

Automated Annotation via Structural Alignment for Hierarchical Classification of Classical Fiqh Texts

Hazim Baroudi¹ , Wassim Ammar², Farid Bouchiba², Shadha Karoumi²,
Christian Müller², and Fabrice Rossi¹ 

¹ CEREMADE, Université Paris Dauphine-PSL, CNRS, Paris, France

² Institut de recherche et d’histoire des textes (IRHT), CNRS, Aubervilliers, France

Abstract

This work presents an automated pipeline for the structural organization and classification of classical Arabic legal corpora. We develop a hierarchical classification framework that segments unpunctuated legal prose into fine-grained semantic units and categorizes them according to the inherent taxonomy of Islamic jurisprudence. To address the scarcity of available segment level annotations for Islamic texts, we use the highly structured Hanafi legal compendium *Mukhtasar al-Qudrī* as a reference. Other texts are aligned to this manual using optimal transport. This process is used together with the SAT segmenter to annotate segments, using the structure of the reference text (chapters and sections) as hierarchical labels. This enables us to train a deep learning model which incorporates hierarchical constraints to ensure consistency across different levels of legal categorization. This approach provides a scalable solution for the large-scale analysis of historical jurisprudence.

Keywords: NLP, Optimal Transport, Text Alignment, Hierarchical Classification

1 Introduction

This work is part of an ERC project, which investigates the evolution of legal authority in Muslim religious law. Sharia, historically perceived as timeless divine law, has been subject to adaptations in legal practice, creating a paradox between immutable rules and their practical application. To study this evolution, the project analyzes a large corpus of legal texts written between the 8th and 19th centuries, with the ultimate goal of identifying specific legal rulings and tracking their development over time.

Classical Arabic presents significant challenges for automated analysis. Manuscripts and classical editions rarely utilize punctuation or short vowels, leading to frequent ambiguity at both the sentence and word levels. Furthermore, the specialized nature of religious and legal vocabulary, combined with Arabic’s rich morphology, complicates fundamental tasks such as text segmentation and term identification.

In previous work, we addressed the foundational task of identifying the global structure of general treatises by focusing on the automatic alignment of chapters across legal texts [3]. By employing partial optimal transport, we successfully mapped heterogeneous treatises onto a reference hierarchy, allowing for the automated thematic labeling of large volumes of unstructured legal text.

The present paper extends this framework toward a much more granular level of analysis. While global alignment provides the thematic context of a Kitāb (section) or a Bāb (subsection), the next critical step for legal history is to categorize the underlying text at the level of the individual legal rule. Our current objective is to move beyond coarse paragraph-level blocks to segment the text into finer semantic units. This pipeline establishes the necessary groundwork for future work aimed at isolating each ruling as a complete logical unit, including its primary normative statement, its underlying conditions (shurūt), and its eventual exceptions.

To achieve this, we propose an end-to-end pipeline that moves from global structural alignment to deep hierarchical classification. First, we enhance the textual granularity by employing a

Funded by the European Union (ERC, MCILRaP, Grant No 101141682).

segmentation model adapted via Low-Rank Adaptation (LoRA) to handle the specificities of unpunctuated classical prose. Second, we leverage the labels generated by our structural alignment to train a hierarchical neural classifier. By integrating a hierarchy-guided objective, the model learns to navigate a complex legal taxonomy while respecting the inherent logical organization of Islamic jurisprudence. This approach provides a robust tool for researchers to index and query classical legal content, enabling the precise extraction of legal segments across different schools of thought, even in texts that cannot be aligned to a reference manual for instance because of their partial nature (focusing only on a subset of Islamic jurisprudence).

2 Related Work

Several works have addressed hierarchical text classification for the Arabic language using both classical machine learning and deep learning approaches. For a broader overview of hierarchical text classification for Arabic, we refer the reader to [1] and [8]; we focus here on the works most directly related to our setting.

Aljedani et al. [1] proposed the HMATC model for hierarchical multi-label Arabic text classification, optimizing the HOMER algorithm by combining a Label Powerset classifier with Support Vector Machines and a balanced k -means clustering strategy over 578 hierarchical labels derived from a corpus of Islamic fatwa requests. Their system achieves competitive predictive performance with a multi-label accuracy of 0.758 and a micro-averaged F-measure of 0.853, while significantly reducing computational cost compared to flat multi-label baselines such as Binary Relevance and Classifier Chains. However, their approach is designed for document-level classification over a broad corpus of Islamic religious content spanning jurisprudence, theology, and doctrine (*aqida*), resulting in a label space of 578 categories that does not align with the fine-grained jurisprudential taxonomy we target.

Hamzaoui et al. [8] proposed a deep learning model combining BERT and BiLSTM for hierarchical Arabic text classification. Evaluated on the WiHArD dataset [4], a Wikipedia-derived hierarchy of three top-level and nine fine-grained categories, their architecture achieves a hierarchical F-score of 0.772, demonstrating the superiority of transformer-based contextualized representations over classical approaches such as Doc2Vec with Decision Trees for capturing the structural dependencies inherent in hierarchical classification. Nevertheless, their taxonomy is derived from Wikipedia and covers general-domain topics such as academic subjects and encyclopedic categories, with no relation to religious or legal Arabic.

Moreover, unlike these works which rely on manually annotated datasets, we automatically construct our training data through structural alignment.

3 Methodology

Our goal is to automatically assign each sentence-level segment of a classical Arabic legal treatise to one or more nodes in a predefined jurisprudential hierarchy. This hierarchy is derived from the *Mukhtasar al-Qudrī*, a concise and highly structured Hanafi legal compendium that organizes Islamic law into two levels: broad thematic sections (*Kitābs*) covering major legal domains such as prayer, commercial transactions, or criminal law, and finer subsections (*Bābs*) that partition each *Kitāb* into specific legal topics. This two-level structure serves as the reference ontology against which all other treatises in our corpus are aligned and classified.

The overall pipeline proceeds in three stages. First, we use partial optimal transport to automatically align chapters of heterogeneous treatises against this reference hierarchy, producing silver-standard labels at the chapter level. Second, we apply a LoRA-adapted segmentation model to split each chapter into sentence-level units. Third, we train a hierarchical neural classifier to

assign each segment to its corresponding node in the *Qudrī* hierarchy, taking into account the parent-child relationships between *Kitābs* and *Bābs*.

3.1 Structural Alignment via Partial OT

Following the methodology established in our previous work [3], we perform an automated structural alignment between the heterogeneous treatises in the Maktaba Shamela (see Section 4.1) and our reference ontology. We represent the textual units as term-frequency vectors to capture the specific legal terminology.

The alignment is formulated as a Partial Optimal Transport (POT) problem [5; 6]. Optimal Transport provides a principled framework for comparing probability distributions by computing the minimum cost of transporting mass from one distribution to another. The partial variant is particularly suited to our task as it relaxes the constraint that all mass must be transported, allowing only the most relevant portions of the source text to be matched against the reference hierarchy while discarding chapters with no counterpart in the reference book.

To adapt this framework to our classification needs, we compute the transport plan T between the content of a source book and a reference book. Each source segment s_i is then assigned the label of the reference segment r_j that receives the maximum transported mass:

$$r_i = \begin{cases} \text{label}(r_{j^*}), & \text{if } T_{ij^*} \geq \tau, \quad j^* = \arg \max_j T_{ij} \\ \text{discarded}, & \text{otherwise} \end{cases}$$

where T_{ij} is the transport mass from source segment s_i to reference segment r_j , $\text{label}(r_j)$ is the label of the reference segment, and τ is a confidence threshold. This approach produces a high-quality “silver-standard” dataset, where only the most reliable matches inherit a label from the reference hierarchy.

The quality of the resulting silver-standard labels is validated against a manually annotated alignment between the *Minhāj al-Tālibīn* and the *Mukhtasar al-Qudrī*, which serves as our gold-standard reference for evaluating the alignment step.

3.2 Granular Segmentation with Adapted Learning

Once the chapters are aligned, we perform a refined segmentation to produce consistent training samples for the subsequent classification stage. Classical Arabic legal texts present a significant challenge for standard NLP tools due to the high density of the prose and the general absence of modern punctuation. To address this, we employ a segmentation model [7] specifically designed to identify sentence and segment boundaries in unpunctuated text.

Rather than using Generative AI for the entire corpus, we restricted its role to the creation of a small, high-quality reference set. We curated 80 specialized training segments from various chapters of *Mukhtaṣar Khalīl* by Sheikh Khalīl ibn Ishāq al-Jundī. These examples were generated using ChatGPT (GPT-4o) with a zero-shot expert-role prompt: “You are an expert in Islamic jurisprudence (Fiqh); segment the following text into distinct legal rules or units of the same subject, preserving the original text exactly.”

This “silver-standard” data, after manual verification to ensure no structural hallucinations or text alterations occurred, served as the ground truth for fine-tuning the model via Low-Rank Adaptation (LoRA). This strategy allows the SAT model to learn specific boundary markers and transition particles typical of Fiqh discourse (e.g., specific legal formulas or introductory phrases) from a limited but semantically rich dataset. As demonstrated in [7], even a small set of 10 to 100 well-segmented training sentences is sufficient to improve segmentation performance considerably for specialized domains, effectively bridging the gap between raw classical prose and the structured input required for the HGCLR classifier.

3.3 Hierarchical Classification Model

For the final stage of our pipeline, we apply the hierarchical text classification model from [9] to our data. The model uses a transformer-based encoder and contrastive learning to account for the hierarchical relationships among the 106 target classes, ensuring that parent-child relationships between Kitābs (sections) and Bābs (subsections) are respected.

4 Experimental Setup

To evaluate the effectiveness of our end-to-end pipeline, we conducted experiments focusing on both alignment accuracy and hierarchical classification performance. This section details the datasets, implementation choices, and the parameters used for the study.

4.1 Datasets and Corpus Statistics

Our experimental corpus (see Table 6) is composed of 48 classical legal treatises covering the four primary Sunni schools of jurisprudence: Hanafi, Maliki, Shafi’i, and Hanbali. We employ the *Mukhtasar al-Quduri* (Hanafi) as our gold-standard reference hierarchy. This choice is motivated by the *Mukhtasar*’s concise and highly standardized structure, which serves as a pivot ontology for aligning more voluminous and structurally diverse works.

Metric	Value
Total Reference Classes	118
Sections (<i>Kitābs</i>)	61
Subsections (<i>Bābs</i>)	57
Max Hierarchy Depth	2

Table 1: Structural properties of the reference hierarchy (*Quduri* mapping).

To evaluate the alignment quality independently of the downstream classification task, we use a manually annotated alignment between the *Minhāj al-Tālibīn* (Shafi’i school) and the *Mukhtasar al-Qudrī* (Hanafi school, reference hierarchy) as our gold standard. This pair was chosen because the two treatises share a broadly similar structure despite belonging to different legal schools, making the alignment non-trivial and representative of the general cross-school matching task.

The annotated corpus is partitioned into training, validation, and test sets using a stratified random split. We first hold out 20% of the data as the test set, then split the remaining 80% into 90% for training and 10% for validation, yielding an approximate 64/16/20 train/valid/test ratio (random seed fixed to 0 for reproducibility).

In addition, the segments of the reference treatise (*Mukhtasar al-Qudrī*) are explicitly included in the training set, as this book defines the target hierarchy and provides clean, unambiguous examples for each jurisprudential category. The full annotated dataset, including segment-level spans and hierarchical labels, will be made publicly available upon publication.

4.2 Implementation Details

For text segmentation, we utilized the `wtpsplit` library with the `sat-121` model, adapted via **LoRA** (Low-Rank Adaptation) to handle the specificities of classical Fiqh prose. The final hierarchical classification task was performed using **bert-base-arabertv2** [2] as the backbone encoder, fine-tuned on the “silver-standard” labels with a hierarchy-guided loss function.

4.3 Hyperparameters

The specific configuration for the alignment (feature extraction and OT), the LoRA-adapted segmentation, and the hierarchical classifier is detailed in Table 2. Hyperparameters for the hierarchical classifier were adopted directly from the original HGCLR model [9], without additional tuning on our data. Specifically, we used the configuration optimized for the Web of Science (WOS) dataset, as its hierarchical structure in terms of depth and label space size is the closest to our jurisprudential taxonomy among the benchmarks reported in the original paper. The other datasets used in that work have significantly different structural properties and their corresponding hyperparameters were therefore not considered. The OT hyperparameters: the confidence threshold τ and the mass ratio m_ratio were tuned via a grid search, with values selected by maximizing chapter-level alignment accuracy against our annotated ground truth. Final values for all parameters are reported in Table 2.

Component	Hyperparameter	Value
Feature Ext.	Max Features	5000
	N-grams	(1, 1)
Partial OT	Lambda (λ)	0.0005
	Threshold	0.002
	Mass Ratio (m_ratio)	0.85
Classifier	Base Encoder	arabertv2-base
	Learning Rate	2×10^{-5}
	Batch Size	12
	Lambda (λ)	0.05
	Threshold	0.02

Table 2: Key hyperparameters for the alignment, segmentation, and classification stages.

4.4 Evaluation Metrics

To assess the performance of our hierarchical classification pipeline, we employ standard multi-label and hierarchical metrics. Given the unbalanced nature of Islamic legal treatises where certain chapters like *Kitāb al-Salāt* (Prayer) are significantly more represented than others, we prioritize **Macro-averaged F1-score** to capture performance across all legal categories, alongside **Micro-averaged F1-score** for overall system effectiveness.

We additionally report two hierarchically-aware metrics. **Exact Match Accuracy (EMA)** measures the proportion of segments for which the model correctly predicts both the level-1 (Kitāb) and level-2 (Bāb) labels simultaneously. **Path Consistency** measures the proportion of predictions that are coherent with respect to the hierarchy, i.e., a predicted level-2 label is only valid if its parent level-1 label is also predicted. All metrics are computed at a classification threshold of 0.5, meaning a label is assigned when the model’s predicted probability exceeds this value.

Furthermore, we evaluate the model at different levels of the hierarchy, as reported in Table 3.

5 Results and Discussion

To provide a qualitative illustration of the classifier’s behavior, we analyze its predictions on a set of segments drawn from *Al-Risāla* by Ibn Abi Zayd Al-Qayrawānī, a Maliki legal treatise. This analysis aims to show that the model captures meaningful thematic structure across texts, and to identify cases where predictions are ambiguous or incorrect.

Table 3: Performance Metrics for Hierarchical Classification (HGCLR) at Threshold 0.5

Metric	Score
Micro-F1	58.41
Macro-F1	42.01
Exact Match Accuracy (EMA)	30.66
Path Consistency	99.66

We present two complementary visualizations. Figure 4 shows a heatmap of the classifier’s confidence scores across 13 consecutive segments, where each row corresponds to a segment and each column to a jurisprudential category. This allows us to observe how the model’s predictions evolve across the text and whether thematic transitions are correctly detected. Figure 3 provides a second example focusing on a different passage, illustrating a contrasting prediction pattern.

The confidence scores across the 13 segments reveal a clear thematic shift across the passages. Segments 1–6 relate to criminal law and legal procedures: Segment 1 scores 68.0% on *kitāb al-janāyāt* (bodily offences), Segment 2 reaches 99.3% and 98.4% on *kitāb al-salāt* and *bāb al-shahīd*, and Segment 3 achieves 90.5% on *kitāb al-janāyāt*. From Segment 7 onward, the classifier shifts toward family and personal status law. Segment 7 scores 99.8% on *kitāb al-hajj* and 63.5% on *kitāb al-nafāqāt*, while Segments 9 and 10 concentrate on *kitāb al-janāyāt* (99.5%) and *bāb al-hudūd* (99.9%) respectively, consistent with the passages regarding waiting periods (*’idda*) and mourning (*ihdād*). Segment 6 exhibits the most distributed confidence scores (63.2% / 49.9% / 44.5%), as it corresponds to a transitional passage located at the boundary between criminal procedure and personal status rulings. These results suggest that the model successfully captures the thematic organization of legal texts.

Figure 3 presents the confidence scores of our hierarchical classifier across the 15 segments detailed in Table 4.

Most segments receive a clearly dominant predicted label, Segment 13 reaches 99.0% on *kitāb al-hudūd* (criminal penalties), Segment 15 scores 99.7% on *kitāb al-sariqah* (theft), and Segment 8 achieves 96.2% on *kitāb al-ma‘āqil* (collective responsibility for blood money). Several segments exhibit more distributed confidence, notably Segment 1 (55.6% / 22.3%) and Segment 11 (40.0% / 35.4%), which correspond to transitional passages straddling two adjacent legal topics.

6 Error Analysis

Table 3 shows that the model achieves a Micro-F1 of 58.41% and a Macro-F1 of 42.01%. The gap between these two scores reflects the class imbalance present in the corpus: the model performs well on frequent categories but struggles with underrepresented ones, which penalizes the Macro-F1. The Exact Match Accuracy of 30.66% indicates that simultaneous correct prediction of both L1 and L2 labels remains challenging, which is expected given the fine-grained nature of the jurisprudential taxonomy. Notably, the Path Consistency of 99.66% confirms that the hierarchical constraints enforced by HGCLR are highly effective: almost all predictions respect the parent-child relationships of the label hierarchy.

A primary source of error stems from the inherent limitations of the unsupervised annotation process. We relied on Partial Optimal Transport (POT) because no manually annotated labels were available for this specific corpus. However, the POT performance is highly dependent on the cost matrix calibration, which in our case was based on TF-IDF vectorization. While this approach effectively identifies broad thematic shifts at the chapter level (L1), it struggles to capture the more granular lexical distinctions required for sub-section alignment (L2). This lack of fine-grained calibration in the POT phase introduced "label noise" into our silver-standard training data, effectively capping the classifier’s performance. The HGCLR model is thus trained on an imperfect teacher, which explains the moderate Exact Match scores despite the high structural consistency of the predictions.

This section investigates two complementary sources of classification error: the effect of segment length on model performance (Section 6.1), and the distribution of correct predictions across jurisprudential categories (Section 6.2). Together, these analyses shed light on the main limitations of the current system and suggest directions for improvement.

6.1 Impact of Segment Length

We investigate whether segment length affects classification performance. Our hypothesis is that very short segments lack sufficient lexical context for reliable jurisprudential classification, since the discriminative markers of Islamic legal discourse typically require several tokens to be unambiguous. To test this, we evaluate the model on subsets of the test data filtered by a minimum token length threshold, ranging from 0 to the maximum observed length in steps of 5 tokens. Figure 1 reports Micro-F1, Macro-F1, and Exact Match Accuracy (EMA) for each threshold, overlaid with a histogram showing the number of segments remaining at each threshold.

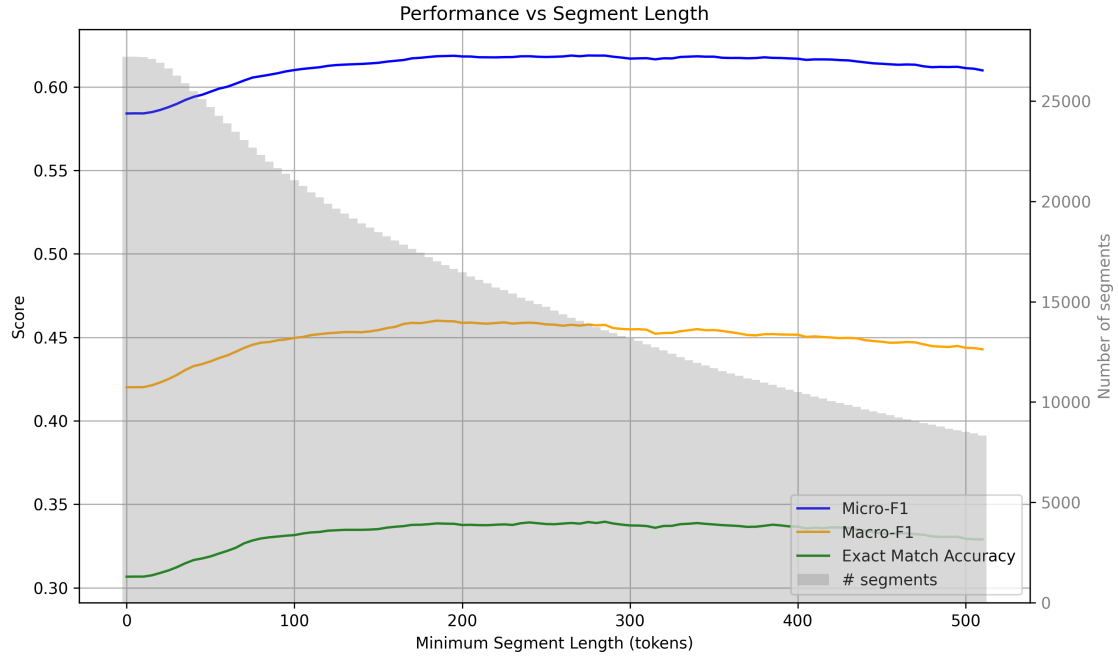


Figure 1: Classification performance as a function of minimum segment length (left axis), overlaid with the distribution of segment lengths in the test set (right axis, gray histogram). Scores improve consistently as very short segments are excluded, confirming that longer segments provide more discriminative context for the classifier.

The results suggest a positive trend between segment length and classification performance, particularly for shorter units. The gray area represents the number of segments evaluated at each threshold, showing that the count decreases sharply as the threshold increases. All three metrics show modest but consistent improvements between 0 and 100 tokens, with Micro-F1 rising by approximately 3 percentage points. The curves then plateau and reach their peak around 285–290 tokens, suggesting that this range captures sufficient discriminative lexical context. Beyond 400 tokens, a slight decline is observed. While this may be partly due to the reduced number of segments available at that threshold, it could also indicate that very long segments become inherently noisier by potentially encompassing multiple legal topics, thus challenging the classifier’s precision. Nevertheless, these results indicate that very short segments, which represent 9.2

6.2 Per-Class Analysis

We now examine which jurisprudential categories are correctly predicted by the model. Table 5 reports, for each class, the per-class F1 score as well as the proportion of examples in the test set (% Total).

At the L1 level, performance is strongly correlated with class frequency: *kitāb al-salāt* (prayer) achieves the highest F1 of 0.864, reflecting its dominance in the training data (29.72% of examples). Underrepresented categories such as *kitāb al-khunthā* (F1=0.145, 0.53% of examples) are nearly never correctly predicted. The same pattern holds at the L2 level, where high-frequency subcategories achieve higher F1 scores while rare subcategories such as *bāb zakāt al-fiddā* (F1=0.019, 0.37% of examples) are almost never recovered. These results highlight the impact of class imbalance and motivate future work on balancing strategies.

These per-class F1 scores are consistent with the global Micro-F1 of 58.41% and Macro-F1 of 42.01% reported in Table 3. The gap between these two global metrics reflects the same imbalance:

frequent classes achieve high F1 and drive the Micro-F1 up, while rare classes pull the Macro-F1 down.

7 Conclusion

This paper presented an end-to-end pipeline for the automated annotation and hierarchical classification of classical Arabic legal texts. By combining partial optimal transport for structural alignment, LoRA-adapted segmentation for unpunctuated prose, and a hierarchy-guided neural classifier, we demonstrated that it is possible to assign fine-grained jurisprudential labels to classical Fiqh segments without manual annotation.

Experimental results show that the model achieves a Micro-F1 of 58.41 and a Macro-F1 of 42.01, with a path consistency of 99.66%. These scores reflect not only the classification difficulty but also the quality of the silver-standard labels produced by the optimal transport alignment, any noise introduced during the annotation stage propagates directly to the classifier. The high path consistency nonetheless confirms that hierarchical constraints are effectively enforced. The confidence analysis further reveals that the model correctly captures the high-level thematic organization of legal treatises, distinguishing criminal law categories from personal status law despite the absence of explicit structural markers in the source text.

Future work will focus on two complementary directions. First, we aim to refine the classification hierarchy itself by restructuring the label space around broad thematic clusters such as worship, transactions, and criminal law before descending to finer subsections, which should improve both alignment quality and classifier performance. Second, we plan to move beyond chapter-level classification toward the automatic identification of the functional components within each ruling, such as conditions, exceptions, causes, and legal statuses. Additionally, the publicly available dataset introduced by [1], which provides a large-scale hierarchical label space for Arabic religious content, represents a promising resource for pre-training our classifier before fine-tuning on our jurisprudential taxonomy. Exploring such transfer learning strategies across Arabic religious datasets constitutes a natural direction for future work. Furthermore, our evaluation of the alignment step is currently limited to a single manually annotated book pair (*Minhāj al-Tālibīn Mukhtasar al-Qudrī*); extending the gold-standard annotation to additional treatises and school pairs constitutes a natural direction for future work.

References

- [1] Aljedani, Nawal, Alotaibi, Reem, and Taileb, Mounira. “HMATC: Hierarchical multi-label Arabic text classification model using machine learning”. In: *Egyptian Informatics Journal* 22, no. 3 (2021), pp. 225–237. ISSN: 1110-8665. DOI: <https://doi.org/10.1016/j.eij.2020.08.004>. URL: <https://www.sciencedirect.com/science/article/pii/S1110866520301523>.
- [2] Antoun, Wissam, Baly, Fady, and Hajj, Hazem. “AraBERT: Transformer-based Model for Arabic Language Understanding”. In: *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, p. 9.
- [3] BAROUDI, HAZIM, Ammar, Wassim, Bouchiba, Farid, Marmardji, Shadha, Mueller, Christian, and Rossi, Fabrice. “Alignment of Islamic Legal Texts”. In: Jan. 2026, pp. 241–246. DOI: 10.14428/esann/2026.ES2026-334.

- [4] Bouchiha, Djelloul, Bouziane, Abdelghani, Doumi, Nouredine, Berbouchi, Farouk Omar, Kebir, Aymen Abdelghani, Mebarki, Nihad, and Benameur, Badiâ Achouak. “WiHArD: Wikipedia Based Hierarchical Arabic Dataset for Text Classification”. In: *2024 4th International Conference on Embedded & Distributed Systems (EDiS)*. 2024, pp. 115–118. DOI: 10.1109/EDiS63605.2024.10783418.
- [5] Caffarelli, Luis A. and McCann, Robert J. “Free boundaries in optimal transport and Monge-Ampère obstacle problems”. English. In: *Ann. Math. (2)* 171, no. 2 (2010), pp. 673–730. ISSN: 0003-486X. DOI: 10.4007/annals.2010.171.673. URL: annals.princeton.edu/annals/2010/171-2/p02.xhtml.
- [6] Chapel, Laetitia, Alaya, Mokhtar Z., and Gasso, Gilles. “Partial Optimal Transport with Applications on Positive-Unlabeled Learning”. In: (Dec. 2020).
- [7] Frohmann, Markus, Sterner, Igor, Vulić, Ivan, Minixhofer, Benjamin, and Schedl, Markus. “Segment Any Text: A Universal Approach for Robust, Efficient and Adaptable Sentence Segmentation”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 11908–11941. DOI: 10.18653/v1/2024.emnlp-main.665. URL: <https://aclanthology.org/2024.emnlp-main.665/>.
- [8] Hamzaoui, Benamar, Bouchiha, Djelloul, Bouziane, Abdelghani, and Doumi, Nouredine. “BERT-BiLSTM model for hierarchical Arabic text classification”. In: *Indonesian Journal of Electrical Engineering and Informatics* 12, no. 3 (2024). Cited by: 2; All Open Access, Gold Open Access, pp. 690–699. DOI: 10.52549/ijeei.v12i3.5659. URL: <https://www.scopus.com/inward/record.uri?eid=2-s2.0-85207514660&doi=10.52549/2fijeei.v12i3.5659&partnerID=40&md5=c704c2ee1c57b53bf175e98c1c5c4b14>.
- [9] Wang, Zihan, Wang, Peiyi, Huang, Lianzhe, Sun, Xin, and Wang, Houfeng. “Incorporating Hierarchy into Text Encoder: a Contrastive Learning Approach for Hierarchical Text Classification”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ed. by Smaranda Muresan, Preslav Nakov, and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7109–7119. DOI: 10.18653/v1/2022.acl-long.491. URL: <https://aclanthology.org/2022.acl-long.491/>.

Appendices

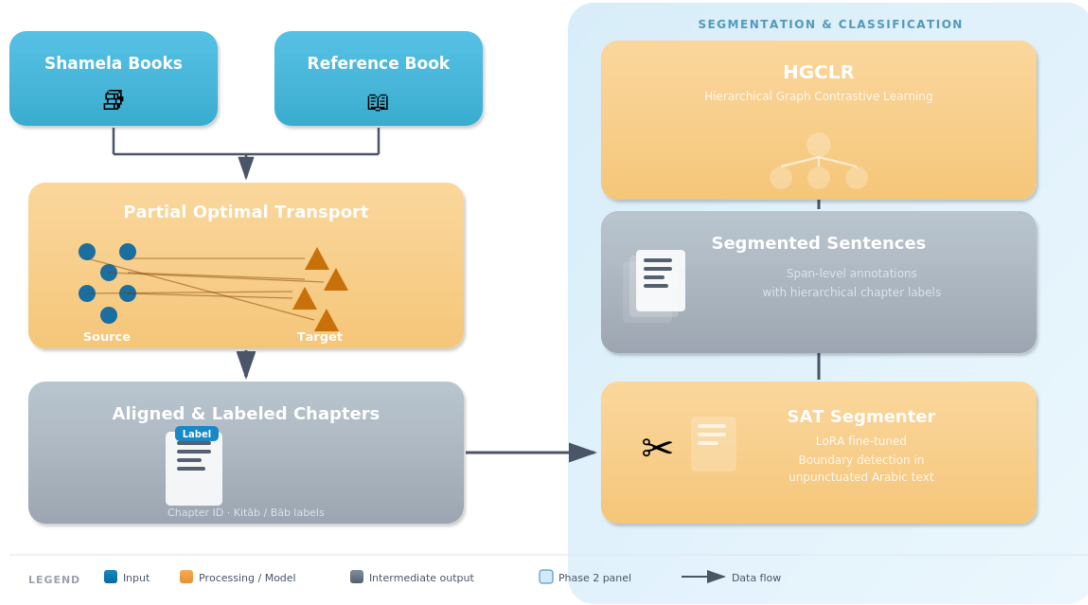


Figure 2: Illustration of the alignment process.

Text	Seg.
ولا تقتل نفس بنفس إلا ببيئة عادلة أو باعتراف أو بالقسامة إذا وجبت يقسم الولاة خمسين يمينا ويستحقون الدم ولا يحلف في العمد أقل من رجلين ولا يقتل بالقسامة أكثر من رجل واحد	1
إنما تجب القسامة بقول الميت دمي عند فلان أو بشاهد على القتل أو بشاهدين على الجرح وإذا نكل مدعو الدم حلف المدعي عليهم خمسين يمينا ولو ادعى القتل على جماعة حلف كل واحد خمسين يمينا	2
لا تحلف امرأة في العمد وتحلف الورثة في الخطأ بقدر ما يرثون من الدية من رجل أو امرأة ويحلفون في القسامة قايما ولا قسامة في جرح ولا في عبد ولا بين أهل الكتاب أو قتل بين الصفيين	3
وقتل الغيلة لا عفو فيه وللرجل العفو عن دمه العمد إن لم يكن قتل غيلة وعفوه عن الخطأ في ثلثه وإن عفا أحد البينين فلا قتل ولمن بقي نصيبهم من الدية ومن عفى عنه في العمد ضرب مائة وحبس عاما	4
الدية على أهل الإبل مائة ومن الذهب ألف دينار ومن الورق اثنا عشر ألف درهم ودية العمد خمسة أو مربعة حسب القبول ودية المرأة على النصف من دية الرجل وكذلك الكتابيين والمجوسي دية ثمانمائة درهم	5
في اليدين أو الرجلين أو العينين الدية وفي كل واحدة منهما نصفها وفي الأنف يقطع مارنه الدية وفي السمع والعقل والصلب والأنتيين واللسان والأسنان مقادير محددة من الدية حسب العضو والمنفعة	6
الموضحة ما أوضع العظم وفيها خمس من الإبل والمنقلة ما طار فراشها من العظم والمأمومة ما وصلت إلى الدماغ ففيها ثلث الدية وكذلك الجائفة وليس فيما دون الموضحة إلا الاجتهاد ولا يعقل جرح إلا بعد البرء	7
لا تحمل العقالة قتل عمد ولا اعترافا به وتحمل من جراح الخطأ ما كان قدر الثلث فأكثر وما كان دون الثلث ففي مال الجاني وعمد الصبي والمجنون كالخطأ وذلك على عاقلتها إن بلغ الثلث	8
تقتل المرأة بالرجل والرجل بها ويقتص لبعضهم من بعض في الجراح ولا يقتل حر بعبد ويقتل به العبد ولا يقتل مسلم بكافر ويقتل به الكافر والسائق والقائد والراكب ضامنون لما وطئت الدابة	9
في جنين الحرة غرة عبد أو وليدة تقوم بخمسين دينارا وتورث على كتاب الله ولا يرث قاتل العمد من مال ولا دية ويقتل الخطأ يرث من المال دون الدية وفي جنين الأمة من سيدها ما في جنين الحرة	10
يقتل الزنديق ولا تقبل توبته وهو الذي يسر الكفر ويظهر الإسلام وكذلك الساحر ويقتل من ارتد إلا أن يتوب ويؤخر للتوبة ثلاثا ومن ترك الصلاة جدا لها فهو كالمرتد يستتاب ثلاثا فإن لم يتب قتل	11
المحارب لا عفو فيه إذا ظفر به فإن قتل أحدا فلا بد من قتله وإن لم يقتل فيسع الإمام فيه اجتهاده بين القتل أو الصلب أو القطع من خلاف أو النفي وكل واحد من اللصوص ضامن لجميع ما سلبوه	12
من زنى من حر محصن رجم حتى يموت والإحصان أن يتزوج نكاحا صحيحا ويطأ فإن لم يحصن جلد مائة وعرب عاما وعلى العبد والأمة خمسون جلدة ولا حد على من لم يحتمل ولا يحد وأطىء أمة والده	13
على القاذف الحر الحد ثمانون وعلى العبد أربعون ومن نفى رجلا من نسبه فطليه الحد ومن شرب خمرأ أو نبيذا مسكرا حد ثمانين سكر أو لم يسكر ويجرد المحنود ولا تحد حامل حتى تضع	14
من سرق ربع دينار أو ثلاثة دراهم قطع من حرز ويقطع في ذلك يد الرجل والمرأة والعبد ثم رجليه من خلاف ولا قطع في الخلصة ولا في ثمر معلق ولا في الغنم الراعية حتى تسرق من مراحمها	15

Table 4: Evaluated segments from the chapter on bloodshed and prescribed punishments (*Bāb fī Ahkam al-Dimā’ wa-l-Ĥudūd*)

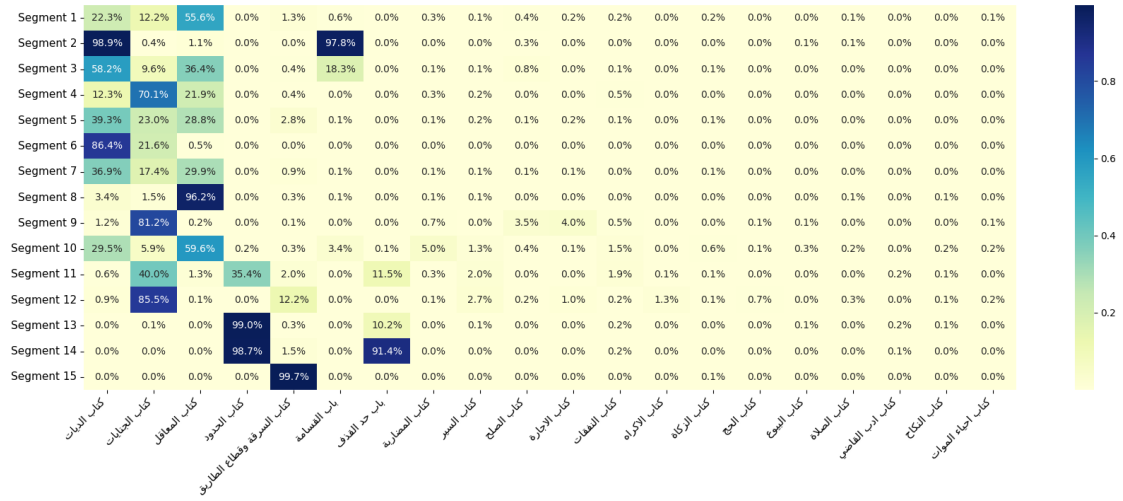


Figure 3: Confidence heatmap across 15 text segments and their predicted jurisprudential categories.

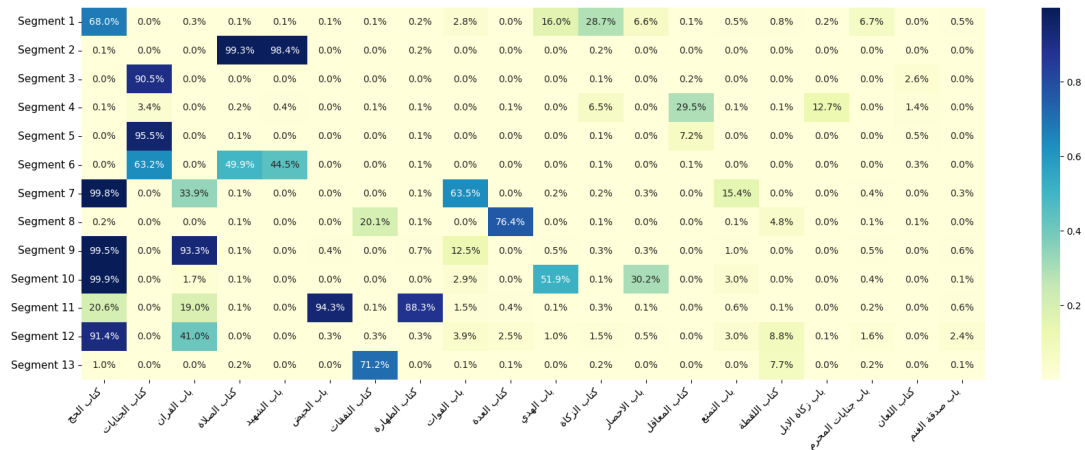


Figure 4: Confidence heatmap across 13 text segments and their predicted jurisprudential categories.

Category (L1)	F1	% Total
<i>kitāb al-salāt</i>	0.864	29.72%
<i>kitāb al-buyū'</i>	0.823	9.78%
<i>kitāb al-tahāra</i>	0.573	7.65%
<i>kitāb al-hajj</i>	0.830	6.44%
<i>kitāb al-zakāt</i>	0.790	6.22%
<i>kitāb al-sūm</i>	0.535	2.02%
<i>kitāb al-talāq</i>	0.564	2.00%
<i>kitāb al-hudūd</i>	0.584	1.64%
<i>kitāb al-nikāh</i>	0.573	1.55%
<i>kitāb al-shahādāt</i>	0.632	1.33%
<i>kitāb al-diyāt</i>	0.469	1.07%
<i>kitāb al-ijāra</i>	0.569	0.86%
<i>kitāb al-janāyāt</i>	0.493	0.75%
<i>kitāb al-rasā'</i>	0.536	1.21%
<i>kitāb al-sariq</i> (theft)	0.689	0.64%
... remaining 46 categories (each < 1% of total)		
<i>kitāb al-khunthā</i>	0.145	0.53%
<i>kitāb al-abāq</i>	0.356	0.09%

Table 5: Per-class F1 and proportion of test examples per L1 category (selection).

Table 6: The full list of the 49 legal works used in our study. Each book is identified by its school, its approximate date of death of the author (AH).

School	Date (AH)	Book (Shamela Identifier)
Hanafi	0428	0428_Quduri_Mukhtasar (Reference)
	0189	0189_Saybani_Asl
	0483	0483_Sarakhsi_Mabsut
	0587	0587_Kasani_Badaic
	0593	0593_Marghinani_Hidaya
	0593	0593_Marghinani_Bidaya
	0683	0683_Musali_Ikhtiyar
	0786	0786_IBabarti_InayaSharhHidaya
	0861	0861_IHumam_FathQadir
	0970	0970_INujaym_BahrRaiq
	1078	1078_ShaykhZadeh_MajmaAnhur
	1252	1252_IAbidin_RaddMuhtar
	1298	1298_Maidani_Lubab
Maliki	0386	0386_IAZaydQayrawani_NawadirCMbab
	0422	0422_QadiAWahhab_Talqin
	0453	0453_Maziri_SarhTalqin
	0463	0463_IAbdBarr_Kafi
	0684	0684_Qarafi_Tanqih
	0684	0684_Qarafi_Dhakhira
	0776	0776_Jundi_Tawdih
	0897	0897_Aabdari_TajIkilil
	0995	0995_Manjur_SharhManhajMuntakhab
	1122	1122_Zarqani_SharhMuwatta
	1230	1230_Dasuqi_SharhKabir
	1335	1335_Azhari_ThamrDani
Shafi'i	0204	0204_Safii_Umm
	0415	0415_IMahamili_Lubab
	0450	0450_Mawardi_Hawi
	0505	0505_Ghazali_Wasit
	0676	0676_Nawawi_Minhaj
	0804	0804_IMulaqqin_UjalatMuhtaj
	0829	0829_TaqiDinHisni_Kifayat
	0973	0973_Huthaymi_Tuhfat
	1004	1004_Ramli_NihayatMuhtaj
	1204	1204_SulaymanJamal_Hashiya
	1310	1310_Dimyati_Hashiya
Hanbali	0311	0311_ABakrKhallal_Wuquf
	0334	0334_Khiraqi_Mukhtasar
	0428	0428_Hashimi_Irshad
	0458	0458_QadiAYacla_Masail
	0597	0597_IJawzi_TahqiqAhadith
	0620	0620_IQudama_Mughni
	0763	0763_IMuflih_FurucTashih
	0884	0884_IMuflih_MubdiSharhMuqni
	0972	0972_INajjarHanbali_Muntaha
	1051	1051_Bahuti_RawdMurbic
	1206	1206_IAWahhab_MajmuHadith
	1319	1319_IYasinLabadi_HashiyaNaylMaarib