understand their linguistic or factual distinctiveness.

To quantitatively assess the quality of the generated clusters, we employed the manual classifications described in Section 3.1 as ground truth labels. Specifically, the numerical identifiers ($0-14$) listed in Table 1 served as the reference for computing external validity metrics, ensuring that the density-based groups discovered by the algorithm are evaluated against expert-defined legal categories.

3.6 Evaluation Metrics

Following established protocols in unsupervised legal document analysis (Marulli et al., 2025; Bastola and Choi, 2025), we employed a multi-metric evaluation framework to assess both topological preservation and semantic coherence.

To validate the dimensionality reduction step, we calculated *Trustworthiness*, which relies on a rank-based comparison of k-nearest neighbors to quantify the preservation of local manifold structures from the high-dimensional space.

For clustering quality, we utilized the expert-annotated labels (Section 3.1) as ground truth to compute four extrinsic metrics. We employed two entropy-based measures: *Homogeneity*, which penalizes clusters containing members of different classes (conditional entropy of class distribution given the cluster), and *Completeness*, which assesses if all members of a class are assigned to the same cluster. Their harmonic mean yields the *V-measure*, providing a balanced view of precision and recall. Finally, to account for random grouping effects, we computed the *Adjusted Rand Score (ARS)*, a pair-counting metric that measures the similarity between the algorithmic and expert partitions, corrected for chance.

4 Results and Discussions

4.1 Presentation of Results

The quantitative evaluation of the clustering pipeline is summarized in Table 3. Beyond the raw values, the empirical data supports two primary conclusions: (1) factual extraction is a prerequisite for effective clustering in the legal domain, and (2) domain-specific embeddings outperform generalist models in creating semantically cohesive groups.

The transition from *Facts* to *Full-text* configurations resulted in a performance drop across all semantic metrics for every model. Notably, the V-measure for Legal-BERT dropped from 0.5148 to 0.3781 when full text was used. This degradation supports the assumption that procedural scaffolding and verbose argumentation in raw petitions may act as noise, hindering semantic representation. By isolating the factual narratives, the model appears to achieve a superior signal-to-noise ratio, allowing the clustering algorithm to separate distinct legal disputes based on events rather than original text.

Legal-BERT-base-uncased (Facts) emerged as the superior model for this task, achieving the highest *V-measure* (0.5148) and *Homogeneity* (0.6144). In the context of pattern discovery, Homogeneity is particularly critical as it measures the purity of the clusters. Legal-BERT's high score indicates that it successfully creates cohesive groups dominated by single legal themes.

Although *Gemini-embedding-001* achieved the highest *Adjusted Rand Score (ARS)* of 0.3980, indicating a strong alignment with ground truth labels, it displayed lower homogeneity. This discrepancy suggests that while Gemini excels at broad categorization, as evidenced by its high completeness, it

Metric	Legal-BERT		Gemini		OpenAI-Large
	Facts	*Full*	*Facts*	*Full*	*Facts*
Trustworthiness	0.9554	**0.9671**	0.9625	0.9562	0.9653
Adjusted Rand Score (ARS)	0.3710	0.1882	**0.3980**	0.1833	0.2219
Homogeneity	**0.6144**	0.5309	0.4860	0.4310	0.5362
Completeness	0.4429	0.2935	**0.5327**	0.3118	0.3962
V-measure	**0.5148**	0.3781	0.5083	0.3618	0.4557

Table 3: Average clustering evaluation metrics across embedding models and input configurations.

may obscure distinctions between subtle sub-topics that the domain-specific Legal-BERT identifies.

Regarding the projection quality, all configurations maintained *Trustworthiness* values above 0.95, confirming that the UMAP dimensionality reduction preserved the local topological structure of the original embeddings.

4.2 Discussion of Clustering Performance

The results presented in Table 3 demonstrate distinct performance across the evaluated embedding models and input configurations. The comparison reveals that the model's specialization and the textual scope of input data play decisive roles in the semantic organization and coherence of clusters.

Among all tested embeddings, Legal-BERT-base-uncased (Facts) achieved the most balanced overall performance, outperforming the others in *V-measure* (0.5148) and *Homogeneity* (0.6144). These metrics indicate that domain-specific embeddings trained on legal corpora provide a richer representation of juridical semantics, enabling better grouping of factual narratives into coherent clusters. This result aligns with prior findings in legal NLP literature, which emphasize that contextual embeddings fine-tuned on legal text capture nuanced patterns that generic models often overlook.

This result aligns with prior findings in legal NLP literature (Chalkidis et al., 2020; Bastola and Choi, 2025), which emphasize that contextual embeddings fine-tuned on legal corpora capture nuanced semantic patterns that generic models often overlook. By leveraging representations optimized for juridical syntax, the model effectively distinguishes between subtle factual variations—such as the difference between a "billing error" and a "contractual revision"—which typically conflate in general-purpose vector spaces.

The gemini-embedding-001 (Facts) configuration exhibited the highest *Adjusted Rand Score* (0.3980) and *Completeness* (0.5327), showing stronger agreement between discovered clusters and reference categories. This suggests that the Gemini embeddings, though not domain-specific, maintain competitive discriminative capabilities when the input text is limited to the factual segments. The removal of rhetorical and procedural elements from the full petition text effectively reduced noise, improving the signal-to-semantic ratio of the representations.

In contrast, all Full-text configurations presented a consistent decline in clustering quality. The decrease in both *V-measure* and *Completeness* values suggests that longer and heterogeneous documents introduce vocabulary dispersion and topic overlap, leading to lower cluster separability. This effect highlights the importance of content selection and segmentation in legal document analysis, particularly in tasks driven by unsupervised learning.

Despite model differences, all configurations achieved *Trustworthiness* values above 0.95, confirming that the UMAP projections preserved the intrinsic neighborhood structure of the high-dimensional embedding space. This ensures that observed clusters in the reduced space are faithful to the underlying semantic relationships of the original embeddings.

Overall, the results reinforce two key findings: (i) embeddings specialized in legal language yield superior intra-cluster cohesion, and (ii) focusing on factual segments rather than full documents enhances cluster clarity and interpretability. These outcomes validate the methodological decision to employ factual narratives as the primary representation for unsupervised clustering in legal texts.

4.3 Evaluation of Thematic Homogeneity

In order to evaluate the quality of the clusters produced by HDBSCAN in relation to every document category, as described in Table 1, we calculated the homogeneity of each of them. This homogeneity was defined as the ratio between the number of

documents belonging to the predominant human-assigned class within a cluster and the total number of documents in that cluster, reflecting the extent to which each group is composed of elements from a single reference category. The goal of this stage was to provide an interpretable assessment of the coherence between the automatic clusters and the thematic categories perceived by legal experts.

Figure 2 shows that several thematic clusters exhibit homogeneity levels above 80%, suggesting a strong correspondence between the patterns discovered by the model and the human-defined categories. A few heterogeneous clusters, such those for themes "Others" or particular cases of "Bank Credit Cards", displayed lower values, reflecting greater lexical and thematic dispersion. Particularly, the class "Others" involved a lot of different documents not included in selected categories. These results reinforce the consistency of the structure produced by HDBSCAN and complement the quantitative metrics by introducing an interpretable dimension grounded in human judgment.

4.4 Qualitative Analysis of Clusters

A qualitative inspection of the UMAP projections, obtained by reducing the embeddings to a two-dimensional space ($n\_components = 2$) specifically for visualization purposes, can be seen in Figure 3. This visualization provides a deeper insight into the topological structure of the discovered groupings. We compared the clusters generated from Factual Narratives (Figure 3a) against those derived from Full-text petitions (Figure 3b), where ground truth labels are represented by different colors and different clusters are shown as labeled instances in the figure. In this way, incoherent clusters can be easily identified as multicolored groups of points or by the spatial dispersion of labels.

In the extracted facts case, we observe a structure with well-defined boundaries and high density. By mapping the visual cluster labels to the ground truth classes, we identified highly homogeneous groups. **Cluster 0** (top-left) is almost exclusively composed of *Utility Billing Error (Amazonas Energia) (Class 7, red points)* cases (Class 7, red points), demonstrating perfect isolation of utility disputes. Similarly, **Cluster 3** (bottom-center) groups all *Judicial Order for Fund Withdrawal (Alvará)* (Class 3, light-green). If we note the clusters based on full texts, only Class 3 (light-green) is properly separated in **Cluster 7**, while Class 7 (red) is mixed with other classes in **Cluster 6**.

Furthermore, the factual clustering revealed semantic mergers that reflect the underlying nature of the harm rather than just the defendant type. **Cluster 4** (right-center), for instance, groups *Undue Negative Credit Report (Banking)* (Class 8, pink points) together with *Undue Negative Credit Report (Telephony)* (Class 11, purple points). Although these involve different entities, they share the same factual trajectory of credit restriction. **Cluster 2** (top-center) presents a more heterogeneous composition dominated by *Unsolicited Consignment Credit Card* (Class 4, green points) but overlapping with *Moral Damages (Banking Service Failure)* (Class 1) and *PASEP Balance Revision* (Class 2). This overlap is legally consistent, as these dispute types frequently share similar narrative structures regarding banking errors. If we note theses classes in the full text scenarios (represented by the same colors green, orange, pink and brown), we can observe that they are very fragmented in different clusters, and their proximity did not correspond to the semantic similarity described before.

In both scenarios, the classes corresponding to *Unauthorized Bank Deductions* (Class 0, dark blue points) and *Moral Damages due to Service Failures* (Class 1, light blue) constitute the largest portion of the corpus. These classes exhibited fragmentation across multiple clusters, particularly in the Full-text configuration. This dispersion is attributable to the intrinsic semantic heterogeneity of these categories: while Class 0 encompasses a wide range of financial sub-products (e.g., undisclosed insurance fees, capitalization bonds, and administrative tariffs), Class 1 acts as a derivative claim arising from diverse factual antecedents rather than a single narrative pattern.

However, a clear distinction arises in the *Factual Narratives* case. Here, Class 0 is primarily present into **Clusters 5 and 7**, while Class 1 is splited in **Clusters 6 and 8**. Notably, these clusters exhibit spatial proximity in the UMAP space, reflecting the frequent co-occurrence of these claims in civil litigation. In contrast, under the Full-text scenario, these categories remain widely dispersed, suggesting that without narrative extraction, the model fails to unify the variations of these high-volume classes into coherent topological regions.

However, in factual clustering class 0 is more concentrated in **Clusters 5 and 7**, and class 1 is more concentrated in **Clusters 6 and 8**, where these cluster are closer to each other. Instead, in the

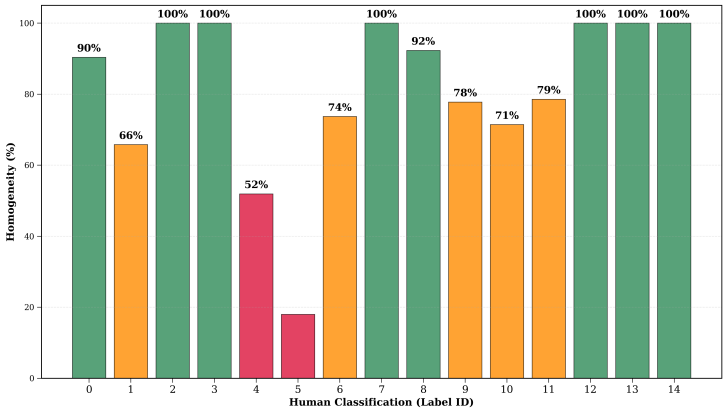
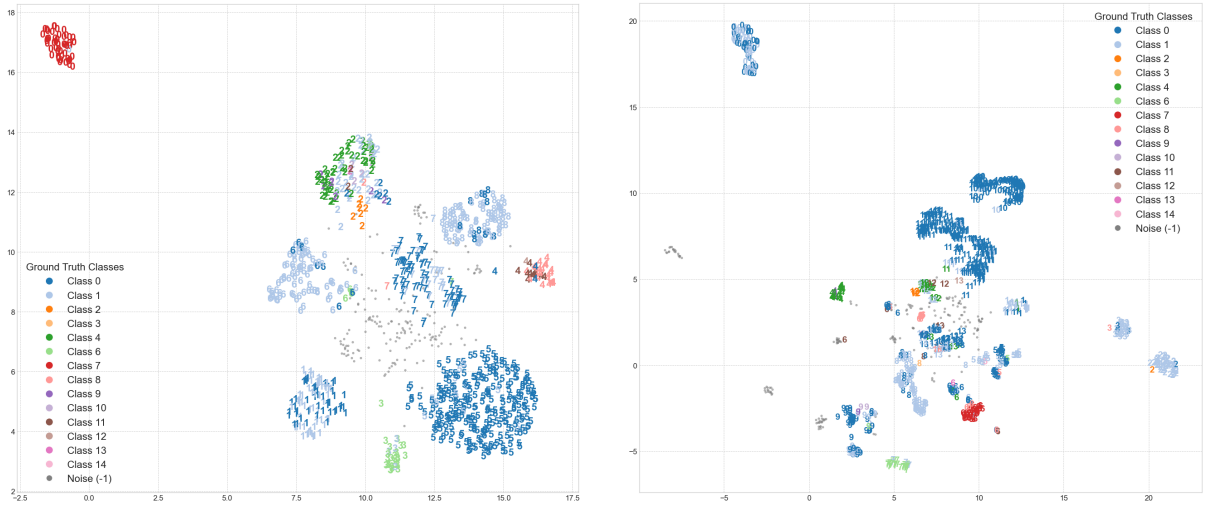


Figure 2: Cluster homogeneity based on the alignment between HDBSCAN labels and human classification.



(a) Facts

(b) Full text

Figure 3: Two-dimensional visualization of the clusters generated from Legal-BERT representations.

full text case, these two classes are very fragmented in several clusters across all the UMAP projections.

Finally, "Noise" (Label -1) detected by the HDBSCAN algorithm are shown as gray points in both figures. Basically, they are less frequent classes - eg. *Defective Products* (Class 12) or *Bailment* (Class 3) - that lack of sufficient density to form independent groups. Visually, the Full-text configuration presents a more diffuse distribution of these outliers, confirming that the use of verbose text in petitions introduces variance that hinders their similarity analysis.

5 Conclusions and future work

This work presented a pipeline for pattern discovery in legal documents, combining factual narrative extraction via Large Language Models and text embeddings to further apply dimensionality reduction and density-based clustering techniques. The methodology was applied to initial petitions from the Court of Justice of Amazonas, comparing the effectiveness of the clustering performed on factual segments versus full texts in terms of quantitative metrics (e.g. V-Measure) and qualitative validation based on specialized legal categories.

Experiments showed that the use of domain-specific embeddings (Legal-BERT) applied exclusively to factual segments outperformed full-text approaches, resulting in significantly more cohesive clusters and well-defined thematic groups. The removal of procedural noise and rhetorical argumentation allowed the method to identify both repetitive demands and atypical cases with greater precision, reducing lexical dispersion and increasing the interpretability of the discovered legal patterns.

For domain professionals, this methodology offers highly practical applications across all stages

of the judicial workflow. Particularly in high-volume litigation like banking disputes, which heavily burden Brazilian courts, clustering factual narratives allows judges and court analysts to identify groups of similar claims. This capability facilitates the batch-processing of repetitive demands, optimizing backlog management and ensuring greater procedural celerity.

Future work will expand this pipeline to diverse judicial contexts and explore automatic cluster labeling via topic modeling. Ultimately, we envision integrating this system-agnostic architecture into electronic lawsuit platforms—such as Projudi, used by the analyzed court, or others—to provide real-time, data-driven decision support, from initial triage to the drafting of standardized rulings.

6 Acknowledgements

This work was also supported by the National Institute of Science and Technology in Responsible Artificial Intelligence for Computational Linguistics, Information Treatment, and Dissemination (INCT-TILDIAR), funded by the Brazilian National Council for Scientific and Technological Development (CNPq), grant no. 408490/2024-1.

References

Deepak Bastola and Woohyeok Choi. 2025. Hybrid topic-semantic labeling and graph embeddings for unsupervised legal document clustering. *Preprint*, arXiv:2509.00990.

Javier Blanco-Portals, Francesca Peiró, and Sònia Estradé. 2022. Strategies for eels data analysis. introducing umap and hdbscan for dimensionality reduction and clustering. *Microscopy and Microanalysis*, 28(1):109–122.

Brasil. 2018. Lei nº 13.709, de 14 de agosto de 2018. `https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm`. Lei Geral de Proteção de Dados Pessoais (LGPD). Diário Oficial da União, Brasília, DF, 15 ago. 2018.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Conselho Nacional de Justiça (CNJ). 2025. Resolução nº 615, de 11 de março de 2025. `https://atos.cnj.jus.br/atos/detalhar/6001`. Dispõe sobre o uso de sistemas de Inteligência Artificial no âmbito do Poder Judiciário e dá outras providências. Diário da Justiça Eletrônico, Brasília, DF, 12 mar. 2025.

Rhedson Esashika, Carlos Figueiredo, and Tiago Melo. 2025. Structuring information from initial petitions using llms: A study in brazilian courts of justice.

Lena Held and Ivan Habernal. 2025. Contemporary LLMs struggle with extracting formal legal arguments. In *Proceedings of the Natural Legal Language Processing Workshop 2025*, pages 292–303, Suzhou, China. Association for Computational Linguistics.

Matteo Marulli, Glauco Panattoni, and Marco Bertini. 2025. A document processing pipeline for the construction of a dataset for topic modeling based on the judgments of the italian supreme court. *Preprint*, arXiv:2505.08439.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

Davoud Moulavi, Pablo Andretta Jaskowiak, Ricardo Campello, Arthur Zimek, and Joerg Sander. 2014. Density-based clustering validation.

Harry Nan, Maarten Marx, and Johan Wolswinkel. 2024. Combining rule-based and machine learning methods for efficient information extraction from enforcement decisions. In *Legal Knowledge and Information Systems*, Frontiers in Artificial Intelligence and Applications, pages 321–326. IOS Press.

Raphael Oliveira and Erick Giovani Sperandio Nascimento. 2021. Clustering by similarity of brazilian legal documents using natural language processing approaches.