ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Alignment of Islamic Legal Texts

Hazim Baroudi[1], Wassim Ammar[2], Farid Bouchiba[2],
Shadha Karoumi Marmardji[2], Christian Müller[2] and Fabrice Rossi[1] *

1- CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine - PSL
Paris, France

2- IRHT-CNRS, Campus Condorcet Paris-Aubervilliers, France

Abstract. This work addresses the automatic alignment of chapters across different Arabic legal texts. We examine two representation strategies: TF-IDF-based lexical embeddings and contextual semantic embeddings generated by AraBERT. These embeddings enable us to assess the semantic proximity between chapters. We then frame the text alignment process as an optimal transport problem, incorporating soft structural constraints. To analyze the impact of method parameters, we use a curated ground-truth dataset derived from a pair of representative texts.

1 Introduction

This work is part of the MCILRaP 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. One of the goals of MCILRaP is to study this evolution via the analysis of a large corpus of legal texts written between the 8th and 19th centuries. One of the key points is to identify rulings and then to track their evolution over time.

Arabic itself presents additional challenges for automatic analysis. Indeed, manuscripts and classical editions rarely use punctuation or short vowels, which often makes sentences and words ambiguous. Moreover, the religious and legal vocabulary is highly specialized and can change meaning depending on the context and too add more difficulties, Arabic's rich morphology and the diversity of word forms complicate basic tasks such as text segmentation, term identification [4].

In the present work, we address part of the ground work needed to build more advanced analyzes. We focus on identifying the global structure of the legal texts studied in the project, more specifically, of general treatises. They cover roughly the same subjects from purification rituals to contract rulings, going through marriage, inheritance, etc. A general subject is covered by a chapter or by a section in a chapter. In contemporary editions of classical texts, those chapters and sections have been identified and reveal the overall structure of each text.

Here, we focus on the automatic alignment of chapters across legal treatises, identifying partial matches between pairs of texts. This approach serves several key purposes. First, it offers a practical way to compare specific treatises at a global level, highlighting the relative importance of different subjects and their logical organization. Second, the resulting distance metric can be used to cluster

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

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

texts or compare groups of treatises. An additional, yet significant, benefit is the ability to compare text representations, as alignments rely on computing distances between chapters. This, in turn, creates opportunities to reconstruct the chapter organization of texts that have not been covered by contemporary editions.

2 Text representation

The foundation of alignment methods is a metric for comparing the objects to be aligned. In our context, we require a way to assess global similarity between texts (i.e., chapters). To this end, we use vector embeddings. Legal treatises are highly structured, as jurists prioritize clarity in their writing. This consistency allows us to use the lexical content of chapters as a robust proxy for their semantic meaning. Because legal discourse typically employs precise and standardized language, ambiguity in terminology or phrasing is minimal, which makes lexical approaches particularly effective.

We use therefore a simple vector text model using standard term-frequency inverse document frequency (TF-IDF) [8]. The texts are preprocessed by removing punctuation, digits, and non-Arabic characters, normalizing spaces, removing common Arabic stopwords, and stripping accents (diacritics). After that, we extract the top P most frequent words in the two texts. Each chapter t is then represented by a vector of dimension m given by

$$t_w = \mathrm{tf}(w,t) \cdot \log \frac{N}{1+\mathrm{df}(w)}, \tag{1}$$

where $\mathrm{tf}(w,t)$ is the frequency of w in chapter t, N is the total number of chapters in the two texts and $\mathrm{df}(w)$ is the number of chapters containing the term w.

As an alternative to this simple and classical approach, we also use the contextual text embeddings generated by AraBERT [1], which can capture paraphrasing or subtle variations in meaning between chapters.

Both approaches represent a chapter by a vector which can be used to compute distances between chapters. We use the cosine similarity for this task.

3 Alignment method

Let us assume we have two legal treatises consisting of m and p chapters, respectively, denoted $A = (a_1, \ldots, a_m)$ and $B = (b_1, \ldots, b_p)$. An alignment is a way to reconcile the two sequences of chapters in such a way that each chapter in A is associated with chapters in B that discuss the same subject, and vice versa. A standard solution would be to use the Levenshtein distance [6], but it is somewhat ill-suited to our setting, as some subjects are treated in full chapters by some jurists and only as a section of a chapter by others. We therefore need a more flexible alignment method.

We propose using a fully flexible matching, in which a given chapter of A can be partially matched to an arbitrary subset of chapters in B. More precisely,

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

an alignment is represented by a matrix $T = (T_{ij})_{1 \leq i \leq m, 1 \leq j \leq p}$ of non-negative terms. T_{ij} indicates "how much" of chapter a_i matches chapter b_j. Given a matrix of matching costs between chapters, C, with the same dimensions as T, the total cost associated with T is $c(T) = \sum_{i,j} T_{ij} C_{ij}$.

Optimizing T under appropriate constraints is an instance of the optimal transport problem [7, 5]. However, a standard balanced transport assumes that all mass from treatise A must be mapped to treatise B. This is suboptimal when chapters are missing or when treatises have different lengths. To address this, we employ partial optimal transport [3, 2] . We assign a weight to each chapter: $\mathbf{u} = (u_i)_{1 \leq i \leq m}$ for treatise A and $\mathbf{v} = (v_j)_{1 \leq j \leq p}$ for B. We then solve the following problem:

$$\min_{T} \sum_{i,j} T_{ij} C_{ij} \quad \text{s.t.} \quad T\mathbf{1}_p \leq \mathbf{u}, \quad T^\top \mathbf{1}_m \leq \mathbf{v}, \quad \mathbf{1}_m^\top T \mathbf{1}_p = m_{trans} \tag{2}$$

where m_{trans} is the total mass to be transported, defined as $m_{trans} = \mathrm{m}_{ratio} \times \min(\sum u_i, \sum v_j)$. This formulation allows the model to leave some chapters unmatched if their matching cost is too high.

To instantiate the problem, we define u_i and v_j as the relative lengths of the chapters measured in words. We then define the cost matrix C as:

$$C_{ij} = (1 - \lambda)\,(1 - \cos(a_i, b_j)) + \lambda \left| \frac{i}{m} - \frac{j}{p} \right|, \tag{3}$$

where a_i and b_j are the vector representations of the chapters. The second term introduces the linear structure as a soft constraint: the cost of matching chapters increases with their relative distance in their respective books.

4 Parameter study

The proposed method uses three key parameters: the vocabulary size P, the trade-off parameter λ, and the transport mass ratio m_{ratio}. In practice, the matrix T_{ij} can be dense, containing many very small values. For result evaluation, we use a decision threshold; a match between chapters a_i and b_j is considered valid only if T_{ij} exceeds this threshold.

To study the effects of these parameters, we use two reference texts[1]. The *Mukhtasar* of al-Quduri is a concise legal treatise written by Imam al-Quduri, a renowned Iraqi scholar and leading figure of the Hanafi school (circa 1000 CE). The *Minhaj al-Talibin* is a comparable treatise written by al-Nawawi, a prominent scholar the Shafi'i school (circa 1250 CE). Both texts are similar in size and structure, containing 56 and 70 chapters respectively. As reference manuals, they are expected to be relatively straightforward to align, focusing on core Islamic rulings. We constructed alignments from A to B (and from B to A) by associating to each chapter in A (or B) all the chapters in B (or A)

[1]The texts used in this study were obtained from the Maktaba Shamela digital library `https://shamela.ws/`.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

that share content with it. We obtain this way a ground truth that consists in a collection of matching pairs of chapters from A and B, where each chapter can appear multiple times.

We assess the quality of a transport plan T, we binarize it according to the chosen threshold and extract all matching pairs (a_i, b_j). This set of pairs is then compared to the ground truth using precision, recall and F1-score. Figure 1 shows the effect of the parameters on these quality metrics for the TF-IDF representation. In summary, the introduction of a small amount of structural constraint $\lambda > 0$ and the fine-tuning of the m_{ratio} improve the results over a straightforward semantic only optimal transport. TF-IDF is favored by the use of a large vocabulary and the transport plan thresholding in necessary to delete spurious noisy match but should be kept to a minimum. Results with AraBERT (not shown here) are slightly worse than the ones obtained by TF-IDF, confirming our hypothesis on the very specific vocabulary used by jurists.

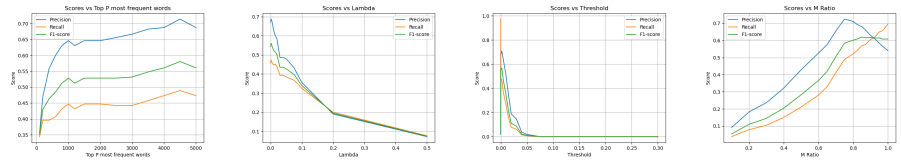

Fig. 1: Evolution of performance metrics (F1-score, Recall, Precision) with the TD-IDF representation

5 Using the alignment

To illustrate the interest of the alignment procedure, we aligned al-Quduri's Mukhtassar with another text, using the optimal parameters according to the ground truth. We used the *At-Talquin* written by Qadi Abd al-Wahhab a seminal figure of the Iraqi Maliki school (circa 1000 CE).

Figure 2 illustrates the TF-IDF alignment between Quduri's *Mukhtassar* and al-Wahhab's *At-Talqin*. Despite distinct *madhahib* and a significant difference in chapter counts (only 22 in *At-Talqin*), the alignment remains globally consistent. The Five Pillars converge perfectly. Relevant thematic links emerge: *Hajj* aligns with *Sacrifice* (*hady* and ritual shaving), and fasting with *dhihar* via the shared requirement of *kafara* (expiation). Structurally, while Quduri is more granular, *At-Talqin* utilizes comprehensive chapters (marriage, justice) that correctly redistribute across Quduri's divided sections. Themes absent in al-Wahhab's work, such as *Kafala* (Suretyship/Guardianship) or *Mafqoud* (Missing Person), are accurately identified as `deleted`. However, some alignment errors occur due to the high sensitivity of the m_{ratio}, which struggles to perfectly balance the significant disparity in total length between the two works.

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

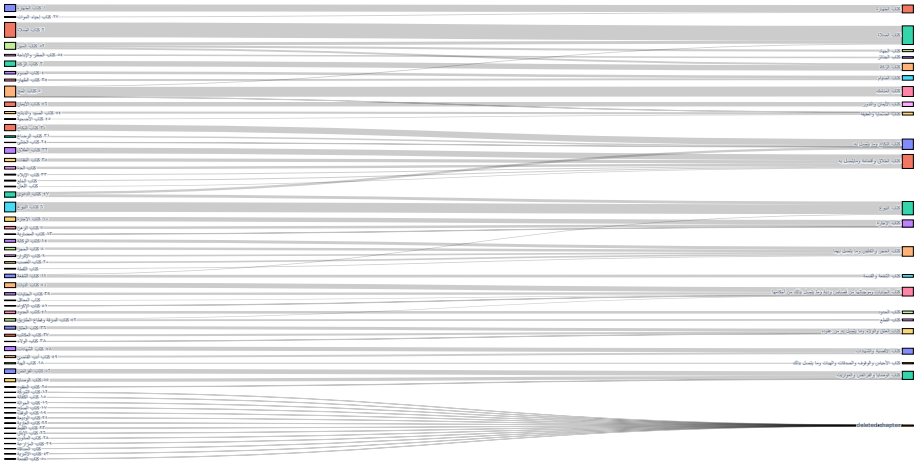

Fig. 2: Chapter Alignment: *Mukhtassar* of al-Quduri → *At-Talquin* of Qadi Abd al-Wahhab (TF-IDF)

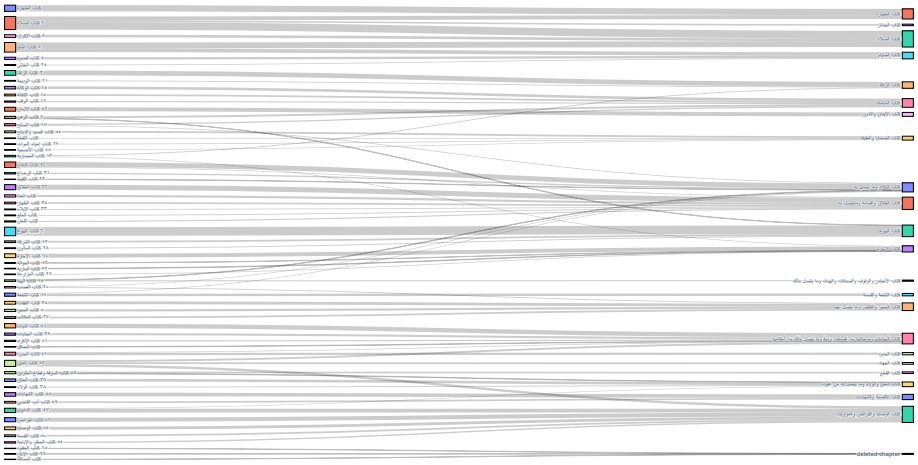

Fig. 3: Chapter Alignment: *Mukhtassar* of al-Quduri → *At-Talquin* of Qadi Abd al-Wahhab (AraBERT)

Figure 3 illustrates the alignment obtained with AraBERT. While main themes such as prayer and ablutions align correctly, this method proves less effective than TF-IDF, exhibiting significantly more errors. The mapping shows a higher frequency of crossing links, indicating less precision in thematic isolation. Furthermore, the model fails to properly identify missing sections, resulting in fewer chapters being correctly flagged as `deleted`. Ultimately, although the core pillars remain recognizable, the increased noise and misaligned granular chapters

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

suggest that AraBERT struggles to reconcile the structural and length disparities between these two legal works.

6 Conclusion

Overall, the experiments indicate that the TF-IDF representation outperforms AraBERT for chapter alignment in these legal texts. TF-IDF provides more consistent results due to the standardized and precise legal terminology, whereas AraBERT may capture indirect semantic links that can lead to misalignments, particularly for smaller chapters or when the language is highly specific. Regarding the alignment methods, the Levenshtein approach is limited to one-to-one correspondences and cannot handle the splitting or merging of chapters. On the other hand, optimal Transport (OT), while flexible in aligning chapters based on semantic similarity and order, does not naturally account for insertions and deletions. Fine-tuning contextual embeddings like AraBERT on historical legal corpora may also improve alignment performance by better capturing the specialized terminology and structure.

References

[1] W. Antoun, F. Baly, and H. Hajj. Arabert: Transformer-based model for arabic language understanding, 2021.

[2] L. A. Caffarelli and R. J. McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Ann. Math. (2)*, 171(2):673–730, 2010.

[3] L. Chapel, M. Z. Alaya, and G. Gasso. Partial optimal transport with applications on positive-unlabeled learning. 12 2020.

[4] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh, H. Bouamor, K. Bouzoubaa, V. Cavalli-Sforza, S. R. El-Beltagy, W. El-Hajj, M. Jarrar, and H. Mubarak. A panoramic survey of natural language processing in the arab world. *Commun. ACM*, 64(4):72–81, Mar. 2021.

[5] L. V. Kantorovich. On translocation of masses. *Doklady Akademii Nauk SSSR (USSR Academy of Sciences Reports)*, 37(7–8):227–229, 1942.

[6] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10:707–710, 1965.

[7] G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, pages 666–704, 1781.

[8] K. SPARCK JONES. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28(1):11–21, 01 1972.