applied sciences

MDPI

Article

Ranked Multi-Label-Augmented Topic Modeling for Legislative Content Profiling

Francesco Invernici *, Andrea Colombo, Flaminia Telese and Anna Bernasconi

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy; andrea1.colombo@polimi.it (A.C.); flaminia.telese@mail.polimi.it (F.T.); anna.bernasconi@polimi.it (A.B.)
* Correspondence: francesco.invernici@polimi.it

Abstract

Navigating extensive legislative corpora is often impeded by the linguistic complexity inherent in legal texts. To address this, we present a novel topic representation learning method designed to facilitate the systematic exploration of legislative content. We demonstrate the efficacy of this approach by applying it to the vast corpus of Italian legislation comprising about 74 k laws with more than 300 k articles. While current topic models group documents by latent semantic similarity, they often lack the granularity required for precise navigation. Our approach augments these representations by integrating our topic modeling framework with multi-label profiles. We enrich the representation of individual laws by extracting and ranking the top 10 keywords based on their relevance to the enclosing topic, subsequently aggregating these rankings to construct a comprehensive, alternative description of the broader legal themes. By bridging latent semantic clusters with explicit, LLM-generated labels, this method yields a highly interpretable representation of the corpus, significantly enhancing the profiling and navigability of complex legislative content. We improve over our baseline representation in 74.67% of cases, showing potential for re-use in highly specialized text corpora.

Keywords: topic modeling; unsupervised learning; augmented representation; multi-label assignment legislative corpus; Italian legislation

1. Introduction

The systematic exploration and analysis of ever-growing legislative corpora present a formidable challenge for legal scholars, practitioners, and the general public. These collections of documents, such as the vast corpus of Italian legislation, are characterized by significant linguistic complexity, the use of technical jargon, and a continuous temporal evolution that reshapes legal language and concepts over time [1]. While digital archives, common ontologies [2], and knowledge graphs [3,4] have improved access to legal texts, navigating these extensive repositories to identify thematic trends, retrieve relevant documents, and understand the structure of legal discourse remains a significant challenge.

Current computational approaches, particularly topic modeling, have shown promise in organizing large textual corpora by identifying latent semantic structures [5–7]. These models group documents based on shared vocabulary and underlying themes, offering a high-level overview of the content. However, they often produce representations that lack the necessary granularity and interpretability for precise analysis and navigation of the corpus. The traditional output of topic models—a list of keywords—may not fully capture

Academic Editor: Douglas O'Shaughnessy

Received: 27 March 2026
Revised: 22 April 2026
Accepted: 26 April 2026
Published: 30 April 2026

the nuanced concepts within legal documents, thereby limiting their practical utility in the legislative domain.

To address these limitations, we introduce a novel topic representation learning method that enriches each topic's representation by extracting fine-grained labels from documents and aggregating them by relevance. Our approach integrates the state-of-the-art topic modeling framework "Topics Evolution That You See" (TETYS) [8] with multi-label profiles generated by Large Language Models (LLMs). We first apply this framework to the extensive corpus of the Italian Republic legislation, spanning from 1948 to the present day, to generate foundational thematic clusters. Subsequently, we enrich the representation of each law by extracting descriptive labels and ranking them according to their relevance to the law's enclosing topic. These individual rankings are then aggregated to construct a comprehensive and highly interpretable profile for each of the initially identified topics.

By bridging—on the one hand—the gap between the latent clusters of documents and—on the other hand—explicit, semantically rich labels, our method produces a navigable and insightful representation of the legislative corpus. The main contributions of this paper are as follows.

- We adapt a domain-agnostic topic modeling pipeline to the complex and linguistically characterized domain of Italian legislation.
- We develop a method for enriching individual law representations by ranking LLM-generated labels based on their relevance to the document's assigned topic, enhancing metadata granularity.
- We aggregate the individual rankings associated with law labels to construct topic profiles, providing an alternative and highly interpretable description of legal themes.
- We run a comprehensive qualitative and quantitative evaluation of our approach, demonstrating its efficacy in improving the profiling and navigability of complex legislative content.

Section 2 reviews relevant literature and prior approaches. Section 3 details the materials and the methodological pipeline, including data extraction, topic modeling, and the label augmentation process. Section 4 presents the results obtained by applying our method to the Italian legislative corpus. Section 5 provides a thorough evaluation of the generated topic representations. Finally, Section 6 summarizes our findings and suggests avenues for future research.

2. Background and Related Work

Topic modeling includes a suite of methods for discovering latent thematic structures in large text corpora. These methods primarily consist of unsupervised learning techniques. Early and influential approaches, such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) [9], are probabilistic models that represent documents as mixtures of topics, where each topic is a distribution over words [10]. While foundational, these are essentially "bag-of-words" models that do not capture the contextual nuances of language. The evolution of the field has led to the development of neural topic models that leverage text embeddings and clustering [11]. A state-of-the-art example is BERTopic [12], which utilizes transformer-based embeddings to cluster documents based on their semantic meaning rather than mere word co-occurrence. TETYS [13], originally developed to work on COVID-19-related scientific abstracts [14] and applied to investigate Sustainable Development Goals [8], evolved as an extension of BERTopic, specifically incorporating LLMs for embedding generation and utilizing self-tuning characteristics that allow managing massive corpora from complex domains with varying characteristics. In this work, we extend TETYS, targeting the challenges of a new application domain (Section 2.1) and exploiting the potential of LLMs (Section 2.2).

2.1. Topic Modeling in the Legislative Domain

The legal field has become a significant application area for topic modeling, as researchers seek to analyze judicial proceedings, legislative trends, and legal scholarship. Basic approaches such as LDA and NMF have been employed to explore political and parliamentary debate [15–17] or legal documents [18,19] in different languages.

BERTopic has also been adapted to this domain. For instance, Silveira et al. [20] customized the BERTopic pipeline to analyze U.S. Supreme Court decisions, using a domain-specific LEGAL-BERT model to generate embeddings and treating individual paragraphs as units of analysis to allow for multi-topic assignments within a single document. Aguiar et al. [21] proposed building thematic models for classifying Brazilian lawsuits using BERTopic. Finally, Didwania et al. [22] conducted a cross-country study applying BERTopic, alongside LDA and NMF, to legislative documents from India and the United Kingdom to unveil thematic content.

Altogether, these studies confirm the utility of modern topic modeling techniques for extracting meaningful patterns from legal texts. However, most of them still employ outdated methods and all of them conclude with the generation of topics represented by lists of keywords, without further steps to refine their interpretability or integrate them with other forms of structured metadata. An organic extension of topic descriptions is a novel objective of this work.

2.2. Augmenting Topic Models with LLMs

While the aforementioned works demonstrate the power of topic discovery, our research addresses a critical next step: enhancing the usability and interpretability of the generated topics.

Traditional approaches like LDA provide interpretable topic–word distributions but often fail to capture semantic relationships between words. Neural approaches, such as the Spherical Embedded Topic (SET) model [23] and NMTF-LTM [24], address this limitation but often sacrifice interpretability. For complex structures, GNN-based neural topic models [25,26] have been developed to maintain interpretability while leveraging network connectivity and graph-based sampling strategies.

The emergence of LLMs has clearly opened up new possibilities, offering both superior semantic understanding and high interpretability, although often requiring significant computational resources. Recent work demonstrates that LLMs can serve as an alternative to traditional topic modeling approaches, offering improved semantic understanding and topic coherence [27]. Frameworks like TopicGPT [28] produce highly interpretable topics with natural language labels, and DeTiME [29] combines LLMs with diffusion models to generate coherent topics and topic-based text. Recent advancements have also shown the potential of LLMs to generate high-quality metadata, such as descriptive labels or summaries for legal documents [30].

Considering the computational challenge, recent work favors efficient hybrid models, coupling embeddings from lightweight pretrained language models with targeted LLM usage for key tasks, like topic representation and evaluation [31]. In our work, we employ parameter-efficient LLMs (up to 7B parameters) for embedding and enhancing the representation of topics.

2.3. Grounding in the Current Literature

The legislative domain presents very specific aspects, including highly technical jargon, long, convoluted grammar structures, and hierarchical, unbalanced distribution of matters. These call for streamlined approaches for lightweight exploratory tasks.

On the one hand, these aspects have not been considered by works presented in Section 2.1, not targeting rich topic representations that easily support further exploration. On the other hand, this need is quite specific, and it is not surprising that existing frameworks (although quite advanced in topic model augmentation proposals—see Section 2.2) have not been crafted toward a legislative text-specific case.

Our proposal is to use LLM-generated labels, appropriately ranked, to extend topic and law representations. Methodologies to represent labels in topics were surveyed some years ago [32], with considerable advancements since LLMs have been introduced in the process [31,33]. We have been inspired by milestone methodologies of advanced topic modeling, incorporating elements from entropy-based literature [34] and vector-space-based approaches [35]. Our approach combines these ideas to build a framework that satisfies the peculiar needs of the legislative domain.

All in all, our main contribution stands in the composition of different techniques to reach a semantically richer and informative topic representation. We do not contrast our results to specific frameworks as comparable approaches do not exist in the literature.

3. Materials and Methods

Our methodology extends TETYS [8], a fully automated topic modeling pipeline. Until now, its efficacy has only been shown in text corpora of scientific publications [36]. These have quite different characteristics and terminology w.r.t. the Italian legislative language we aim to analyze. First, we adapted the pipeline for the new specific domain. Then, we extended it with a representation–augmentation step, aimed at decorating topics with a roster of descriptive, domain-specific, and highly interpretable labels—note that labels also serve to enrich the metadata of single laws. To achieve this, our methodology follows the multi-stage pipeline shown in Figure 1.

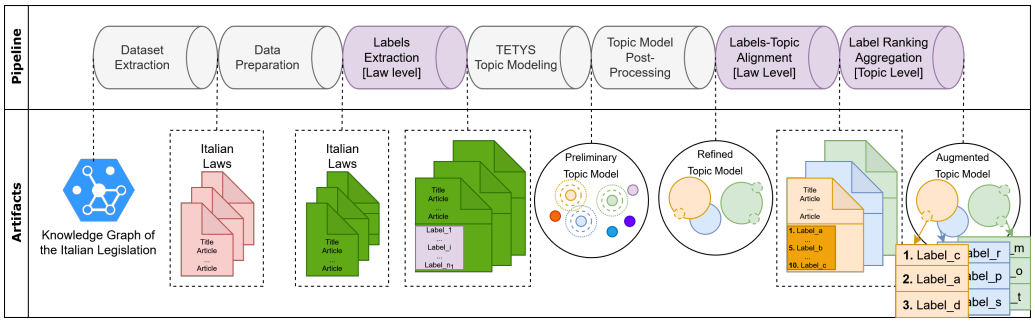

Figure 1. Ranked multi-label augmented topic modeling pipeline for legislative text. Modules in purple are specific to the multi-label topic augmentation. It begins with the extraction of the legislative corpus, followed by the generation of descriptive labels for each document using a fine-tuned LLM. We then employ an adapted topic modeling framework to identify latent thematic clusters. Finally, our augmentation technique, which leverages these clusters to rank the LLM-generated labels, creates enriched profiles for both individual laws and the broader topics they constitute.

3.1. Legislative Corpus and Data Preparation

The dataset employed contains the complete corpus of the Italian legislation from 1948 to 2025, extracted from the Knowledge Graph of Italian Legislation [30]. The corpus includes 74,184 documents, each representing a single law with its associated articles and corresponding texts. As characterizing features, the texts have a considerable average total length (850 words), average single-sentence length (>20 words), and linguistic complexity. The dataset includes Latin expressions, highly technical terminology (e.g., '*contumacia*'), and context-specific formulations whose meaning diverges from common usage (e.g., '*buon padre di famiglia*', a legal and old-fashioned term to refer to the main person who represents their family, not used in current language) [1].

Data extraction was performed by querying the graph database to retrieve the full text for each law. The raw text underwent a preprocessing phase to prepare it for topic modeling. This included standard text cleaning, such as removing symbols and special characters. Moreover, we implemented a two-tiered stopword removal process: First, a standard list of Italian stopwords was applied, followed by a custom-curated list of domain-specific stopwords (e.g., '*legge*' (law), '*articolo*' (article), '*decreto*' (decree)) that are highly frequent but offer little semantic value for thematic differentiation. Finally, all texts were lemmatized using the `it_core_news_lg` model from spaCy [37] to normalize word forms.

3.2. Extraction of Labels from Laws

To create descriptive labels for each law, we implemented a label extraction methodology [30]. We used a fine-tuned `Mistral-7B` model to identify and extract a concise list of descriptive keywords (labels) from the titles and contents of each law. The model was fine-tuned on a dataset generated using a larger LLM (`Mixtral-8x22B`) with few-shot prompting, followed by heuristic filtering and lemmatization to ensure high-quality, normalized labels. This step provides a set of multi-label descriptors for every document in the corpus, serving as the foundation for our augmentation procedure.

3.3. Topic Modeling Framework

The topic modeling approach embedded in the pipeline stems from the TETYS framework [8], which is built upon the BERTopic library [12]. This framework identifies topics in the latent embedding space after dimensionality reduction and clustering.

Document Embedding. Our strategy to handle the specific linguistic characteristics of Italian legal text diverges from the standard SBERT models [38] of BERTopic. We employed the *gte-Qwen2-7B-instruct* model [39] to generate dense vector representations (embeddings) of 3584 dimensions for each preprocessed document. At the time of writing (March 2026), this model achieves a reasonable compromise between performance and efficiency for clustering, as documented in the Multilingual Leaderboard of the Massive Text Embedding Benchmark [40].

Clustering and Topic Identification. The high-dimensional embeddings were first reduced using Uniform Manifold Approximation and Projection (UMAP) [41] to preserve both local and global data structures. Subsequently, Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [42] was applied to the reduced embeddings to identify dense clusters of semantically similar documents. Each resulting cluster represents a distinct topic. In the first run, we obtained 515 topics.

Topic Representation and Post-Processing. Each topic is initially represented by a list of the most relevant words, calculated using a class-based TF-IDF (c-TF-IDF) variant (see [12]). This method treats all documents within a cluster as a single composite document and identifies terms that are characteristic of that topic. To enhance interpretability, we used the KeyBERTInspired algorithm to refine these representations, ensuring keyword diversity and relevance. A hyperparameter optimization search, guided by the Density-Based Clustering Validation (DBCV) score [43], was conducted to select the optimal UMAP and HDBSCAN parameters. Furthermore, to address the issue of an initially large number of semantically similar topics, we performed a topic reduction step, merging the most similar clusters based on the cosine similarity of their c-TF-IDF vectors. This reduced the number of topics from 515 to a more distinct and manageable set of 300, significantly increasing topic diversity. The ranges and optimal parameters applied in this process are reported in Table 1.

Quantitative Assessment. To quantitatively assess the topic model in terms of diversity and coherence, we adopted four additional metrics, as implemented in the OCTIS library [44]. Topic diversity measures how varied the top words across all topics are (closer to 1 is better) [45]. Inversed RBO evaluates topic distinctiveness within a single model by calculating the reciprocal of Rank-Biased Overlap (RBO) [45]. C_V Coherence uses a variant of Non-Pointwise Mutual Information (NPMI) with a weighted sliding window to measure how semantically related the top words in a topic are [46]. U_{Mass} coherence is an asymmetric measure that calculates the log ratio of the co-document frequency of word pairs against the document frequency of a single word, estimating the degree of semantic coherence based on co-occurrence in the corpus [46].

Table 1. Parameters, their corresponding ranges for the random search space, and the values found for the highest DBCV score.

Parameters	Ranges	Optimal Value
umap_n_neighbors	[2, 5, 7, 10, 15, 20, 30]	30
umap_min_dist	[0.0]	[0.0]
umap_n_components	[5, 10, 15, 20, 30]	20
hdbscan_min_samples	[5, 10, 15, 20, 30]	30
hdbscan_min_cluster_size	[5, 10, 15, 20, 25]	10
hdbscan_cluster_selection_method	[“eom”, “leaf”]	“eom”
hdbscan_metric	[“euclidean”]	“euclidean”

3.4. Multi-Label Augmentation of Topics

The core novelty of our method lies in augmenting the topic model’s output by integrating LLM-generated labels. This process operates at two levels: enriching metadata of individual laws and creating new representations for topics.

Alignment of Labels and Topics for Individual Laws. For each law, we leverage a set of unordered *labels* and an assigned topic, which is described by a list of weighted-by-relevance keywords (commonly referred to as *subtopics*). To create a relevance-based ranking of the labels, we first embedded both the labels and the subtopics using the same lightweight LLM used in the previous modules of the pipeline, i.e., *gte-Qwen2-7B-instruct*, (which also includes support for Italian). We then constructed a cosine similarity matrix between the labels and the top 10 subtopics of the law’s parent topic, as shown in Figure 2. The labels were then ranked iteratively by selecting the label with the highest similarity score, removing it from consideration, and repeating the process until all labels are ordered. This yields a fine-grained, ranked profile for the content of each law, as exemplified in Figure 3.

Aggregating Label Rankings for Topics. To create an alternative, label-based representation for each topic, we aggregated the ranked label lists of all laws within that topic. We treated each law’s ranked list as a ballot and employed a rank aggregation algorithm to find a consensus ranking. Specifically, we used the Sequential Winner algorithm with a Plurality Score [47], where in each round the label appearing most frequently at the top of the unranked lists is selected. This process generates a final, aggregated ranking of the top 10 most representative labels for the entire topic, providing a complementary and highly interpretable description of its core theme.

	Law L						
	Label 1	**Label 2**	**Label 3**	…	**Label** j	…	**Label m**
Subtopic 1	$k_{1,1}$	$k_{1,2}$	$k_{1,3}$	…	$k_{1,j}$	…	$k_{1,m}$
Subtopic 2	$k_{2,1}$	$k_{2,2}$	$k_{2,3}$	…	$k_{2,j}$	…	$k_{2,m}$
Subtopic 3	$k_{3,1}$	$k_{3,2}$	$k_{3,3}$	…	$k_{3,j}$	…	$k_{3,m}$
…	…	…	…	…	…	…	…
Subtopic i	$k_{i,1}$	$k_{i,2}$	$k_{i,3}$	…	$k_{i,j}$	…	$k_{i,m}$
…	…	…	…	…	…	…	…
Subtopic n	$k_{n,1}$	$k_{n,2}$	$k_{n,3}$	…	$k_{n,j}$	…	$k_{n,m}$

Figure 2. Similarity matrix produced for the alignment of labels and topics for a generic law L. Each $k_{i,j}$ is the value of the cosine similarity between the embedding vectors of the Subtopic i and the Label j.

	Law L
Rank	**Label**
1	Label $a = argmax_{j,1}(k_{i,j})$
2	Label $b = argmax_{j,2}(k_{i,j})$
3	Label $c = argmax_{j,3}(k_{i,j})$
…	…
m	Label $s = argmax_{j,m}(k_{i,j})$

Figure 3. Final ranking produced for the alignment of labels and topics for a generic law L. As in Figure 2, each $k_{i,j}$ is the value of the cosine similarity between the embedding vectors of the Subtopic i and the Label j.

4. Results

Here, we present the outcomes of the proposed multi-label-augmented topic modeling pipeline, when applied to the Italian legislative corpus. We first detail the characteristics of the preliminary topic model, then describe the improved model resulting from our topic reduction process, and finally showcase the enriched representations generated by our label ranking and aggregation method. The overall execution took about 200 min for the modules described in Section 4.1; ∼100 s for the post-processing in Section 4.2; and ∼10 min for the modules described in Section 4.3) on a node of a High-Performance Computing cluster equipped with four NVIDIA Ampere A100 64 GB GPUs, a single-socket 32-core Intel Xeon Platinum 8358 CPU, and 512 GB of RAM.

4.1. Preliminary Topic Model

The initial application of our adapted topic modeling framework yielded a model with 515 topics. The optimal parameters selected via random search resulted in a DBCV score of 0.527. This preliminary model achieved a C_V coherence of 0.77 and a topic diversity score of 0.425, indicating reasonably coherent but somewhat redundant topics. A significant portion of the documents (24,435) were assigned to "Topic-1", which usually—in BERTopic models—collects outliers (noisy documents that do not contain enough information to be positioned within other thematic clusters). In this case, it primarily contained documents of a highly bureaucratic or specific nature that precluded clear thematic classification.

In particular, it primarily comprises three types of documents: (i) highly bureaucratic texts, such as administrative provisions, procedural regulations, or notifications of expired decrees (e.g., '*Decaduto*'); (ii) extremely short and specific legislative acts whose

unique content—such as specific financial grants to individual entities—prevents association with broader thematic clusters; and (iii) documents potentially misclassified due to residual noise in the model. Due to this lack of thematic unity and its function as a "catch-all" for non-interpretable items, the topic was excluded from the subsequent detailed qualitative analyses.

Despite the high number of topics, many were thematically coherent. For example, Topic 6 effectively grouped 693 laws related to the military, with representative subtopics such as '*esercito*' (army), '*militare*' (military), and '*ufficiale*' (officer). Similarly, Topic 7 captured 614 laws concerning fiscal matters, described by subtopics like '*imposta*' (tax), '*reddito*' (income), and '*tributario*' (tributary). However, manual inspection and analysis of the hierarchical clustering tree revealed significant semantic overlap between many topics. For instance, multiple distinct topics were generated for regulations concerning the Catholic Church, all sharing highly similar subtopics like '*chiesa*' (church), '*parrocchiale*' (parochial), and '*giuridico*' (legal), indicating a need to reduce the number of topics.

4.2. Post-Processed Topic Model

To address redundancy in the preliminary model, a post-processing step for topic reduction was performed, merging semantically similar clusters to produce a final model with 300 topics. This reduction preserved the overall cluster structure (DBCV score remained 0.527) while substantially improving the model's quality according to our evaluation metrics. As shown in Table 2, topic diversity increased significantly from 0.425 to 0.567, and the U_{Mass} coherence score improved from -0.842 to -0.7. The C_V coherence saw a negligible decrease to 0.754, confirming that the reduction process enhanced topic distinctiveness without sacrificing internal coherence.

Table 2. Comparison of evaluation metrics before and after the topic reduction step, demonstrating a significant improvement in topic diversity.

Metric	Preliminary Model	Post-Processed Model
C_V Coherence	0.770	0.754
U_{Mass} Coherence	-1.842	-0.700
Topic Diversity	0.425	0.567
Inversed RBO	0.990	0.994

The final model produced highly specific and interpretable topics. The intertopic distance map of the final model confirmed greater separation between clusters compared to the preliminary version, indicating a more distinct and navigable thematic structure. A relevant portion of the map can be inspected in Figure 4, where only the two primary components of the UMAP projection are plotted; word clouds of the 20 largest topics support domain interpretation, with several well-separated semantic clusters. For instance, Topic 5 focused on military personnel and ranks, Topic 8 on medical faculties and specializations, Topic 10 on post-war land reform and colonization in central Italy (specifically the Maremma region), and Topic 12 on taxation and fiscal declarations. Even when specific subtopics (i.e., keywords) are in common in multiple word clouds (e.g., 'parrocchia' in Topics 1, 3, 7), other subtopics provide different context (i.e., referring to the recognition of parochial jurisdiction, the union of parishes, or the construction of new ceremonial buildings).

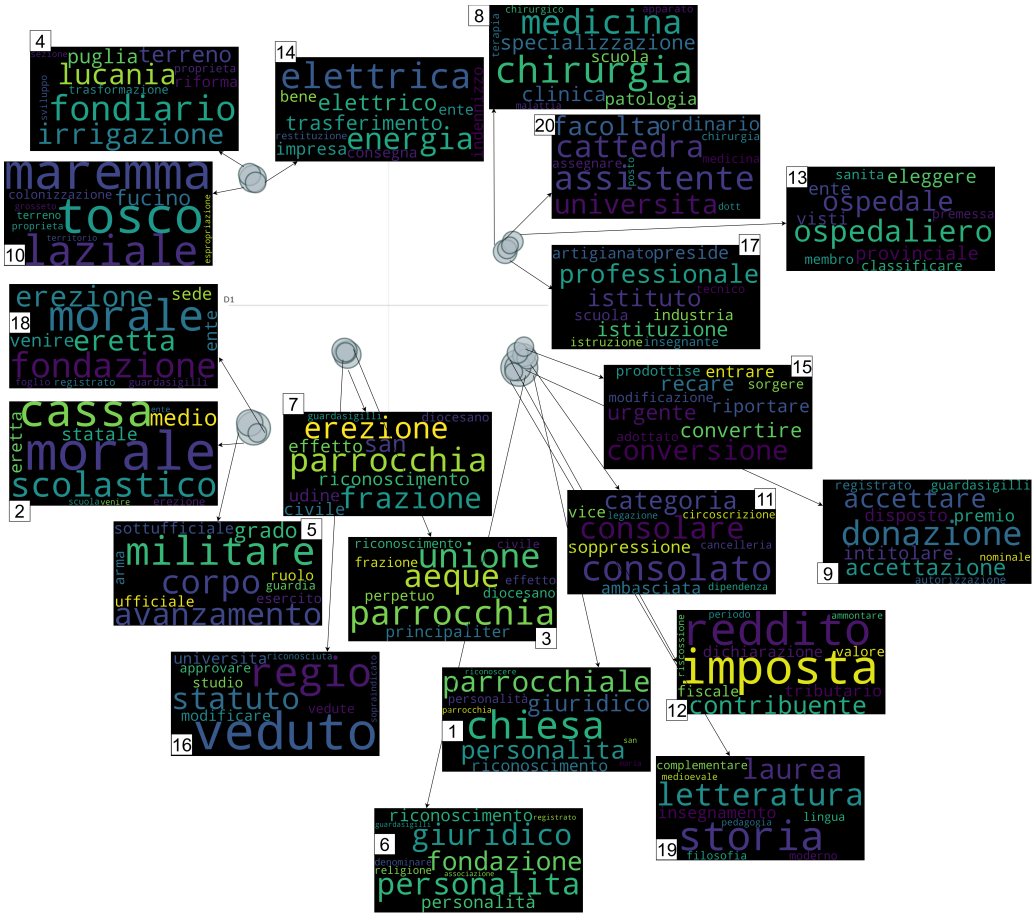

Figure 4. Portion of the intertopic distance map for the post-processed model, with word clouds for the 20 largest topics.

4.3. Augmented Law and Topic Representations

After the post-processing, we applied the last step of our method. For each law, we obtained a ranked list of labels, providing a more granular description w.r.t. an unordered set of keywords. For each topic, instead, we obtained a ranked multi-label representation of augmenting keywords that preexisted in the corpus of the topic.

Ranked Label Profiles for Individual Laws. For example, within Topic 5 (Military), a 1988 decree concerning the recall of military personnel for training (`DECRETO DEL PRESIDENTE DELLA REPUBBLICA 17 febbraio 1988, n. 39`, i.e., "Decree of the President of the Republic February 17, 1988, No. 39") received the ranked labels: (1) '*esercito*' (army); (2) '*militare*' (military); (3) '*addestramento*' (training); (4) '*aggiornamento*' (update); (5) '*difesa*' (defense). This ranking accurately reflects the document's primary focus, moving from the general domain to the specific actions legislated. Similarly, a 1986 law modifying the statute of the University of Naples medical school (Topic 8) was assigned the labels: (1) '*medicina*' (medicine); (2) '*università*' (university); (3) '*studente*' (student), prioritizing the core subject over the institutional context.

Aggregated Label Profiles for Topics. By aggregating the individual law rankings, we constructed the multi-label profile for each of the 300 topics. These aggregated profiles often provided a more intuitive and holistic description of a topic's content than the default c-TF-IDF subtopics, since they are not constrained to words already contained in the original laws, but can rely on more meaningful and conceptual keywords extracted by an LLM. Figure 5 shows four example topics with, side-by-side, (i) the original word cloud representation and (ii) the new aggregated ranking.

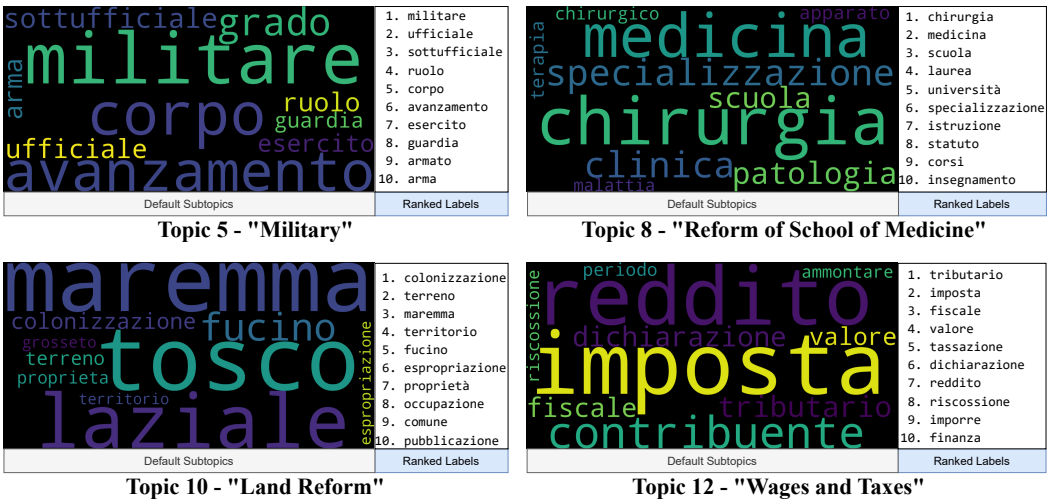

Figure 5. Baseline word clouds of representations alongside their new ranked profiles.

For Topic 5 (Military), the aggregated ranking, that is (1) '*guardia*' (guard), (2) '*militare*' (military), (3) '*sottufficiale*' (non-commissioned officer), emphasizes the focus on personnel and ranks, achieving a more specific representation of personnel management than the subtopic list. For Topic 8 (Reform of School of Medicine), the aggregated ranking is topped by '*medicina*' (medicine)—also prominent in the word cloud—but it also integrates '*laurea*' (degree) and '*specializzazione*' (specialization). These terms, alongside '*università*' (university), now fifth, capture the specific academic and certification requirements of medical reform that are sometimes less distinct in basic word cloud representations. For Topic 10 (Land Reform), the aggregated ranking is topped by (1) '*colonizzazione*' (colonization), (2) '*terreno*' (land/terrain), (3) '*Maremma*' (Maremma region), and (4) '*espropriazione*' (expropriation), directly highlighting the key legal action central to the topic, a nuance not as prominent in the subtopic representation. For Topic 12 (Wages and Taxes), there is strong agreement on '*tributario*' (tax-related), '*imposta*' (tax), and '*fiscale*' (tax-related/fiscal), which, together, constitute the most specific and representative terminology for this legal domain in Italian. These results suggest that the aggregated label rankings serve as a powerful, human-interpretable summary of the core legislative themes within each topic.

5. Evaluation

To validate the efficacy and added value of our multi-label augmentation approach, we designed a comprehensive evaluation framework. On one side, we compare the value of the metrics presented in Section 3 (e.g., coherence, diversity) across different studies, to assess the structural quality of the topic model. On the other side, we run an evaluation to measure the semantic and qualitative improvement of our final topic representations; we adopt a quantitative methodology using an LLM as a proxy for human judgment, complemented by statistical analysis to ensure the robustness of our findings.

5.1. Topic Coherence Reference Values for the Legal Domain

We compared the structural quality of our model against the findings of [22], who applied LDA, NMF, and BERTopic to legislative documents from India and the UK. Their evaluation, utilizing C_V and U_{Mass} coherence, was performed on three specialized datasets on Indian judgments, Indian legal summaries, and British Supreme Court cases.

The choice to adopt cross-national studies as a benchmark is motivated by the shared structural and linguistic complexity inherent to large-scale legislative corpora, regardless of jurisdiction. Legal texts are universally characterized by specialized terminology and dense, hierarchical organizations that present identical challenges for automated thematic

extraction. To the best of our knowledge, no previous state-of-the-art approach has been evaluated on the entire Italian legislative corpus for this specific task. Consequently, our work represents the first attempt to provide a baseline for the Italian domain. By benchmarking against established results from diverse legal systems, such as those in [22], we aim to demonstrate the robustness and generalizability of our framework in handling the universal intricacies of legal data.

We observe in Table 3 that our model obtained similar C_V coherence to the best results. Furthermore, our model achieved better U_{Mass} coherence, which is a noteworthy finding given that our analysis utilizes a significantly larger corpus, a factor that typically tends to skew U_{Mass} scores downward.

Table 3. Topic coherence scores for different models in India and the UK [22] versus our post-processed model for Italian laws.

Country	Model	C_v Coherence	U_{mass} Coherence
India	LDA	0.596	−1.03
	NMF	0.763	−1.915
	BERTopic	0.781	−1.846
UK	LDA	0.526	−0.91
	NMF	0.732	−0.915
	BERTopic	0.769	−1.554
Italy **(ours)**	BERTopic	0.770	−0.842
	TETYS	0.754	−0.70

5.2. Qualitative Evaluation Framework: LLM-as-a-Judge

We adopted the "LLM-as-a-Judge" paradigm [48] to conduct a scalable, human-like assessment of our model's output. For this purpose, we utilized Google's `Gemini Pro 2.5` model [49]. The evaluation was performed on all 299 meaningful topics (i.e., excluding the residual Topic 1).

To ensure the robustness of these automated judgments, we implemented specific mitigation strategies to address known cognitive biases inherent in LLM evaluators, as categorized by Yu et al. [50]. Specifically, we identified three primary concerns, position bias, verbosity bias, and self-enhancement bias, which we addressed in the following ways:

- **Position bias**: To prevent the LLM from favoring responses based on their order of appearance, we employed a randomized swapping mechanism. The two inputs (baseline and our method) were positioned randomly to avoid placing the baseline always before the proposed method's output.
- **Verbosity bias:** Recognizing that LLMs often equate length with quality, we evaluated lists of the same length and we checked that the number of tokens was uniform across all topics' prompts.
- **Self-enhancement bias:** Given that `Gemini Pro` was used as the evaluator, there was a risk of it favoring outputs that mirror its own stylistic patterns. To mitigate this, we utilized LLMs from other providers trained following different guidelines, like Alibaba's `Qwen` for topic modeling and our previously fine-tuned `Mistral` model for label extraction. This was also useful to reduce the risk of preference leakages [51] across all our experimental setups.

For each topic, the LLM was presented with the preprocessed text of five randomly selected laws belonging to that topic. It was then given two distinct ordered lists of descriptive terms: **List 1 (Subtopics)**, representing the baseline representation, consisting of

the top 10 keywords generated by the standard c-TF-IDF method; **List 2 (Ranked Labels)**, as our proposed representation, consisting of the top 10 labels from the aggregated ranking. The LLM was tasked with answering four questions to compare the two lists:

- $Q_{Subtopics}$: Does the first list (subtopics) adequately and consistently describe the provided laws? (YES/NO)
- Q_{Labels}: Does the second list (ranked labels) adequately and consistently describe the provided laws? (YES/NO)
- $Q_{Comparison}$: Does the second list describe the laws better than the first, considering overall relevance to the topic? (YES/NO)
- Q_{Score}: Assign a score from 1 (no contribution) to 7 (maximum contribution) reflecting how much the second list improves the description of the topic compared to the first.

The first three questions aim to assess the adequacy of the baseline and our method, and accept binary answers. The last question asks the LLM to perform a direct comparison between the quality of the two representations using a 7-point Likert scale. This scale runs from 1 (indicating there is no improvement in the topic representation) to 7 (indicating the multi-label profile strongly improves the topic's representation), with 4 representing no perceptible difference in quality.

All results were manually re-checked by a collaborator of our group, an expert in legislative data management, validating the LLM's judgment.

The collected feedback allowed for a direct comparison of the interpretability and semantic relevance of our augmented representations against the baseline topic modeling output, as presented in the two following sections.

5.3. Analysis of LLM Judgments

The LLM-as-a-Judge experiment yielded clear and statistically significant results demonstrating the tangible advantage of the multi-label-augmented representations.

Overall, the LLM judged both representations to be of high quality, affirming the validity of the underlying topic model. The subtopic list was deemed adequate in 70.00% of cases, while the ranked label list was considered adequate in 90.67% of cases. Indeed, when asked for a direct comparison, the LLM showed a clear preference for our method. As shown in Table 4, the ranked labels were judged to be a better description of the topic in 74.67% of cases.

Table 4. Results of the LLM-as-a-Judge evaluation, on the top in percentages and on the bottom in absolute values.

Response	$Q_{Subtopics}$ (Adequate)	Q_{Labels} (Adequate)	$Q_{Comparison}$ (Labels Better)
NO	30.00%	9.33%	25.33%
YES	70.00%	90.67%	74.67%
NO	90	28	76
YES	210	272	224

The scoring question (Q_{Score}) provided further evidence of the significant contribution of our method. The mean score across all topics was 4.62 out of 7, with a median of 5, indicating a substantial improvement over the baseline. A breakdown of the scores reveals a compelling trend: in cases where the ranked labels were preferred, the median improvement score was 5 ("significant contribution"). In the minority of cases where they were not preferred, the median score was 3 ("minor contribution"), indicating that even when not superior, our method produces representations of comparable quality to the baseline.

5.4. Statistical Validation

To ensure the statistical validity of these findings, we conducted further tests on the LLM's responses. A McNemar test was performed to compare the proportions of "YES" answers for $Q_{Subtopics}$ and Q_{Labels}. The resulting p-value of 1×10^{-6} ($p < 0.01$)) indicates that there is a statistically significant difference in the proportions of adequacy of the two lists. This finding suggests that our method has a significant advantage in quality with respect to the baseline.

Furthermore, we fitted a logistic regression model to determine if the responses to $Q_{Subtopics}$ and Q_{Labels} could predict the outcome of $Q_{Comparison}$. The model yielded a highly significant result (Odds Ratio of $Q_{Labels} = 35.765 \gg 1$), confirming a strong predictive relationship. This demonstrates that the perceived quality of the two lists directly influences the preference judgment.

Finally, an analysis of cases where the baseline subtopics were deemed inadequate (79 topics) showed that our ranked labels provided a satisfactory and preferred description in 77 of those instances (97.47%). This highlights the robustness of our method, as it offers a substantial improvement precisely when the traditional topic representation is weakest. In summary, the comprehensive evaluation confirms that our multi-label-augmented approach yields a more interpretable, relevant, and preferred representation of legislative content.

6. Conclusions

Navigating the vast and complex landscape of legislative corpora requires sophisticated tools that can distill intricate legal texts into coherent, interpretable, and navigable thematic structures.

In this paper, we introduced a novel method, *ranked multi-label-augmented topic modeling*, designed to address the limitations of traditional topic models in the legislative domain on large-scale datasets. By combining latent topic representations with explicit, LLM-generated multi-label profiles, our approach produces a significantly enriched and more intuitive framework for legislative content profiling.

First, we successfully adapted and applied a state-of-the-art topic modeling pipeline to the challenging corpus of Italian legislation, demonstrating its effectiveness in a linguistically distinct and complex domain. Then, we developed a novel methodology to augment these latent representations by creating ranked, relevance-based profiles for individual laws. This was achieved by leveraging the semantic context of a law's parent topic to organize its associated labels, transforming a flat set of keywords into a structured, granular description. Finally, by aggregating these individual rankings, we constructed comprehensive and nuanced topic profiles that serve as a powerful alternative to standard keyword-based topic descriptors, enriching topics' representation for other prospective use.

The evaluation of our method, conducted through a state-of-the-art LLM-as-a-Judge framework, confirmed its efficacy. Our augmented representations were found to be preferable to traditional subtopic lists in almost 75% of cases, providing a substantial improvement in semantic clarity and relevance. This approach proved particularly valuable in instances where baseline topic representations were weak, highlighting its robustness and practical utility. This evaluation was designed to provide first-level feedback on whether this kind of topic augmentation would provide useful results. While the outcome was indeed promising, we plan to run extensive evaluations in the future.

A limitation of this work lies in the handling of a large, heterogeneous cluster of documents (Topic -1) that could not be thematically resolved. In future developments, we plan on mitigating this by implementing iterative re-ranking strategies or semi-supervised "outlier-reduction" passes to redistribute these documents into more specialized sub-clusters. Furthermore, we intend to explore hierarchical classification strategies to better parse highly bureaucratic or domain-specific texts that currently resist standard semantic segmentation.

Regarding the broader outcome of this research, we advocate for a responsible and sustainable implementation of AI-driven legislative tools. In accordance with the European Union Artificial Intelligence Act, which classifies tools supporting the search and interpretation of legislative documents as high-risk, we ensure high standards of transparency and auditability. Our modular, step-by-step pipeline allows for the monitoring and verification of results at each stage—from latent topic modeling to label augmentation—ensuring the system does not operate as a "black box." To this end, our methodology mitigates the risk of LLM hallucinations by using a frequentist approach (TF-IDF) as a baseline; by grounding the AI-generated labels in the actual statistical distribution of the corpus, we ensure that the augmented representations remain faithful to the source text. Since our methodology performs AI-driven knowledge management on a corpus of enacted laws rather than political discourse or parliamentary debates, the risk of ideological bias is significantly minimized; the task remains one of structural organization rather than opinion synthesis. Finally, to ensure environmental sustainability, further research will focus on optimizing the inference efficiency of the underlying LLMs, reducing the computational footprint required for large-scale legal processing without compromising semantic accuracy.

In conclusion, our research demonstrates that bridging latent semantic structures with explicit, structured metadata yields a highly effective representation of complex textual corpora. The proposed method not only enhances the interpretability of topic models but also lays the groundwork for future information retrieval systems and practically relevant analytical dashboards for legislative data. As an example, we foresee its inclusion in platforms such as LegisSearch [52], enhancing the search and retrieval in complex linked datasets. By enabling more precise and intuitive exploration of legal content, this work offers a valuable artifact for legal scholars, policymakers, and citizens seeking to understand the intricate fabric of law.

Author Contributions: Conceptualization, F.I., A.C. and A.B.; methodology, F.I., A.C. and A.B.; software, F.I. and F.T.; validation, F.I. and F.T.; formal analysis, F.I.; investigation, F.I.; resources, F.I.; data curation, F.I. and A.C.; writing—original draft preparation, F.I.; writing—review and editing, A.C. and A.B.; supervision, A.B.; project administration, A.B.; funding acquisition, A.B. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by TEThYS, a beneficiary of the EU-funded NGI Search project (Sub-grant Agreement SEARCH OC2_18).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.19243620.

Acknowledgments: During the preparation of this study, the authors used Google's `Gemini Pro 2.5` model for the purposes of topic evaluation. The authors have reviewed the output and take full responsibility for its content.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

DBCV	Density-Based Clustering Validation
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
LDA	Latent Dirichlet Allocation
LLM	Large Language Model
NMF	Non-Negative Matrix Factorization
NPMI	Non-Pointwise Mutual Information
RBO	Rank-Biased Overlap
TETYS	Topic Evolution That You See
TF-IDF	Term Frequency–Inverse Document Frequency
UMAP	Uniform Manifold Approximation and Projection

References

1. Siino, M. Exploring the use of LLMs in the Italian legal domain: A survey on recent applications. *Comput. Law Secur. Rev.* **2025**, *58*, 106164 . [CrossRef]
2. Palmirani, M.; Vitali, F. Akoma-Ntoso for legal documents. In *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 75–100.
3. Colombo, A. Leveraging knowledge graphs and LLMs to support and monitor legislative systems. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, Boise, ID, USA, 21–25 October 2024; pp. 5443–5446.
4. Colombo, A.; Cambria, F.; Invernici, F. Legislative knowledge management with property graphs. In Proceedings of the Workshops of the EDBT/ICDT 2025 Joint Conference Co-Located with the EDBT/ICDT 2025 Joint Conference, Barcelona, Spain, 25 March 2025; Volume 3946, pp. 1–8.
5. Abdelrazek, A.; Eid, Y.; Gawish, E.; Medhat, W.; Hassan, A. Topic modeling algorithms and applications: A survey. *Inf. Syst.* **2023**, *112*, 102131. [CrossRef]
6. Schirmag, T.; Wedemeyer, J.H.; Stechemesser, A.; Wenz, L. Neural topic modeling reveals German television's climate change coverage. *Commun. Earth Environ.* **2025**, *6*, 441. [CrossRef]
7. Falkenberg, M.; Galeazzi, A.; Torricelli, M.; Di Marco, N.; Larosa, F.; Sas, M.; Mekacher, A.; Pearce, W.; Zollo, F.; Quattrociocchi, W.; et al. Growing polarization around climate change on social media. *Nat. Clim. Change* **2022**, *12*, 1114–1121. [CrossRef]
8. Invernici, F.; Curati, F.; Jakimov, J.; Samavi, A.; Bernasconi, A. Capturing research literature attitude towards sustainable development goals: An LLM-based topic modeling approach. *J. Big Data* **2025**, *12*, 139. [CrossRef]
9. Obadimu, A.; Mead, E.; Agarwal, N. Identifying latent toxic features on YouTube using non-negative matrix factorization. In Proceedings of the Ninth International Conference on Social Media Technologies, Communication, and Informatics, Valencia, Spain, 24–28 November 2019.
10. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *J. Mach. Learn. Res.* **2003**, *3*, 993–1022.
11. Sia, S.; Dalmia, A.; Mielke, S.J. Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 1728–1736.
12. Grootendorst, M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv* **2022**, arXiv:2203.05794. [CrossRef]
13. Bernasconi, A.; Invernici, F.; Ceri, S. TETYS: Towards the Next-Generation Open-Source Web Topic Explorer. In Proceedings of the CEUR PROCEEDINGS, CEUR-WS, Limassol, Cyprus, 3–7 June 2024; Volume 3692, pp. 26–33.
14. Invernici, F.; Bernasconi, A.; Ceri, S. Exploring the evolution of research topics during the COVID-19 pandemic. *Expert Syst. Appl.* **2024**, *252*, 124028. [CrossRef]
15. Greene, D.; Cross, J.P. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Anal.* **2017**, *25*, 77–94. [CrossRef]
16. Ristilä, A.; Elo, K. Observing political and societal changes in Finnish parliamentary speech data, 1980–2010. with topic modelling. *Parliam. Estates Represent.* **2023**, *43*, 149–176. [CrossRef]
17. Herget, K.; Yeung, J.E.; Rufrano, M.; Alegre, T. Exploring Semantic-Thematic Fields and Lexical Patterns in Parliamentary Debates—Topic Modeling Across Comparable Corpora. 2025. Available online: https://gredos.usal.es/handle/10366/164227 (accessed on 27 March 2026).
18. O'Neill, J.; Robin, C.; O'Brien, L.; Buitelaar, P. An analysis of topic modelling for legislative texts. In Proceedings of the CEUR Workshop Proceedings, London, UK, 16 June 2017.

19. Vīksna, R.; Kirikova, M.; Kiopa, D. Exploring the use of topic analysis in latvian legal documents. In Proceedings of the First International Workshop "CAiSE for Legal Documents" (COUrT 2020) Co-Located with the 32nd International Conference on Advanced Information Systems Engineering (CAiSE 2020), Grenoble, France, 9 June 2020; Volume 2690, pp. 39–47.
20. Silveira, R.; Fernandes, C.G.G.; Monteiro Neto, J.A.; Furtado, J.J.V.P.; Pimentel Filho, J.E. Topic Modelling of Legal Documents via LEGAL-BERT. In Proceedings of the First International Workshop RELATED—Relations in the Legal Domain Co-Located with the 18th International Conference on Artificial Intelligence and Law (ICAIL), São Paulo, Brazil, 25 June 2021.
21. Aguiar, A.; Silveira, R.; Furtado, V.; Pinheiro, V.; Neto, J.A.M. Using topic modeling in classification of Brazilian lawsuits. In *Proceedings of the International Conference on Computational Processing of the Portuguese Language*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 233–242.
22. Didwania, K.; Toshniwal, D.; Agarwal, A. Unveiling Themes in Judicial Proceedings: A Cross-Country Study Using Topic Modeling on Legal Documents from India and the UK. In Proceedings of the Joint Ontology Workshops (JOWO) Co-Located with the 14th International Conference on Formal Ontology in Information Systems (FOIS), Enschede, The Netherlands, 15–19 July 2024.
23. Wang, R.; Wang, Y.; Liu, X.; Huang, H.; Sun, G. Bridging spherical mixture distributions and word semantic knowledge for Neural Topic Modeling. *Expert Syst. Appl.* **2024**, *256*, 124850. [CrossRef]
24. Lei, Z.; Liu, H.; Yan, J.; Rao, Y.; Li, Q. NMTF-LTM: Towards an Alignment of Semantics for Lifelong Topic Modeling. *IEEE Trans. Knowl. Data Eng.* **2023**, *35*, 10616–10632. [CrossRef]
25. Zhang, D.C.; Lauw, H.W. Topic Modeling on Document Networks With Dirichlet Optimal Transport Barycenter. *IEEE Trans. Knowl. Data Eng.* **2024**, *36*, 1328–1340. [CrossRef]
26. Luo, Z.; Liu, L.; Ananiadou, S.; Xie, Q. Graph Contrastive Topic Model. *Expert Syst. Appl.* **2024**, *255*, 124631. [CrossRef]
27. Mu, Y.; Dong, C.; Bontcheva, K.; Song, X. Large language models offer an alternative to the traditional approach of topic modelling. *arXiv* **2024**, arXiv:2403.16248. [CrossRef]
28. Pham, C.M.; Hoyle, A.; Sun, S.; Resnik, P.; Iyyer, M. TopicGPT: A Prompt-based Topic Modeling Framework. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Mexico City, Mexico, 16–21 June 2024; pp. 2956–2984.
29. Xu, W.; Hu, W.; Wu, F.; Sengamedu, S. DeTiME: Diffusion-Enhanced Topic Modeling using Encoder-decoder based LLM. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 6–10 December 2023; pp. 9040–9057.
30. Colombo, A.; Bernasconi, A.; Ceri, S. An LLM-assisted ETL pipeline to build a high-quality knowledge graph of the Italian legislation. *Inf. Process. Manag.* **2025**, *62*, 104082. [CrossRef]
31. Theocharopoulos, P.C.; Anagnostou, P.; Georgakopoulos, S.V.; Tasoulis, S.K.; Plagianakos, V.P. Large language models for efficient topic modeling. *Neural Comput. Appl.* **2025**, *37*, 24421–24439. [CrossRef]
32. Truica, C.O.; Radulescu, F.; Boicea, A. Comparing different term weighting schemas for topic modeling. In Proceedings of the 2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Timisoara, Romania, 24–27 September 2016; pp. 307–310.
33. Wang, H.; Prakash, N.; Hoang, N.K.; Hee, M.S.; Naseem, U.; Lee, R.K.W. Prompting large language models for topic modeling. In Proceedings of the 2023 IEEE International Conference on Big Data (BigData), Sorrento, Italy, 15–18 December 2023; pp. 1236–1241.
34. Li, X.; Zhang, A.; Li, C.; Ouyang, J.; Cai, Y. Exploring coherent topics by topic modeling with term weighting. *Inf. Process. Manag.* **2018**, *54*, 1345–1358. [CrossRef]
35. Dieng, A.B.; Ruiz, F.J.; Blei, D.M. Topic modeling in embedding spaces. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 439–453. [CrossRef]
36. Invernici, F.; Bernasconi, A.; Curati, F.; Jakimov, J.; Samavi, A. TETYS: Configurable Topic Modeling Exploration for Big Corpora of Text Documents. In Proceedings of the 28th International Conference on Extending Database Technology (EDBT), Barcelona, Spain, 25–28 March 2025; Volume 28, pp. 1114–1117.
37. Honnibal, M.; Montani, I.; Van Landeghem, S.; Boyd, A. spaCy: Industrial-Strength Natural Language Processing in Python. Available online: https://spacy.io/ (accessed on 18 March 2026).
38. Reimers, N.; Gurevych, I. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 4512–4525.
39. Li, Z.; Zhang, X.; Zhang, Y.; Long, D.; Xie, P.; Zhang, M. Towards general text embeddings with multi-stage contrastive learning. *arXiv* **2023**, arXiv:2308.03281. [CrossRef]
40. MTEB Leaderboard. Available online: https://huggingface.co/spaces/mteb/leaderboard (accessed on 18 March 2026).
41. McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2018**, arXiv:1802.03426. [CrossRef]
42. McInnes, L.; Healy, J.; Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2017**, *2*, 205. [CrossRef]

43. Moulavi, D.; Jaskowiak, P.A.; Campello, R.J.; Zimek, A.; Sander, J. Density-Based Clustering Validation. In Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, PA, USA, 24–26 April 2014; pp. 839–847.
44. Terragni, S.; Fersini, E.; Galuzzi, B.G.; Tropeano, P.; Candelieri, A. OCTIS: Comparing and Optimizing Topic models is Simple! In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations, Online, 19–23 April 2021; pp. 263–270.
45. Bianchi, F.; Terragni, S.; Hovy, D. Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Online, 1–6 August 2021; pp. 759–766.
46. Rahimi, H.; Mimno, D.; Hoover, J.; Naacke, H.; Constantin, C.; Amann, B. Contextualized Topic Coherence Metrics. In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, 17–22 March 2024; pp. 1760–1773.
47. Boehmer, N.; Bredereck, R.; Peters, D. Rank aggregation using scoring rules. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 5515–5523.
48. Zheng, L.; Chiang, W.L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E.; et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 46595–46623.
49. Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv* **2025**, arXiv:2507.06261. [CrossRef]
50. Ye, J.; Wang, Y.; Huang, Y.; Chen, D.; Zhang, Q.; Moniz, N.; Gao, T.; Geyer, W.; Huang, C.; Chen, P.Y.; et al. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv* **2024**, arXiv:2410.02736. [CrossRef]
51. Li, D.; Sun, R.; Huang, Y.; Zhong, M.; Jiang, B.; Han, J.; Zhang, X.; Wang, W.; Liu, H. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv* **2025**, arXiv:2502.01534. [CrossRef]
52. Colombo, A.; Bernasconi, A.; Bellomarini, L.; Guiso, L.; Michelacci, C.; Ceri, S. LegisSearch: Navigating legislation with graphs and large language models. *Artif. Intell. Law* **2025**, 1–27. [CrossRef]