Published paper available at: https://doi.org/10.1016/j.ins.2023.118956

Multi-view overlapping clustering for the identification of the subject matter of legal judgments

Graziella De Martino[a], Gianvito Pio[a,b], Michelangelo Ceci[a,b,c]

[a]Dept. of Computer Science, University of Bari Aldo Moro, Bari (Italy)
[b]Big Data Lab. - National Interuniversity Consortium for Informatics, Rome (Italy)
[c]Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana (Slovenia)

Abstract

The legal field is generally burdened by paper-heavy activities, and the management of massive amounts of legal judgments without the adoption of computational tools may compromise the effectiveness and efficiency of administration processes. In this paper, we propose MOSTA, a novel unsupervised method to support the automated identification of groups of legal judgments with similar characteristics, with the goal of reducing the manual effort necessary for the management of legal judgments.

Methodologically, MOSTA learns two different embedding models for legal judgments. The first aims to represent the semantics of the textual content, while the second aims to represent co-citations of legal acts, also considering the granularity of the citations. Such representations are then fused through a multi-view approach based on an autoencoder, and the obtained representation is finally exploited by a novel overlapping clustering algorithm. The latter is an additional strong point of MOSTA, since, contrary to existing approaches, does not rely on additional input parameters that inherently influence the degree of overlap of the resulting clusters.

Our experiments, performed on three textual datasets, including a real-world legal dataset provided by EUR-Lex, proved that the proposed representation of cited legal acts, the adopted multi-view fusion strategy, and the novel overlapping clustering algorithm implemented in MOSTA provide a positive contribution to the quality of the identified clusters. Finally, MOSTA demonstrated to be able to outperform by a great margin existing complete solutions based on fine-tuned BERT embedding models and existing overlapping clustering algorithms.

Keywords: Multi-view overlapping clustering, Embedding, Legal judgments

1. Introduction

The *law* can be considered an ensemble of governance rules aiming to guarantee that the rights of the members of a community are not abused by other members, corporations, or authorities. These rules inherently define a framework for good governance that binds the society to ensure the safety and the justice of day-to-day activities. However, the legal sector is burdened by paper-heavy activities, and the manual management of massive amounts of legal documents may compromise the effectiveness and efficiency of administration processes. In this context, computational approaches, possibly based on Artificial Intelligence (AI) techniques, can support the transformation of slow, paper-based, processes into smart and efficient workflows, through the automated integration and analysis of massive amounts of data.

In the literature, we can find several works that proposed the application of AI techniques to solve different tasks in the legal field. For example, in [3], the application of solutions based on information technology in the legal field was deeply investigated. The authors first discussed how the law appears as a body of rules that can be represented and understood through automated reasoning, emphasizing the challenges raised by the presence of ambiguities and *open texture*. The authors also suggested the adoption of ontologies to represent crucial legal relationships and to support machine learning algorithms. Finally, the authors proposed LUIMA, an architecture based on the UIMA framework that proved to be able to perform the conceptual markup of legal documents considering the semantics. In [28] the authors proposed a tool that notifies lawyers and consumers about potentially unfair clauses listed in terms of service of online platforms. Mandal et al. [31] proposed a measure to assess the similarity between textual legal court documents to improve the accuracy and the scalability of legal document retrieval systems. Another relevant example is the work presented in [32], where the authors designed an automated data collection framework that detects eviction judgments issued by Dutch courts. The authors performed two experiments, where the emphasis was on locating eviction-related judgments and the resolution of the cases in the judgments, respectively.

Following this line of research, in this paper, we propose MOSTA (Multi-view Overlapping cluSTering of legAl judgments), a novel AI method that can identify groups of legal judgments with similar traits, possibly corresponding to *subject matter(s)*[1], thus reducing the necessary human effort for navigating, organizing, and classifying large quantities of legal judgments. Note that even if subject matters could be considered as labels/categories in supervised machine learning tasks, in real-world scenarios, labeled legal judgments are scarcely available. This is the main motivation for which we designed MOSTA as a novel unsupervised clustering approach. Specifically, MOSTA falls in the category of *overlapping* clustering methods, that is, it is able to assign each document to more than one cluster. The adoption of an overlapping clustering approach in this scenario is motivated by the fact that legal documents tend to be related to multiple subject matters [12, 30, 40], and restricting legal judgments to belonging to a single cluster would lead to disregard relevant secondary topics.

Another peculiarity of legal documents is that their complex semantics is not entirely described by their textual content, but also by cited legal acts, such as regulations, directives, decisions, recommendations, and opinions. Moreover, legal citations can also guarantee that the judgment conclusion is not based solely on the magistrate's choice, but takes into account the information conveyed by entrenched precedents. This aspect is particularly important in legal systems based on Common law, which apply the *stare decisis* principle. In such systems, the similarity of the scenario with respect to that of precedents is exploited to push the decision towards a similar outcome [27]. Note that although precedents play a less decisive role in Civil law systems, they are frequently used to back, support and defend specific outcomes, but also to show how a similar legal problem was previously dealt with. Therefore, even if the textual content can properly be represented by resorting to existing embedding techniques (e.g., BERT [15]), possibly focused on the legal domain (e.g., LEGAL-BERT [10]), ignoring the information conveyed by cited legal acts would lead to disregard relevant aspects for the identification of the subject matters.

Although existing general-purpose overlapping clustering approaches can overcome the limitation of a single cluster assignment, they usually require additional input parameters, mainly to define the desired degree of overlap [47]. Moreover, existing methods for document clustering are not able to specifically take advantage of the complimentary information represented by

[1] A *subject matter* denotes the substance of the arguments, reasoning and informal fallacies presented for consideration during a judgment hearing.

the cited legal acts, together with the textual content. Consequently, they cannot accurately grasp the similarity between legal judgments.

In this context, the method MOSTA proposed in this paper solves all the above-mentioned limitations. In particular, MOSTA is based on a multi-view approach that fuses content-based embeddings with citation-based embeddings by means of a stacked autoencoder [5]. For the former, we adopt a word embedding method able to consider the semantics of the textual content, as well as the contextual information. For the latter, we represent the granularity of each citation (e.g., a whole act, an article, a sub-article, etc.) through a tree-based structure, and exploit an embedding strategy based on the similarity among trees. Note that MOSTA can work with any kind of citations, both linking to precedents (typical of Common Law systems) and linking to regulations, directives, and decisions (typical of Civil Law systems). Finally, MOSTA exploits a novel overlapping clustering method that does not require additional input parameters, and is able to automatically estimate the proper degree of overlap from data.

The rest of the paper is structured as follows. In Section 2 we describe existing works related to the present paper, while in Section 3 we describe in detail the proposed method MOSTA. In Section 4 we describe our experimental evaluation, showing and discussing the obtained results. Finally, in Section 5 we draw some conclusions and outline possible future work.

2. Related Work

In the following subsections, we briefly discuss existing approaches related to the present paper. Specifically, we discuss existing clustering methods applied in the legal field, and works that proposed multi-view document clustering approaches, even if not specifically tailored for the legal field.

2.1. Clustering of legal documents

Most of the activities in the legal field are based on the management and analysis of large amounts of textual documents. During the last years, the increased availability of legal databases outlined new opportunities for automated data-driven approaches. In particular, in the literature we can find several methods for cluster analysis, whose primary objective is the reduction of the complexity of repetitive tasks, by facilitating the navigation and the organization of large collections of legal documents. A relevant example is [11], where the authors applied clustering techniques to automatically group case law petitions submitted to electronic trial systems. The authors adapted the hard clustering algorithm initially proposed in [1], and introduced the paradigm of *bag of terms and law references*. This paradigm is based on a domain thesaurus to identify legal terms, and on regular expressions (RE) to extract law references. Although this approach, similarly to MOSTA, somehow considers citations, it treats them as textual words in the bag, without properly considering their granularity. Moreover, the adopted clustering method does not allow each document to fall into multiple clusters.

Lu et al. [30] proposed an overlapping clustering algorithm based on a built-in topic segmentation approach that leverages legal metadata about several types of legal documents. In addition to showing the scalability of the proposed solution, the authors emphasized the ability to move from traditional lexical approaches toward the exploitation of topics, citations, and click-stream data from behavior databases. However, the textual content is represented through the classical bag-of-words model, with TF-IDF weighing, and the similarity among documents in terms of citations is based on the Jaccard measure, without taking into account their granularity.

Conrad et al. [12] performed a comparative study between hard and overlapping clustering solutions on three different legal datasets, using the CLUTO clustering toolkit [50]. The results showed the effectiveness of overlapping and hierarchical clustering, in terms of both internal and external quality measures, as well as in terms of the usefulness of the extracted clusters for human legal experts. Similarly, Sabo et al. [40] explored the application of approaches based on hard clustering (K-means and Affinity Propagation), overlapping hierarchical clustering, and soft clustering (Lingo) to sparse numerical vectors (obtained using the Bag of Words model) related to cases dealing with airline service failure claims. The results showed the superiority of hierarchical clustering in terms of entropy, purity, and legal experts' feedback. It is noteworthy that, in this case, possible overlaps among clusters can occur only at different hierarchical levels, i.e., clusters can overlap simply because of inclusive parent relationships. On the other hand, the considered soft clustering solution requires a user-defined threshold to decide whether a legal judgment belongs to a given cluster or not.

Existing general-purpose overlapping clustering approaches (e.g., [20]), even if not specifically tailored for the legal field, can provide alternative solutions if applied to a proper representation of legal documents. However, analogously to soft clustering approaches, they require additional input parameters, that explicitly or implicitly influence the final degree of overlap.

An exception is represented by [47], which also proposes some strategies to estimate the value of such additional parameters from data. For this reason, in Section 4, we will consider it as a competitor with respect to the novel clustering method implemented in MOSTA.

2.2. Multi-view document clustering

The need to take into account multiple perspectives/views of a document could straightforwardly be satisfied by concatenating the features associated with each different view. However, approaches based on feature concatenation usually cannot differentiate the contribution provided by each view, and could easily over-estimate the weight of a given view simply because it is represented by a high number of features. Therefore, in the literature, several multi-view document clustering approaches have been proposed, that aim to overcome the limitations of methods based on simple feature concatenation.

A relevant example is the work by Gao et al. [18], that extends the information bottleneck algorithm to cluster web documents represented by multiple distinct feature sets. Their experiments on two real datasets demonstrated the effectiveness of the proposed approach, specifically when the views represent the textual content, anchor texts, and URLs.

Other approaches are based on ensemble strategies. In particular, Kim et al. [24] adopted an incremental algorithm to cluster multi-lingual documents, where each view provides a representation of documents in a different language. In the first stage, the authors apply the Probabilistic Latent Semantic Analysis (PLSA) [21] independently on each view, constraining each clustering model to identify the same number of groups (topics). Then, they identify the final clustering model such that documents falling in the same group share similar patterns in terms of the probabilities returned by PLSA. Wahid et al. [46] exploited a multi-objective optimization technique based on the Non-Dominated Sorting Genetic Algorithm-II (NSGA-II) [14], aiming to identify a clustering solution, among those returned by multiple clustering methods applied to all the available views, that simultaneously minimizes the number of obtained clusters, the number of words that are not in common among documents in the same cluster, and the inter-cluster similarity. Hussain et al. [23] aggregated (by average) a cluster-based similarity matrix, a pairwise similarity matrix, and an affinity matrix, computed through different approaches on the different views. A further clustering step is then applied on the combined similarity matrix to obtain the final result. Finally, Zamora and Sublime [48] combined clustering results obtained from different views using an information theory model based on Kolmogorov complexity.

It is noteworthy that ensemble-based approaches (that work on the output spaces) may suffer from similar issues with respect to approaches based on feature concatenation (that work on the input spaces). Indeed, while in the latter case each feature has the same importance, leading to possible biases towards high-dimensional views, in ensemble-based approaches, each view has the same importance, independently of the actual contribution it provides. On the contrary, the approach implemented in MOSTA combines the contribution provided by the features describing each view, without introducing specific biases (see [2] for an overview on the effect of different kinds of biases on the learned models).

Some other attempts to overcome this issue have been made in more recent works. For example, Zhan et al. [49] proposed the multi-view graph-regularized concept factorization (MVCF) method, based on concept factorization. In addition to exploiting multi-view features, similarly to the system SAIRUS [36], MVCF achieves superior clustering performances with respect to previously-proposed methods by reducing the dimensionality of data and by learning different weights for each view. Similarly, Bai et al. [4] designed a deep neural network that learns a semantic mapping from a high-dimensional to a low-dimensional feature space. In particular, the authors exploited a neighbor-based autoencoder model and a cross-view autoencoder model to involve neighbor-wise (within the same view) and view-wise complementary information in the clustering process.

Although the above-mentioned methods can be considered as multi-view clustering approaches, since they properly weigh the contribution provided by different views, they are neither able to identify overlapping clusters nor to properly capture the different granularities of legal citations we can find in legal documents. In this respect, to the best of our knowledge, MOSTA can be considered the first method that adopts a multi-view learning approach able to properly model both the textual content and citations of legal acts, also considering their granularity, and that exploits a novel overlapping clustering approach to identify their subject matters, without the need of specifying additional parameters that influence the degree of overlap.

3. The proposed method MOSTA

Before describing the steps performed by our method MOSTA, we briefly introduce some useful notation (see Appendix A for a compact view of all the used symbols) and formally define the solved task. Let:

- J be a set of legal judgments, that also cite legal acts;
- k be the desired number of clusters, possibly representing legal subject matters.

The task solved by MOSTA consists in the identification of k, possibly overlapping, clusters of the legal judgments J, taking into account *i)* the semantics of their textual content and *ii)* the legal acts they cite, at different levels of granularity (e.g., a whole act, an article, a sub-article, etc.). As per the definition of overlapping clustering, each legal judgment $J_i \in J$ can possibly be assigned to multiple clusters, representing the fact that it may be related to multiple subject matters.

Our method consists of four main phases, namely:

1. **Embedding of the textual content of legal judgments**, that consists in *i)* learning an embedding model from J, capable to represent the *semantics of the textual content* of the judgments into a numerical feature space, and *ii)* adopting the learned model to represent each judgment $J_i \in J$ in the learned feature space.
2. **Embedding of the citations of legal judgments**, that consists in *i)* learning an embedding model from J, capable of representing the *co-citation network* (also considering the granularity of the citations) of legal judgments towards legal acts into a numerical feature space, and *ii)* adopting the learned model to represent each judgment $J_i \in J$ in the learned feature space.
3. **Multi-view embeddings fusion**, that is the construction of a fused, multi-view representation for each judgment $J_i \in J$ through a stacked autoencoder that exploits both the content-based and the citation-based embeddings identified in phases 1 and 2.
4. **Identification of overlapping clusters of legal judgments**, that consists in the adoption of a novel overlapping clustering approach, that discovers k homogeneous groups of legal judgments according to their fused embeddings, without requiring additional input parameters to determine the degree of overlap.

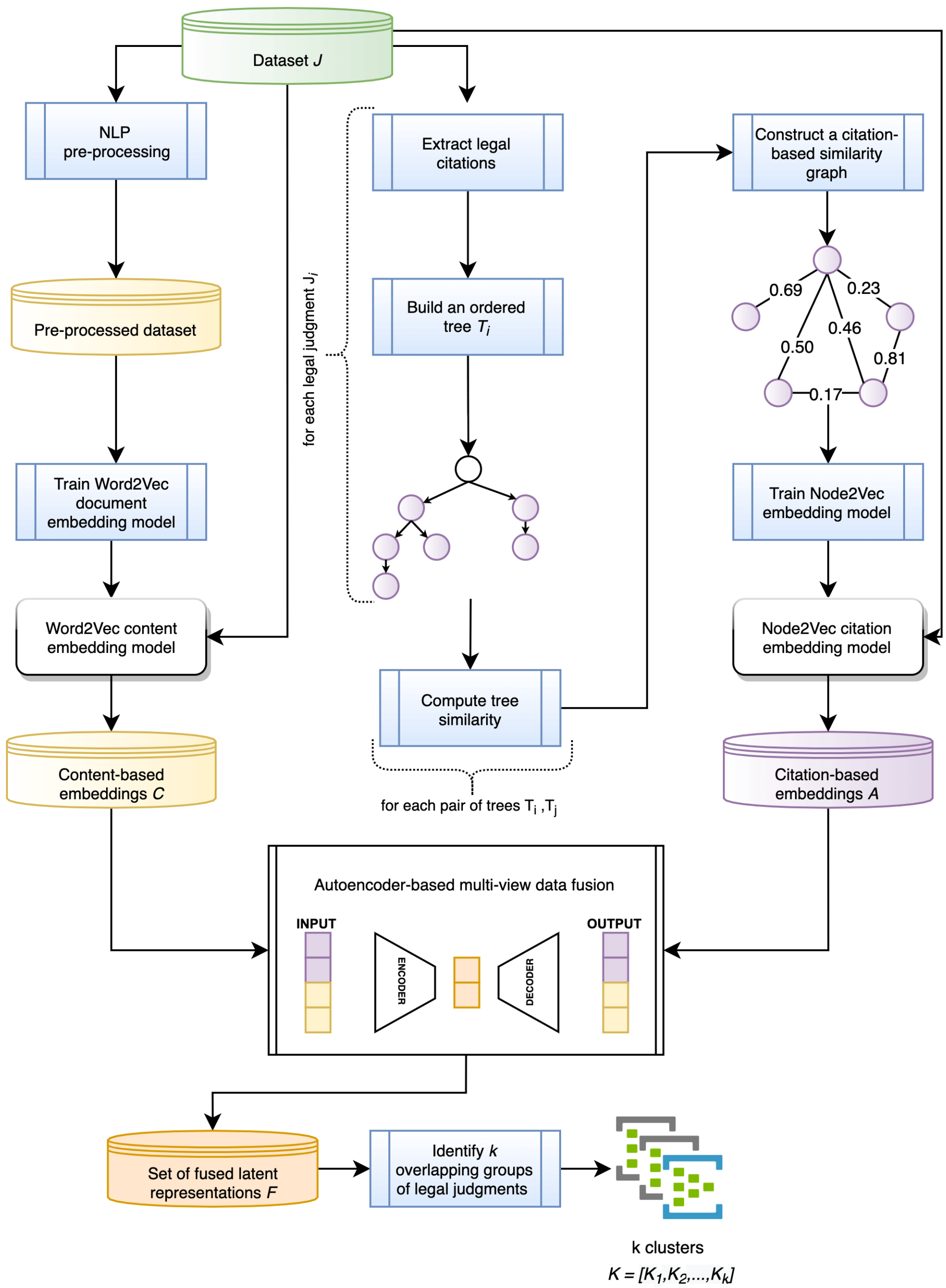


Figure 1: General workflow of the method MOSTA.

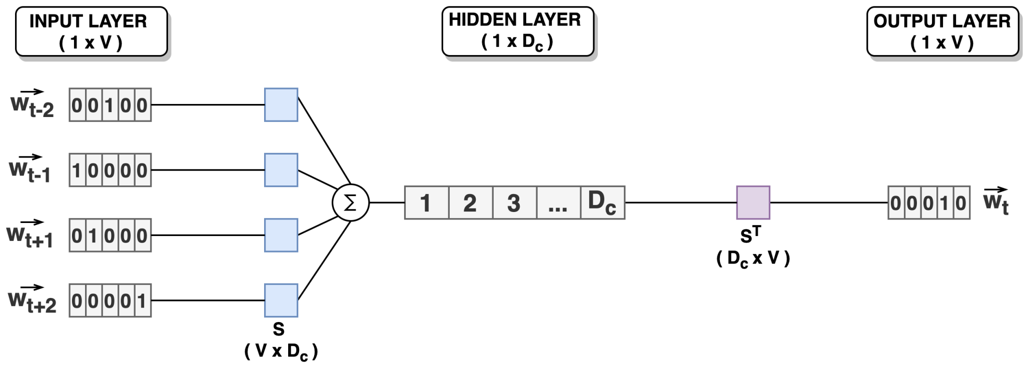


Figure 2: Graphical representation of the Word2Vec CBOW architecture. Note that there is only one matrix S, that is repeated multiple times only for explanatory purposes.

In the remainder of this section, we describe the approach followed by MOSTA to perform each phase, which is also globally depicted in Fig. 1.

3.1. Embedding of the textual content of legal judgments

In this section, we describe the steps followed by MOSTA for the representation of the textual content of legal judgments in a numerical feature space. Initially, MOSTA adopts standard Natural Language Processing (NLP) [8] pre-processing techniques, namely, lowercasing, punctuation and digits removal, lemmatization, and removal of stopwords and rare words. Subsequently, the pre-processed legal judgments are used to train an embedding model. In particular, MOSTA adopts the neural network (NN) architecture implemented in Word2Vec [33], given its proven superiority over traditional counting-based and other document-based embedding approaches, even in presence of noise in the data [31, 13, 26]. Word2Vec relies on two different shallow NN architectures, namely the Continuous-Bag-of-Words (CBOW) architecture and the Skip-gram (SG) architecture. Although both architectures are able to capture complex syntactic and semantic relationships among words, they adopt distinct learning processes. Specifically, CBOW aims to predict a target word from a surrounding context, while SG aims to predict the surrounding words of a given target word. The CBOW architecture is able to represent rare words more accurately, although it usually requires a slightly higher execution time than SG [37, 43]. Therefore, in MOSTA, we adopt the CBOW architecture, whose description is reported as follows.

Given a sequence of words $\langle w_{t-h}, ..., w_t, ..., w_{t+h} \rangle$ describing a target word

w_t and its context of size $2h$, the CBOW architecture takes as input the one-hot vector representation $\vec{w_i}$ of size V for each context word w_i, where V is the size of the vocabulary observed in the set of legal judgments J. The learning phase aims to identify the optimal values for the matrix $S \in \mathbb{R}^{V \times D_C}$, where D_C represents the desired embedding dimensionality. The one-hot vector representation of each w_i is multiplied by S to obtain $2h$ vectors in $\mathbb{R}^{D_C}$. The hidden layer represents the embedding of the target word w_t obtained by aggregating the $2h$ vectors associated with the context words as follows:

$$\sum_{w_i \in \{w_{t-h}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+h}\}} \vec{w_i} \cdot S \tag{1}$$

The output layer is obtained by multiplying the embedding of the target word w_t by $S^\top$, and corresponds to the one-hot vector representation $\vec{w_t}$ of the target word w_t. This means that the values of the matrix S are optimized so that the one-hot vector representation of the target word w_t is accurately reconstructed, given the one-hot vectors of the context words as inputs. The learned matrix S can therefore be used to embed any word into a numerical feature space of size D_C, given its context words.

Word2Vec naturally provides an embedding for each word. Therefore, in order to identify an embedding for the document in J, as suggested in [33], we adopt a mean aggregation strategy. The output of this phase is the set of embedded documents C, according to their textual content.

3.2. Embedding of the citations of legal judgments

During the redaction of legal documents, legal experts usually cite pertinent legal acts, such as statutes, regulations, decisions, or directives [41]. A legal citation provides a direct link to a recognized source that *i)* references a legal act and/or a legal act section through which some conclusions are inferred; *ii)* supports the impartiality of the judgment, providing possible links to similar contexts and precedents.

Given the importance of legal act citations, in MOSTA we define an approach that extracts a set of citation-based embeddings A from the legal judgments J. The goal is to identify a complimentary representation, with respect to that based on the textual content, that takes into account co-cited legal acts, as well as the granularity of the citations.

In detail, for each legal judgment $J_i \in J$, MOSTA represents cited legal acts as an ordered tree T_i (see Fig. 3). Note that cited legal acts may already

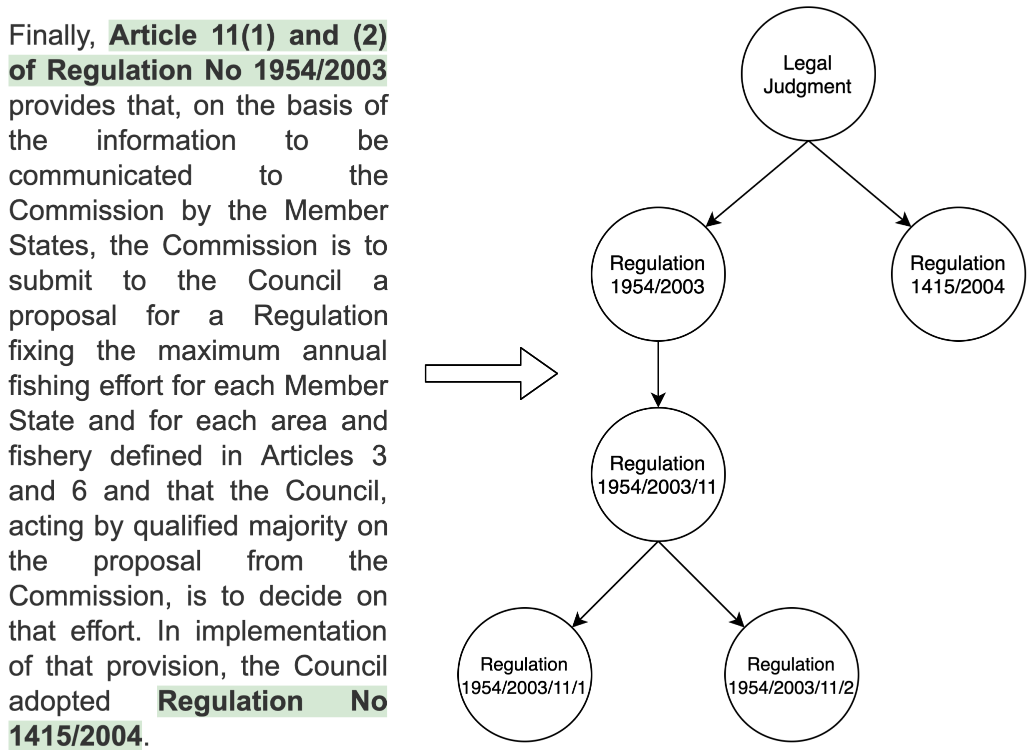


Figure 3: Representation of cited legal acts through an ordered tree.

be available as structured data in the dataset, or may need to be extracted, e.g., using Regular Expressions (RE). Since each legal system has its distinctive characteristics and there is no uniformity across all jurisdictions, in Section 4.1, we define in detail the specific techniques used to extract legal citations from the dataset used in our experiments.

Once an ordered tree has been constructed for each $J_i \in J$, MOSTA computes the pairwise similarity between judgments. More formally, given two ordered trees T_i and T_j, extracted from the judgments $J_i \in J$ and $J_j \in J$, respectively, the tree similarity $s(T_i, T_j)$ is computed as:

$$s(T_i, T_j) = 1 - \frac{\delta(T_i, T_j)}{|T_i| + |T_j| - 2}, \qquad (2)$$

where:

- $\delta(T_i, T_j)$ is the tree edit distance [35] defined as the minimum-cost sequence of node edit operations, i.e., deletion, insertion, and relabeling of nodes[2], needed to transform T_i into T_j[3];

- the factor $(|T_i| + |T_j| - 2)$, where $|\cdot|$ denotes the number of nodes of a tree, corresponds to the maximum number of edit operations needed to transform T_i into T_j, assuming that they are totally different trees.

To compute $\delta(T_i, T_j)$, MOSTA adopts the memory-efficient algorithm APTED [35]. In Fig. 4, we report a step-by-step example of the computation of the tree edit distance, while in Fig. 5 we report multiple examples of the similarity computed between different pairs of trees.

After computing the similarity between documents in terms of their citations, MOSTA builds a weighted graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where the set of nodes $\mathcal{N}$ corresponds to the judgments J, and each edge $\langle J_i, J_j \rangle \in \mathcal{E}$ represents the fact that J_i and J_j co-cited some legal acts. Moreover, each edge $\langle J_i, J_j \rangle \in \mathcal{E}$ is associated with a weight corresponding to the similarity of their citations, namely to $s(T_i, T_j)$, computed through Eq. (2).

Starting from such a weighted graph, we learn a numerical representation for each node of the graph (i.e., for each judgment), where the new numerical feature space aims to preserve the closeness relationships in the graph, also according to the defined edge weights. In this way, the learned representation for a given judgment encodes the information about the fact that other judgments co-cite the same legal acts, taking into account the granularity of such co-citations thanks to the similarity measure defined in Eq. (2).

For the learning phase of the numerical representation from such a graph, MOSTA exploits PecanPy [29], a memory-efficient implementation of the method Node2Vec [19]. Node2Vec is a neural network architecture that learns continuous feature representations for each node in a graph, by sampling some representative nodes (in its neighborhood) following r 2^{nd}-order random walks of fixed length l, biased by a hybrid Depth-First (DFS) / Breadth-First (BFS) search approach. In particular, assuming that a given random walk traverses the edge $\langle J_i, J_j \rangle$, the transition probability from J_j to the node representing another judgment J_k, via the edge $\langle J_j, J_k \rangle$, is computed as

$$s(T_j, T_k) \cdot \beta(J_i, J_k) \tag{3}$$

[2]The cost of the relabeling operation is considered the double of the cost required for insertion or deletion operations, since it corresponds to a deletion of a node and to an insertion of a new node with a different label.

[3]Note that the considered node distance measure is symmetric. Therefore, the cost of transforming T_i into T_j is the same as that required to transform T_j into T_i.

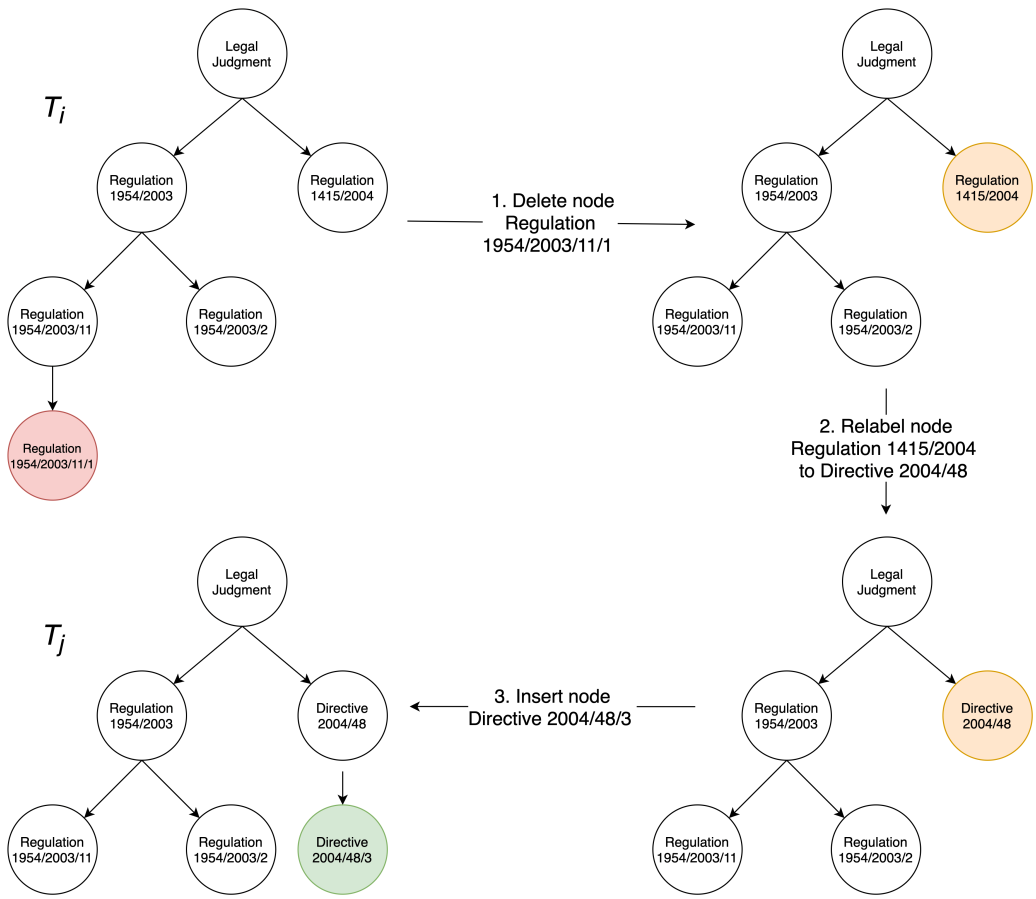
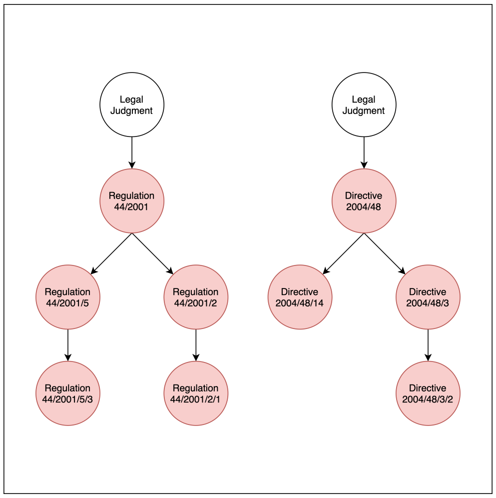


Figure 4: Graphical representation of the minimum-cost sequence of node edit operations needed to transform T_i into T_j. In the example, the distance between T_i and T_j is 4, which derives from the cost (1) for a node deletion operation (red) + the cost (2) for a node relabeling operation (orange) + the cost (1) for a node insertion operation (green).

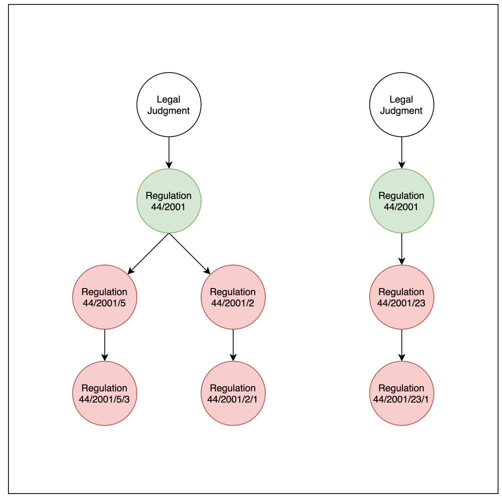
where:

$$\beta(J_i, J_k) = \begin{cases} \dfrac{1}{p} \text{ if } g(J_i, J_k) = 0 \qquad (i.e., J_i = J_k) \\ 1 \text{ if } g(J_i, J_k) = 1 \\ \dfrac{1}{q} \text{ if } g(J_i, J_k) = 2 \end{cases} \tag{4}$$

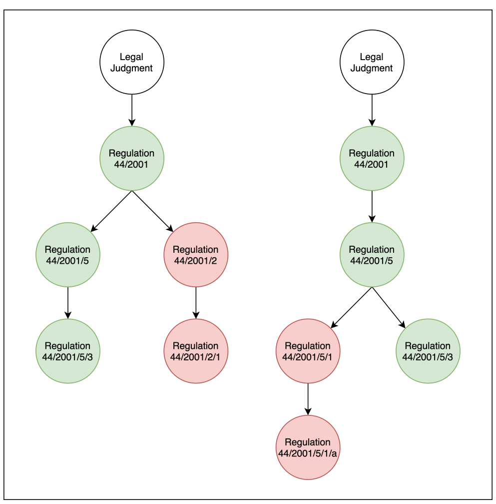
In Eq. (4), $g(J_i, J_k)$ is the distance (in terms of steps in the graph) between the nodes representing the judgments J_i and J_k; p is a parameter that controls



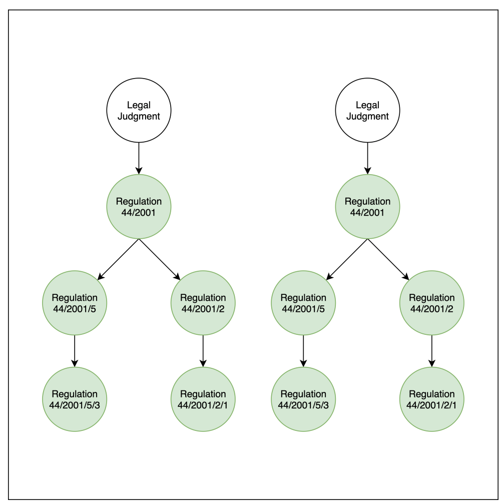
(a) $s(T_i, T_j) = 0.00$



(b) $s(T_i, T_j) = 0.25$



(c) $s(T_i, T_j) = 0.60$



(d) $s(T_i, T_j) = 1.00$

Figure 5: Examples of tree similarity scores computed between two ordered trees T_i, T_j. Green nodes represent matched citations, while red nodes represent differences.

the likelihood of immediately revisiting a node; q is a parameter that controls how far the random walk should progress from J_i.

Subsequently, sampled random walks starting from each judgment are considered as sequences of words representing its context, and are used to learn a Word2Vec model. The embedding layer of this model is finally used to extract the citation-based embeddings A for all the judgments J.

3.3. Multi-view embeddings fusion

The exploitation of multiple perspectives/views for the same units of analysis has attracted increasing attention in the research community since, when available, they can offer complimentary representations that may boost the performance of the learned models. However, simple approaches, such as feature concatenation, may introduce additional issues, namely feature redundancy and collinearity [17], if the considered views are not completely independent/orthogonal, and the curse of dimensionality, if the final number of features is significantly higher than the available observations. As shown in Section 2.2, more advanced approaches can be adopted, to properly capture the contribution coming from the available views. In MOSTA, we adopt an Autoencoder (AE) [5] to learn a low-dimensional fused representation from the D_C-dimensional content-based embeddings C and from the D_A-dimensional citation-based embeddings A.

An AE is an unsupervised feedforward neural network that learns a *compressed* representation, such that the original data can be accurately reconstructed. It comprises an *encoding* part, that maps the original input data into the compressed space, and a *decoding* part, that reconstructs the original data from its compressed version.

Methodologically, MOSTA initially concatenates content-based and citation-based embeddings, leading to a feature vector in $\mathbb{R}^{D_C+D_A}$ for each judgment. The input layer of the AE takes such a concatenated representation, which is compressed into a D_F-dimensional feature space, where $D_F < D_C + D_A$. The specific architecture of the adopted AE is depicted in Fig. 6. Note that, in general, multiple hidden layers can be defined in the AE architecture before reaching the bottleneck layer that represents the final embeddings. The choice of the number of additional hidden layers, as well as of the number of their neurons, usually depends on the difference between the input dimensionality ($D_C + D_A$, in our case) and the desired embedding dimensionality (D_F, in our case).

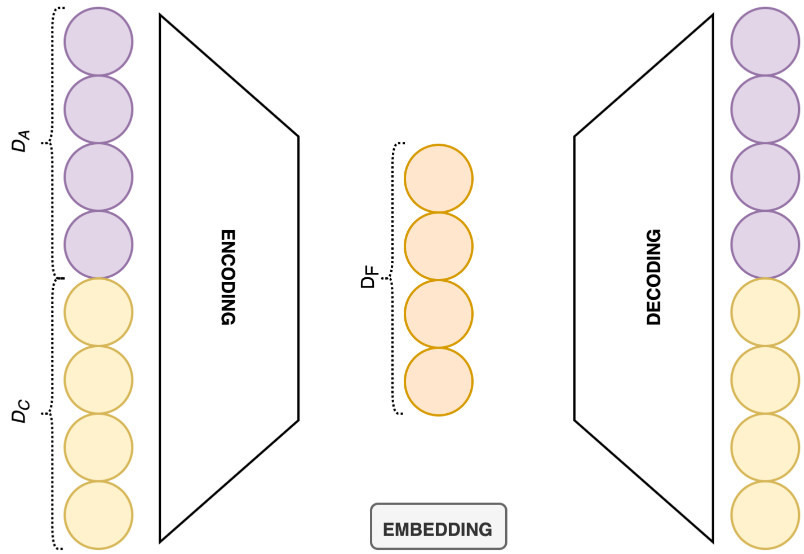


Figure 6: Graphical representation of the architecture of the adopted AE.

Note that the goal of the learning phase of the autoencoder is to optimize the weights of the neurons, such that the reconstruction errors, i.e., the *loss* between the input and the output layer, is minimized. The loss is usually based on common measures like Root Mean Square Error (RMSE). In MOSTA, we adopt a different customized measure, that is able to provide different importance to the different sets of features belonging to each view. Specifically, we adopt a weighted variant of the RMSE, defined as follows:

$$\theta = \sqrt{\frac{1}{|J|} \sum_{J_i \in J} \gamma \cdot (\hat{x}_i - x_i)^2} \tag{5}$$

- x_i is the input $(D_C{+}D_A)$-dimensional feature vector representing the judgment J_i;

- $\hat{x}_i$ is the $(D_C{+}D_A)$-dimensional feature vector representing the judgment J_i, returned by the output layer of the AE;

- $\gamma = \lambda^{1 \times D_C} \oplus (1-\lambda)^{1 \times D_A}$ defines the weights for the features coming from each view. If $\lambda = \frac{D_C}{D_C+D_A}$, θ corresponds to the standard RMSE.

Note that γ influences the importance given to each view in the computation of the loss function θ. Therefore, even if $\lambda = 0$ (resp., $\lambda = 1$) does not formally mean that the AE discards the features of the content-based (resp., citation-based) embeddings, it implies that they are ignored in the computation of the loss function. In particular, when the loss function is required to ignore the features related to a specific view, such features would actually provide a negligible contribution to the obtained fused embeddings, meaning that we can consider the configuration $\lambda = 0$ (resp., $\lambda = 1$) equivalent to the scenario in which only the citation-based embeddings (resp., the content-based embeddings) are considered.

The result of this phase is the set of embedded judgments F, represented in a new D_F-dimensional feature space that fuses the contribution of the initial embeddings learned in the previous phases. The embedded judgments will be the input of the final clustering phase, that is described in the following subsection.

3.4. Identification of overlapping clusters of legal judgments

This subsection describes the novel clustering method that we implemented in MOSTA to identify k overlapping groups of legal judgments from F. A common approach adopted by existing overlapping clustering methods, such as Neo K-Means [47], consists in the application of hard clustering solutions and in the assignment of additional clusters to each object, according to some criteria. However, as mentioned in Section 2.1, such criteria are usually based on a user-defined parameter that defines the degree of overlap, or the number of additional cluster assignments to perform. In MOSTA, we overcome this issue by adopting an approach based on outlier detection. Specifically, after applying a hard clustering method (i.e., k-means), MOSTA computes the Euclidean distance between each judgment and the centroid of each identified cluster. Assuming a Normal distribution of the distances, MOSTA identifies the judgment-cluster pairs, not already identified by the initial run of k-means, whose distance can be considered as an outlier. Specifically, following the 3-σ rule, MOSTA assigns a judgment to a given cluster if their distance is less than $d_{max} = \bar{d} - 3\sigma$ (see Fig. 7), where $\bar{d}$ and σ are the average distance and the standard deviation of distances, respectively, between a judgment and a cluster centroid identified by k-means.

In Alg. 1, we report a pseudocode description of the clustering algorithm implemented in MOSTA. The algorithm starts by adopting the k-means clustering algorithm to partition F into k non-overlapping clusters (Alg. 1, line 2). Then, the Euclidean distance between each judgment and each centroid of each cluster is computed (Alg. 1, line 3), in order to compute the mean $\bar{d}$ and

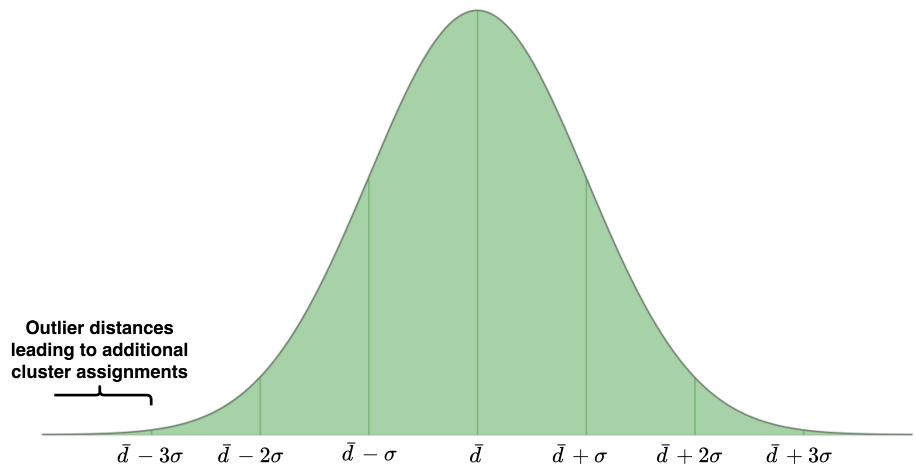


Figure 7: Representation of the outlier distances leading to additional cluster assignments.

the standard deviation σ (Alg. 1, line 4), and the threshold d_{max} to consider a distance between a judgement and a cluster as an outlier (Alg. 1, line 5). Finally, MOSTA performs additional cluster assignments when the observed judgment-cluster distance is less than the threshold d_{max} (Alg. 1, lines 6-9). We stress the fact that this strategy allows MOSTA to identify overlapping clusters by solely exploiting the observed distribution of distances, without imposing a pre-defined degree of overlap among clusters, or a pre-defined number of cluster assignments per judgment.

3.5. Time complexity analysis

In this subsection, we discuss the time complexity of the proposed method MOSTA, analyzing the time complexity of each phase, separately.

The first phase is the embedding of the textual content of legal documents, which corresponds to learning a Word2Vec model and to adopt it to embed all the documents. The time complexity of the training phase of Word2Vec is $O(|J| \cdot log(V))$ [33], where V is the size of the vocabulary observed in the set of legal documents J. Once the learning phase of the Word2Vec model is completed, the embedding of each document requires $O(len \cdot 2h \cdot D_C \cdot V)$ where len is the average number of words of a document, $2h$ is the size of the context and D_C is the dimensionality of the embedding. This complexity depends on a matrix multiplication between the input and the hidden layer of the Word2Vec architecture, performed for each word of the context of each word of the document. Considering that h and D_C are constant values, the complexity of embedding all the documents can be approximated to $O(|J| \cdot len \cdot V)$.

The second phase of MOSTA requires the computation of a pairwise tree-based similarity between the documents, based on the citations. The computation of the similarity between two trees T_i and T_j (see Eq. (2)) has a time complexity $O(max(|T_i|, |T_j|)^2)$ [35]. Considering that each tree is generally very small, compared to the number of documents, the complexity of the pairwise tree-based similarity between the documents can be approximated to $O(|J|^2)$. Moreover, the second phase requires learning an embedding model through Node2vec and its adoption to embed all the documents of the collection. Considering that Node2Vec is based on Word2Vec, where the vocabulary corresponds to the set of documents, its training complexity is $O(|J| \cdot log|J|)$, while the cost for embedding all the documents is $O(|J|^2)$ [33]. Therefore, the time complexity of the second phase can asymptotically be approximated to $O(|J|^2)$.

The third phase of MOSTA consists of learning a stacked autoencoder and using it to fuse the content-based and citation-based embeddings of each document. The time complexity of this step (see [6]) depends linearly on the number of training examples ($|J|$ in our case) and quadratically on the number of input features (in our case $(D_C + D_A)^2$). Therefore, considering that D_C and D_A are constant values and that $(D_C + D_A)^2 \ll |J|$, the time complexity of this phase can be approximated to $O(|J|)$.

Finally, we need to estimate the complexity of the clustering phase. The classical k-means algorithm linearly depends on both the number of examples and on the number of features. This means that, in our case, the time complexity of running k-means is $O(|J| \cdot D_F)$, where D_F is the fused embedding dimensionality. The identification of additional cluster assignments requires the computation of the distance between each document and each cluster centroid. Therefore, such a computation requires $O(|J| \cdot k \cdot D_F)$ operations. Finally, such distances are scanned once to identify those falling below the automatically generated threshold d_{max}. Since D_F and k are constants generally much smaller than $|J|$, we can conclude that the time complexity of this phase can be approximated to $O(|J|)$.

Summing up the time complexity of each phase performed by MOSTA, we obtain $O(|J| \cdot len \cdot V) + O(|J|^2) + O(|J|) + O(|J|)$. If $(len \cdot V) \leq |J|$, the overall complexity of MOSTA can be approximated to $O(|J|^2)$, whereas if $(len \cdot V) > |J|$, the overall complexity of MOSTA can be approximated to $O(|J| \cdot len \cdot V)$.

Therefore, we can conclude that the time complexity of MOSTA is either quadratic in the number of documents to be processed or linear in the number of legal documents, in the average size of each document, and in the size of the vocabulary considered. Obviously, the worst-case analysis requires us to consider the highest time complexity among the two cases.

Algorithm 1: MOSTA overlapping clustering approach

Data:

· F: set of fused vector representations of the legal judgments J

· k: desired number of clusters to identify

Result:

· K: set of k overlapping clusters of legal judgments

begin

    `/* Identify` k `non-overlapping clusters through` k`-means */`

    $K \leftarrow k\text{-means}(F, k)$;

    `/* Compute judgment-cluster pairwise distances */`

    $PD \leftarrow computePairwiseDistances(F, K)$;

    `/* Compute mean and standard deviation of the distances */`

    $\bar{d} \leftarrow \frac{1}{|PD|} \cdot \sum_{d \in PD} d; \qquad \sigma \leftarrow \sqrt{\frac{1}{|PD|} \sum_{d \in PD} (d - \bar{d})^2}$;

    `/* Compute the threshold to consider a distance value as an outlier */`

    $d_{max} \leftarrow \bar{d} - 3 \cdot \sigma$

    `/* Identify overlapping clusters: perform additional judgment-cluster assignments when their distance appears as an outlier */`

    foreach $F_i \in F$ **do**

        foreach $K_j \in K$ **do**

            if $distance(F_i, K_j) < d_{max}$ ***and*** $F_i \notin K_j$ **then**

                $K_j \leftarrow K_j \cup \{F_i\}$;

            end

        end

    end

    return K;

end

---

4. Experiments

We performed the experiments along three different dimensions of analysis. Specifically, we first evaluated the effectiveness of the proposed *overlapping clustering method* implemented in MOSTA on three textual datasets, in comparison with existing overlapping and soft clustering approaches.

Subsequently, we evaluated the effectiveness of the *multi-view fusion strategy* adopted by MOSTA, and its ability to also capture the information conveyed by citations. This evaluation was performed on the EUR-Lex dataset[4], whose documents fall in the legal domain and provide both textual content and citations.

Finally, on the same dataset, we compared the overall performance exhibited by MOSTA with those achievable by complete competitor solutions based on fine-tuned BERT embedding models and on the best overlapping clustering method identified in the first phase of our experiments.

In the following subsections, we first detail the adopted datasets, the competitor systems, the experimental setting and the evaluation measure. Then, we show and discuss the obtained results for all the experiments.

4.1. The considered datasets

EUR-Lex. The first dataset that we considered in our experiments was provided by EUR-Lex[4]. This dataset contains 4176 non-empty official public EU legal judgments that were finalized between 2008 and 2018, categorized in one or more *subject matters*[5], that fall within the case-law sector and the Court of Justice. In the dataset, we can find 133 distinct subject matters.

In order to build the set of citation-based embeddings A, we adopted a custom strategy to extract citations from the dataset, since they were not available as structured data. In particular, we reached EUR-Lex to identify common rules adopted for citations in the legal judgments of this specific dataset. Following their indications, we pre-processed the set of judgments J by: *i)* lowercasing the text, *ii)* removing punctuation except for the forward slash and the parenthesis (commonly used in citations), and *iii)* removing stop words except for the word *of* (commonly used in citations). Subsequently, we designed custom regular expressions (see Appendix B) to

[4]https://eur-lex.europa.eu/homepage.html

[5]For evaluation purposes, we discarded legal judgments not associated with any *subject matter* in the original dataset.

ID	Act Name	Article Number	Sub-Level 1	Sub-Level 2	Sub-Level 3	Sub-Level 4
62015CJ0005	Dir. 87/344	Dir. 87/344/4	Dir. 87/344/4/1	-	-	-
62015CJ0005	Dir. 87/344	-	-	-	-	-
62015CJ0005	Dir. 87/344	Dir. 87/344/4	Dir. 87/344/4/1	Dir. 87/344/4/1/a	-	-
62015CJ0005	Dir. 87/344	Dir. 87/344/3	Dir. 87/344/3/2	Dir. 87/344/3/2/c	-	-
62015CJ0005	Dir. 87/344	Dir. 87/344/3	Dir. 87/344/3/2	Dir. 87/344/3/2/a	-	-
...	...	...	...	...	...	...
62007CJ0416	Dir. 91/628	Dir. 91/628/5	Dir. 91/628-5/a	Dir. 91/628/5/a/1	Dir. 91/628/5/a/1/a	-
62007CJ0416	Reg. 806/2003	Reg. 806/2003/5	Reg. 806/2003/5/a	Reg. 806/2003/5/a/2	Reg. 806/2003/5/a/2/d	Reg. 806/2003/5/a/2/d/i
62007CJ0416	Dir. 90/425	-	-	-	-	-
62017CJ0530	Dec. 2015/143	Dec. 2015/143/2	Dec. 2015/143/2/1	-	-	-

Table 1: Examples of the structure of citations extracted from the judgments in the EUR-Lex dataset. *Dir.*, *Reg.* and *Dec.* are abbreviations of *Directive*, *Regulation* and *Decision*.

extract citations towards *Directives*, *Decisions*, and *Regulations*, following the numbering rules for articles and sub-levels.

In Tab. 1, we show some examples of the structure of the citations as confirmed by EUR-Lex. A total of 36,116 unique legal citations were extracted, leading, on average, to 8.65 cited acts per legal judgment.

Considering that this dataset specifically falls in the legal domain, and that has both the textual content and citations, it has been exploited for all the performed experiments, namely, *i)* for the evaluation of the effectiveness of the proposed overlapping clustering approach implemented in MOSTA, *ii)* for the evaluation of the performance of its fusion strategy, and *iii)* for the comparison with existing complete solutions for the final task of identifying the subject matters of legal documents.

Reuters-21578. This dataset consists of the train split of the ModHayes Reuters-21578 subset[6], which contains 9873 textual documents associated with one or more topics, collected from the Reuters financial newswire service during the 1987. The number of distinct topics in this dataset is 118.

ArXiv. This dataset consists of the train split of the arXiv dataset[7], which contains 4998 arXiv abstracts of submitted papers, associated with one or

[6]https://huggingface.co/datasets/reuters21578
[7]https://huggingface.co/datasets/arxiv_dataset

more system tags. In this dataset, we can find a total of 166 distinct tags.

For both Reuters-21578 and ArXiv, we pre-processed the text by *i)* lowercasing the text, *ii)* removing punctuation, and *iii)* removing stop words. Considering that these datasets only contain textual content (i.e., with no citations) and that each document is possibly associated with multiple topics/tags, they were considered appropriate only for the first part of the experiments, i.e., for the evaluation of the proposed overlapping clustering approach.

4.2. Experimental setting and competitor systems

In MOSTA, the embedding dimensionality of both the Word2Vec model for the content-based embedding and the Node2Vec model for the citation-based embedding was set to 256, i.e., $D_C = D_A = 256$, which is a pretty standard value adopted for these architectures [45, 38]. The remaining parameters for Node2Vec were left to their default value, i.e., $p = 1$, $q = 1$, $l = 80$ (length of random walks), and $r = 10$ (number of random walks).

For the evaluation of the clustering performance, we considered two competitor algorithms. The first is Neo K-Means [47], which identifies overlapping clusters on the basis of a user-defined input threshold α. In Neo K-Means, this parameter represents the average number of additional cluster assignments per document. For the estimation of the optimal value of α, we adopted the automatic strategy proposed in [47]. Moreover, we also evaluated the results obtained when the optimal value of α is known a-priori, by relying on the true number of cluster assignments in the dataset. Of course, the results obtained in such a configuration are over-optimistic, since such information is usually unknown in real scenarios. The second considered competitor algorithm is Fuzzy C-Means (FCM) [7]. FCM is a soft-clustering approach that returns the degree according to which each document belongs to each cluster. In order to determine the cluster assignments, it requires a user-defined threshold on such degrees. Since there is no automatic strategy to determine such a threshold, we normalized the membership degrees in $[0; 1]$ and collected the results with different thresholds, i.e., 0.3, 0.5, and 0.7.

As regards the parameter k, common to MOSTA and its competitors, we run the experiments with different values following the rule of thumb, namely, $k \in \{\sqrt{|J|}/2, \sqrt{|J|}, 2\sqrt{|J|}, 4\sqrt{|J|}, 8\sqrt{|J|}, 16\sqrt{|J|}\}$. The results with additional low values of k (e.g., $\sqrt{|J|}/4$, $\sqrt{|J|}/8$, and $\sqrt{|J|}/16$) are not reported, since the obtained results appeared to be consistently worse with respect to adopting higher values, for all the considered systems and parameter configurations. On the other hands, we do not report the results with values for k higher than $16\sqrt{|J|}$, because from $32\sqrt{|J|}$ the performance of MOSTA naturally started to decrease since k was quickly degenerating to $|J|$ (note that, in the EUR-Lex dataset, $|J| = 4176$, and $32\sqrt{|J|} = 2068$).

To specifically evaluate the effectiveness of the proposed multi-view fusion strategy, we compared the results obtained by the AE implemented in MOSTA with those achieved by other fusion approaches. The AE implemented in MOSTA was structured with a simple 3-layers architecture with only one hidden layer, corresponding to the bottleneck layer, with a dimensionality of 256, namely, $D_F = 256$, and *sigmoid* as activation function. We also evaluated the influence on the final results of the weight λ of the custom loss function θ defined in Eq. (5). In particular, we performed the experiments with $\lambda \in \{0.0, 0.1, 0.2, \ldots, 0.8, 0.9, 1.0\}$. As competitor approaches, we considered the simple concatenation of content-based and citation-based embeddings, i.e., $A \oplus C$, and a feature weighting approach applied on such a concatenation, which weighs each feature according to desired importance to apply to its source view (i.e., content or citation embeddings). For this competitor approach, we considered the following configurations of the weights for content and citation embeddings, respectively: $\langle 0.1, 0.9\rangle$, $\langle 0.3, 0.7\rangle$, $\langle 0.7, 0.3\rangle$, and $\langle 0.9, 0.1\rangle$. Intuitively, $\langle 0.1, 0.9\rangle$ gives more importance to the content than to citations, while $\langle 0.9, 0.1\rangle$ does the opposite. Note that we did not consider the configuration $\langle 0.5, 0.5\rangle$ because it corresponds to the simple concatenation approach $A \oplus C$, since it provides the same weight to all the features.

Finally, as mentioned at the beginning of Sec. 4, we compared the results achieved by MOSTA with those achievable by complete competitor solutions, on the EUR-Lex dataset. Specifically, for the construction of content-based embeddings, we considered pre-trained BERT-based models fine-tuned for the legal field [10], namely LEGAL-BERT BASE (768-dimensional embeddings), LEGAL-BERT EURLEX (768-dimensional embeddings), and LEGAL-BERT SMALL (512-dimensional embeddings). Note that LEGAL-BERT EURLEX is specifically fine-tuned on the dataset adopted in this evaluation, which, in principle, could provide it some advantages. Since BERT-based models support the embedding of documents with maximum 512 tokens [16], we adopted two strategies [44, 34]: TS_1, that preserves the first 512 tokens of each legal judgment, and TS_2 that preserves the first and the last part of each document, cutting off the middle part. For the clustering phase, we considered Neo K-Means, since it provided the best results among the competitors in the first part of our experiments aimed at evaluating the clustering performance.

4.3. Evaluation measure

Since the datasets contain the true topics/subject matters assigned to each document, as evaluation measure, we collected the F1-score averaged over the clusters, computed after the identification of the best cluster-topic matching through the Hungarian algorithm [25]. Therefore, for each cluster:

- a True Positive (TP) is a document/judgment that is labeled with the topic/subject matter matched with the cluster;
- a False Positive (FP) is a document/judgment falling in the cluster which is not labeled with the topic/subject matter matched with the cluster;
- a True Negative (TN) is a document/judgment that did not fall in the cluster and is not labeled with the topic/subject matter matched with the cluster;
- a False Negative (FN) is a document/judgment that did not fall in the cluster, but is labeled with the topic/subject matter matched with the cluster.

Note that this evaluation setting is coherent with that usually adopted for multi-label classification tasks [42].

The adoption of the average F1-score, instead of other measures like the accuracy, is motivated by its ability to evaluate the quality of the result without being biased by data unbalancing. Indeed, in the considered datasets, we can notice a strong unbalancing (see Fig. 8). For the clustering task at hand, the presence of unbalanced data corresponds to the fact that a few dominant topics/subject matters, whose documents may be widely and unevenly dispersed in the feature space, may partially obscure other topics/subject matters that are less prominent and equally dispersed [9], making their modeling by clustering algorithms much more difficult. This observation further motivates the adoption of the F1-score as evaluation measure.

Note that, thanks to the availability of the ground truth in the datasets, also in that specifically related to the legal field (EUR-Lex), we had the

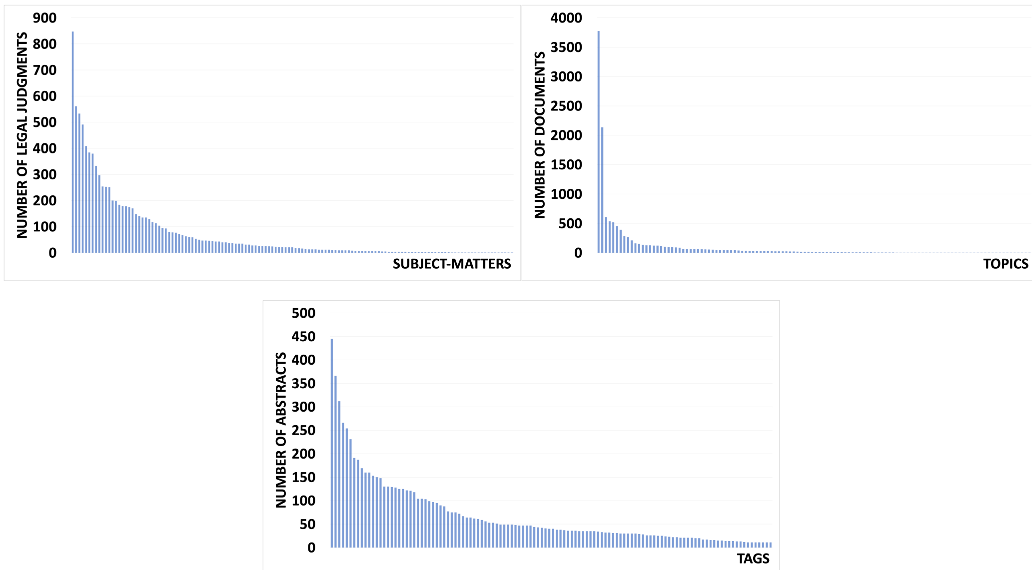


Figure 8: Number of legal judgments, documents and abstracts assigned to each subject matter, topic and tag, respectively, in the datasets EUR-Lex, Reuters and arXiv.

possibility to avoid the adoption of internal clustering quality measures, such as clustering agreement measures [39], since, when applied to overlapping clustering tasks, they tend to reward specific patterns in the resulting clusters (e.g., a low/high overlapping degree among clusters).

4.4. Results and Discussion

In Tab. 2, we report the F1-score results related to the evaluation of the overlapping clustering method implemented in MOSTA (see Sec. 3.4), applied only on the textual content of the three considered datasets.

We compared the results with those obtained by *i)* Neo K-Means with the automatic estimation of its parameter α, indicated as **N (est.** α**)**; *ii)* Neo K-Means with the optimal value of its parameter α, indicated as **N (opt.** α**)**; *iii)* Fuzzy C-Means, with different values of the threshold p applied on the membership degrees, indicated as $\boldsymbol{FCM_p}$, with $p \in \{0.3, 0.5, 0.7\}$. The reported results refer to the F1-score obtained with different values of k. In the same table, we also report the average rank achieved by a given configuration, with respect to the clustering algorithm (last column of each sub-table) and k (last row of each sub-table).

Focusing on the value of k, we can observe higher F1-score results with higher values of k. This is probably due to the high unbalancing in the dataset (see Fig. 8), which makes the clustering algorithms more capable of modeling the high amount of poorly represented subject matters in the dataset when requiring a higher number of (thus, generally smaller) clusters.

Looking at the results obtained by different clustering algorithms, we can easily conclude that MOSTA generally outperforms all the competitors. This is also clear by observing the average ranks (last column of each sub-table). The only case in which a competitor, i.e., Neo K-Means, is able to compete with MOSTA is on the EUR-Lex dataset, but only when fed with the ground value of its parameter α. The results obtained by FCM, with all the considered values of its threshold, generally appear below those achieved by MOSTA and Neo K-Means. Therefore, all the subsequent analyses have been performed only considering these two algorithms.

In Tab. 3, we report the F1-score results obtained on the EUR-Lex dataset, considering both the textual content and the citations, with different values of λ for the multi-view fusion phase. In the same table, we also report the average rank achieved by a given configuration, with respect to λ (last column of each sub-table) and k (last row of each sub-table). As we can observe from the table, also when using other algorithms for the clustering phase, i.e., N (opt. α) and N (est. α), there is some influence coming from the value of λ. Specifically, the best overall results were achieved with $\lambda = 0.1$ for MOSTA, $\lambda = 0.4$ for N (opt. α) and $\lambda = 0.2$ for N (est. α). This result proves the usefulness of considering the information conveyed by the citations in the multi-view fusion phase, irrespectively from the algorithm adopted for the clustering phase. Therefore, citation-based embeddings can be considered a useful complement to content-based embeddings, since they positively contribute to the clustering results.

Overall, we can observe that the F1-score values obtained by MOSTA are much higher than those obtained by N (est. α) and N (opt. α). In Tab. 4, we make a direct comparison between MOSTA, N (est. α) and N (opt. α), considering the best values of λ for each of them. As we can see from the results, independently on the value of k, MOSTA consistently outperforms N (est. α), and outperforms N (opt. α) in 4 out 6 cases, even if the latter exploits the true value of α that, in principle, cannot be known a-priori. The clear dominance of the clustering algorithm implemented in MOSTA (on average, 10% higher F1-scores than N (opt. α) and 149% higher F1-scores than N (est. α)), also confirmed by the average ranks (see the last row of Tab. 4), confirms the effectiveness of the proposed outlier-based approach.

	alg. \ k	$\sqrt{\|J\|}/2$	$\sqrt{\|J\|}$	$2\sqrt{\|J\|}$	$4\sqrt{\|J\|}$	$8\sqrt{\|J\|}$	$16\sqrt{\|J\|}$	**AvgRank**
EUR-LEX	MOSTA	*0.102*	*0.147*	*0.191*	0.225	0.254	0.278	**1.50**
	N (opt. α)	0.079	0.127	0.184	*0.242*	*0.283*	*0.314*	**1.50**
	N (est. α)	0.053	0.068	0.081	0.101	0.126	0.149	3.00
	$FCM_{0.3}$	0.029	0.034	0.037	0.039	0.043	0.046	4.42
	$FCM_{0.5}$	0.028	0.031	0.035	0.039	0.041	0.047	5.25
	$FCM_{0.7}$	0.028	0.034	0.035	0.037	0.041	0.046	5.33
AvgRank		6.00	5.00	4.00	3.00	2.00	**1.00**	

	alg. \ k	$\sqrt{\|J\|}/2$	$\sqrt{\|J\|}$	$2\sqrt{\|J\|}$	$4\sqrt{\|J\|}$	$8\sqrt{\|J\|}$	$16\sqrt{\|J\|}$	**AvgRank**
REUTERS	MOSTA	*0.079*	*0.095*	*0.148*	*0.195*	*0.226*	*0.304*	**1.00**
	N (opt. α)	0.066	0.082	0.130	0.166	0.217	0.278	2.00
	N (est. α)	0.034	0.040	0.048	0.058	0.079	0.095	3.00
	$FCM_{0.3}$	0.024	0.025	0.025	0.025	0.025	0.025	6.00
	$FCM_{0.5}$	0.024	0.025	0.025	0.025	0.025	0.025	5.00
	$FCM_{0.7}$	0.024	0.025	0.025	0.025	0.025	0.025	4.00
AvgRank		6.00	5.00	3.33	2.83	2.17	**1.67**	

	alg. \ k	$\sqrt{\|J\|}/2$	$\sqrt{\|J\|}$	$2\sqrt{\|J\|}$	$4\sqrt{\|J\|}$	$8\sqrt{\|J\|}$	$16\sqrt{\|J\|}$	**AvgRank**
ARXIV	MOSTA	*0.061*	*0.095*	*0.137*	*0.164*	*0.193*	0.209	**1.17**
	N (opt. α)	0.050	0.083	0.120	0.157	0.187	*0.224*	1.83
	N (est. α)	0.027	0.039	0.049	0.070	0.106	0.129	3.17
	$FCM_{0.3}$	0.018	0.025	0.030	0.033	0.033	0.033	5.17
	$FCM_{0.5}$	0.023	0.031	0.036	0.038	0.039	0.040	4.17
	$FCM_{0.7}$	0.028	0.015	0.020	0.018	0.016	0.017	5.50
AvgRank		5.17	5.17	3.67	2.83	2.33	**1.83**	

Table 2: F1-score results obtained on the textual content of the EUR-Lex, Reuters and arXiv datasets by the clustering algorithm implemented in MOSTA, Neo K-Means (opt. α), Neo K-Means (est. α) and Fuzzy C-Means with different thresholds applied on the membership degrees. Best column-wise results are emphasized with a gray background.

	λ \ k	$\sqrt{\lvert J\rvert}/2$	$\sqrt{\lvert J\rvert}$	$2\sqrt{\lvert J\rvert}$	$4\sqrt{\lvert J\rvert}$	$8\sqrt{\lvert J\rvert}$	$16\sqrt{\lvert J\rvert}$	**AvgRank**
MOSTA	0.0	0.104	0.139	0.201	0.239	0.252	0.281	10.33
	0.1	0.124	*0.205*	0.250	0.280	*0.305*	0.301	**2.00**
	0.2	0.122	0.196	*0.261*	*0.286*	0.287	2.92	3.00
	0.3	0.121	0.194	0.254	0.263	0.298	0.294	3.50
	0.4	*0.127*	0.181	0.238	0.264	0.272	0.306	3.67
	0.5	0.116	0.181	0.228	0.256	0.277	*0.313*	4.17
	0.6	0.112	0.156	0.220	0.254	0.277	0.300	6.50
	0.7	0.120	0.164	0.216	0.253	0.270	0.296	6.50
	0.8	0.115	0.162	0.205	0.251	0.277	0.285	7.67
	0.9	0.115	0.161	0.216	0.246	0.272	0.293	8.00
	1.0	0.102	0.147	0.191	0.225	0.254	0.278	10.67
AvgRank		6.00	5.00	4.00	3.00	1.82	**1.18**	

	λ \ k	$\sqrt{\lvert J\rvert}/2$	$\sqrt{\lvert J\rvert}$	$2\sqrt{\lvert J\rvert}$	$4\sqrt{\lvert J\rvert}$	$8\sqrt{\lvert J\rvert}$	$16\sqrt{\lvert J\rvert}$	**AvgRank**
N (opt. α)	0.0	0.057	0.097	0.158	0.174	0.188	0.245	11.00
	0.1	0.087	0.151	0.230	0.259	0.255	0.262	6.17
	0.2	0.085	0.159	0.211	0.239	0.286	0.282	6.83
	0.3	*0.098*	0.150	*0.238*	0.268	0.291	0.299	3.83
	0.4	0.096	*0.161*	0.238	*0.282*	0.290	0.323	**2.50**
	0.5	0.086	0.151	0.213	0.274	0.295	*0.324*	3.67
	0.6	0.092	0.142	0.206	0.266	*0.296*	*0.324*	4.17
	0.7	0.080	0.143	0.207	0.246	0.291	0.317	6.67
	0.8	0.091	0.151	0.194	0.258	0.294	0.305	5.50
	0.9	0.092	0.125	0.184	0.257	0.269	0.323	7.00
	1.0	0.079	0.127	0.184	0.242	0.283	0.314	8.67
AvgRank		6.00	5.00	4.00	2.91	2.00	**1.09**	

	λ \ k	$\sqrt{\lvert J\rvert}/2$	$\sqrt{\lvert J\rvert}$	$2\sqrt{\lvert J\rvert}$	$4\sqrt{\lvert J\rvert}$	$8\sqrt{\lvert J\rvert}$	$16\sqrt{\lvert J\rvert}$	**AvgRank**
N (est. α)	0.0	0.031	0.043	0.055	0.078	0.115	0.137	8.83
	0.1	0.065	0.076	0.084	0.096	*0.144*	*0.149*	5.83
	0.2	0.066	0.081	0.090	0.097	0.136	0.142	5.00
	0.3	*0.070*	*0.085*	0.094	0.100	0.122	0.129	**3.17**
	0.4	0.067	0.084	0.095	*0.105*	0.111	0.114	**3.17**
	0.5	0.065	0.082	*0.096*	0.102	0.108	0.112	4.17
	0.6	0.064	0.081	0.093	0.103	0.106	0.109	6.17
	0.7	0.064	0.080	0.091	0.098	0.105	0.109	7.83
	0.8	0.064	0.079	0.091	0.099	0.105	0.110	7.33
	0.9	0.063	0.078	0.092	0.098	0.104	0.110	8.00
	1.0	0.053	0.068	0.081	0.101	0.126	*0.149*	6.50
AvgRank		6.00	5.00	3.91	3.09	2.00	**1.00**	

Table 3: F1-score results obtained on EUR-Lex (both content and citations) with different values of λ and k. The last column of each sub-table is the average rank of a given value of λ (by varying k), while the last row of each sub-table is the average rank of a given value of k (by varying λ) Best column-wise results are emphasized with a gray background.

k	**MOSTA**	**N** (opt. α)	**N** (est. α)
$\sqrt{\lvert J\rvert}/2$	*0.124*	0.096	0.067
$\sqrt{\lvert J\rvert}$	*0.205*	0.161	0.084
$2\sqrt{\lvert J\rvert}$	*0.250*	0.238	0.095
$4\sqrt{\lvert J\rvert}$	0.280	*0.282*	0.105
$8\sqrt{\lvert J\rvert}$	*0.305*	0.290	0.111
$16\sqrt{\lvert J\rvert}$	0.301	*0.323*	0.114
AvgRank	**1.33**	1.67	3.00

Table 4: F1-score results obtained on EUR-Lex (both content and citations) by N (opt. α) and N (est. α), and by the clustering algorithm implemented in MOSTA, with the best fusion strategy (i.e., the AE implemented in MOSTA, as shown in Tab. 5), with their respective best value for λ, i.e., $\lambda = 0.1$ for MOSTA, $\lambda = 0.4$ for N (opt. α)), and $\lambda = 0.4$ for N (est. α)). Best row-wise results are emphasized with a gray background.

k	**MOSTA**	**C⊕A**	**C⊕A** $\langle 0.1, 0.9\rangle$	**C⊕A** $\langle 0.3, 0.7\rangle$	**C⊕A** $\langle 0.7, 0.3\rangle$	**C⊕A** $\langle 0.9, 0.1\rangle$
$\sqrt{\lvert J\rvert}/2$	*0.124*	0.119	0.098	0.115	0.100	0.093
$\sqrt{\lvert J\rvert}$	*0.205*	0.178	0.135	0.162	0.148	0.139
$2\sqrt{\lvert J\rvert}$	*0.250*	0.214	0.200	0.214	0.194	0.185
$4\sqrt{\lvert J\rvert}$	*0.280*	0.249	0.257	0.249	0.243	0.230
$8\sqrt{\lvert J\rvert}$	*0.305*	0.262	0.276	0.269	0.270	0.260
$16\sqrt{\lvert J\rvert}$	0.301	0.285	0.279	0.292	*0.309*	0.304
Avg. Rank	**1.33**	3.50	4.17	3.17	3.67	5.17

Table 5: F1-score results obtained on EUR-Lex (both content and citations) by the AE-based fusion strategy implemented in MOSTA ($\lambda = 0.1$) and by other fusion strategies based on the simple concatenation and on feature weighting, with different values of k. Best row-wise results are emphasized with a gray background.

In Tab. 5, we report the results of a further analysis aiming to specifically evaluate the contribution of the AE-based multi-view fusion strategy implemented in MOSTA. In particular, we compare it with the concatenation of the embeddings $C \oplus A$, as well as with an approach based on feature weighting, considering different weights for each view (see Sec. 4 for details).

The results show that the proposed AE-based fusion strategy outperforms the other considered techniques in almost all the situations (i.e., for almost all the considered values of k). The influence of the weight on the feature importance adopted for the considered competitor approach does not appear to influence the results in a consistent way. In other words, determining the best weight appears to be very challenging and dependent on the value of k. For this specific analysis, we can conclude that the superiority of the AE-based fusion strategy implemented in MOSTA is clear, and also confirmed by the observed average ranks (see the last row of Tab. 5). These results confirm that the proposed approach is able to significantly alleviate the issues possibly introduced by the curse of dimensionality and to identify a fused feature space that properly represents the complementary information conveyed by the textual content and by cited legal acts.

Finally, in Tab. 6 we report the results of a comparison between the whole method MOSTA and possible combinations of competitor systems that could be adopted to solve the considered task on the EUR-Lex dataset. Specifically, as described in Sec. 4, we adopted different BERT-based embedding models, and N (est. α) as the clustering algorithm. Note that, in this case, a comparison with N (opt. α) would be totally unfair, since in real-world scenarios, we cannot assume to know the true value of α. On the contrary, both N (est. α) and MOSTA automatically identify the best estimate for their parameters.

The F1-scores shown in Tab. 6 emphasize that MOSTA always outperforms all the competitors, independently on the adopted embedding model, truncation strategy, and value of k. Indeed, MOSTA always ranks as the first (best) method, in all the configurations (see the last row of Tab. 6). On average, we can observe an improvement of 203%, 151% and 186% over the results obtained when adopting LEGAL-BERT BASE, LEGAL-BERT SMALL, and LEGAL-BERT EURLEX, respectively, as embedding models. It is noteworthy that, among the competitor approaches adopted for the embedding, LEGAL-BERT SMALL appears to be the best solution, even if not specifically fine-tuned on the considered EUR-Lex dataset as LEGAL-BERT EURLEX. This is probably due to the slightly lower number of features of its embeddings (512 instead of 768), that alleviates the issues possibly introduced by the curse of dimensionality. This observation further confirms the appropriateness of the approach adopted by MOSTA.

Together with the specific analyses on the contribution provided by the proposed overlapping clustering algorithm, by the citation-based embeddings, and by the multi-view AE-based fusion strategy, these final results

	LEGAL-BERT BASE		LEGAL-BERT SMALL		LEGAL-BERT EURLEX		**MOSTA**
k	$\mathbf{TS}_1$	$\mathbf{TS}_2$	$\mathbf{TS}_1$	$\mathbf{TS}_2$	$\mathbf{TS}_1$	$\mathbf{TS}_2$	
$\sqrt{\lvert J\rvert}/2$	0.046	0.041	0.057	0.047	0.048	0.042	*0.124*
$\sqrt{\lvert J\rvert}$	0.058	0.054	0.070	0.064	0.058	0.055	*0.205*
$2\sqrt{\lvert J\rvert}$	0.074	0.064	0.080	0.083	0.071	0.069	*0.250*
$4\sqrt{\lvert J\rvert}$	0.091	0.083	0.100	0.105	0.089	0.100	*0.280*
$8\sqrt{\lvert J\rvert}$	0.109	0.109	0.133	0.148	0.127	0.122	*0.305*
$16\sqrt{\lvert J\rvert}$	0.144	0.151	0.176	0.192	0.169	0.177	*0.301*
AvgRank	5.17	6.83	2.83	2.50	4.67	5.00	**1.00**

Table 6: F1-score results obtained on EUR-Lex (both content and citations) by MOSTA (λ = 0.1) and existing complete solutions, where the embedding is based on different BERT-based models, using the different truncation strategies TS_1 and TS_2, and clustering is performed by N (est. α). Best row-wise results are emphasized with a gray background.

prove that the whole workflow implemented in MOSTA, that simultaneously exploits the information conveyed by the textual content and by cited legal acts, as well as its novel overlapping clustering method, can be considered a precious tool for the unsupervised identification of the subject matters of legal judgments.

5. Conclusions

In this paper, we proposed MOSTA, a novel method to identify groups of legal judgments according to their characteristics. MOSTA is able to identify a fused representation that considers both the textual content of legal judgments and the legal acts they cite, properly taking into account the granularity of the citations. Moreover, MOSTA adopts a novel overlapping clustering method that does not require additional input parameters to define the desired degree of cluster overlap, but automatically identifies additional cluster assignments by exploiting an outlier-based strategy.

The specific evaluation of the performance of the proposed clustering algorithm on three textual datasets proved that MOSTA is able to outperform Neo K-Means and Fuzzy C-Means, also considering different values of their input parameters. Moreover, the experiments performed on a real legal dataset provided by EUR-Lex emphasized that *i)* properly taking into account citations can provide a positive contribution to the quality of the identified clusters; *ii)* the proposed AE-based fusion strategy generally outperforms concatenation-based approaches, including those that exploit feature weighting; *iii)* the clustering algorithm implemented in MOSTA outperforms Neo K-Means, even when providing it with the optimal value of its input parameter; *iv)* the whole method implemented in MOSTA outperforms existing complete solutions based on the combination of pre-trained models for document embedding and clustering.

For future work, we will take into account the aspects related to the explainability of the output, in order to make the clusters extracted by MOSTA understandable and trustable [22]. Moreover, we will investigate the possibility to exploit the groups of legal judgments identified by MOSTA to provide actual suggestions during the preparation of new legal judgments. In particular, we will explore the application of process mining techniques to clusters of sequences of paragraphs to suggest the next paragraph to add to a legal judgment under preparation.

Appendix A. Symbols

Symbol	Description
J	A set of legal judgments
k	The number of clusters/groups of legal judgments to identify
Embedding of the textual content of legal judgments	
C	Content-based embeddings of the legal judgments J
D_C	Dimensionality of the content-based embeddings
w_i, $\vec{w}_i$	A context word and its one-hot vector representation
w_t, $\vec{w}_t$	A target word and its one-hot vector representation
h	Size of the context window
V	Size of the vocabulary observed in the set of legal judgments J
S	Weight matrix learned by the Word2Vec model
Embedding of the citations of legal judgments	
A	Citation-based embeddings of the legal judgments J
D_A	Dimensionality of the citation-based embeddings
T_i	An ordered tree representing the citations of the document J_i
$s(T_i, T_j)$	Tree similarity between the ordered trees T_i and T_j
$\delta(T_i, T_j)$	Tree edit distance between the ordered trees T_i and T_j
$\mathcal{G} = (N, E)$	A weighted graph. $\mathcal{N}$ = legal judgments J; $\mathcal{E}$ = co-citations of legal acts
r, l	Number and length of Node2Vec random walks for each node
$\beta(J_i, J_k)$	The function defining the likelihood to reach the node J_k starting from J_i
$g(J_i, J_k)$	The distance (i.e., number of steps) between J_i and J_k in the graph
p, q	Node2Vec parameters to bias random walks
Multi-view embeddings fusion	
F	Fused, compressed, embeddings
D_F	The dimensionality of the fused latent representation
θ	The loss function adopted in the AE
λ	Importance of content-based embeddings in the AE loss
γ	The vector of weights used by the AE loss, based on the parameter λ
Identification of overlapping clusters of legal judgments	
k	Number of overlapping clusters of judgments to identify
K	Set of overlapping clusters of legal judgments
$\bar{d}, \sigma$	The mean and the standard deviation of the judgment-cluster distances

Appendix B. Regular expressions for the extraction of citations

In the following, we report the Regular Expressions adopted to extract the citations from the legal judgments of the EUR-Lex dataset:

```
1: (?<!of\s)(council\s)*(?<!of\scouncil\s)(regulation|decision|directive
   ↪ )((\s\((cfsp|ec|ecsc|eec|eu|euratom|jha|op_dat%pro)\))*\s\d+/\d
   ↪ +(/(cfsp|ec|ecsc|eec|eu|euratom|jha|op_datpro))*)+(((\s(-|{| | ))
   ↪ *(\s((articles?)|(arts?))(\s\d+)+%)+((\s\d+)|([a-z])|(\(\w+\))|(\
   ↪ s\(\w+\)))*))*,

2: (((articles?)|(arts?))\s\d+([a-z])*)+((\s\d+)|([a-z])|(\(\w+\))|(\s
   ↪ \(\w+\)))*(\sof)*\s(council\s)*(regulation|decision|directive)((\
   ↪ s\((cfsp|ec|ecsc|eec|eu|euratom|jha|op_datpro)\))*(\s\d+/\d+(/(
   ↪ cfsp|ec|ecsc|eec|eu|euratom|jha|op_datpro))*))
```

Availability: The system, the dataset and all the results are available at: https://osf.io/a9jm2/?view_only=471428680ce5483abc358fa17a99ad5f.

Acknowledgements: This work was partially supported by the project FAIR - Future AI Research (PE00000013), Spoke 6 - Symbiotic AI (CUP H97G22000210007), under the NRRP MUR program funded by the NextGenerationEU. The authors also acknowledge the support of the project "START UPP – Modelli, sistemi e competenze per l'implementazione dell'ufficio per il processo" (CUP H29J22000390006) funded by the Italian Minister of Justice.

References

[1] Aggarwal, C.C., Gates, S.C., Yu, P.S., 2004. On using partial supervision for text categorization. IEEE Transactions on Knowledge and Data Engineering 16, 245–255.

[2] Alessa, A., Jianlong, Z., Kevin, T., Fang, C., Andreas, H., 2022. Fairness and Explanation in AI-Informed Decision Making. Machine Learning and Knowledge Extraction 4, 556–579.

[3] Ashley, K.D., 2017. Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge University Press.

[4] Bai, R., Huang, R., Chen, Y., Qin, Y., 2021. Deep multi-view document clustering with enhanced semantic embedding. Inf. Sci. 564, 273–287.

[5] Ballard, D.H., 1987. Modular learning in neural networks, in: Proceedings of the 6th National Conference on Artificial Intelligence. Seattle, WA, USA, July 1987, Morgan Kaufmann. pp. 279–284.

[6] Barracchia, E.P., Pio, G., Bifet, A., Gomes, H.M., Pfahringer, B., Ceci, M., 2022. LP-ROBIN: Link prediction in dynamic networks exploiting incremental node embedding. Information Sciences 606, 702–721.

[7] Bezdek, J.C., Ehrlich, R., Full, W., 1984. Fcm: The fuzzy c-means clustering algorithm. Computers & Geosciences 10, 191–203.

[8] Bird, S., Klein, E., Loper, E., 2009. Natural Language Processing with Python. O'Reilly.

[9] Carrington, A.M., Fieguth, P.W., Qazi, H., Holzinger, A., Chen, H.H., Mayr, F., Manuel, D.G., 2020. A new concordant partial AUC and partial c statistic for imbalanced data in the evaluation of machine learning algorithms. BMC Medical Informatics Decis. Mak. 20, 4.

[10] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I., 2020. LEGAL-BERT: The muppets straight out of law school, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online. pp. 2898–2904.

[11] de Colla Furquim, L.O., de Lima, V.L.S., 2012. Clustering and categorization of brazilian portuguese legal documents, in: PROPOR 2012, Coimbra, Portugal, April 17-20, 2012. Proceedings, pp. 272–283.

[12] Conrad, J.G., Al-Kofahi, K., Zhao, Y., Karypis, G., 2005. Effective document clustering for large heterogeneous law firm collections, in: Proceedings of the 10th International Conference on Artificial Intelligence and Law, p. 177–187.

[13] De Martino, G., Pio, G., Ceci, M., 2022. PRILJ: an efficient two-step method based on embedding and clustering for the identification of regularities in legal case judgments. Artificial Intelligence and Law 30, 359–390.

[14] Deb, K., Agrawal, S., Pratap, A., Meyarivan, T., 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. IEEE Trans. Evol. Comput. 6, 182–197.

[15] Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805.

[16] Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding, in: ACL 2019, Association for Computational Linguistics. pp. 4171–4186.

[17] Draper, N., Smith, H., 1966. Applied regression analysis. Wiley series in probability and mathematical statistics, Wiley.

[18] Gao, Y., Gu, S., Xia, L., Fei, Y., 2006. Web document clustering with multi-view information bottleneck, in: CIMCA 2006, International Conference on Intelligent Agents, Web Technologies and Internet Commerce (IAWTIC 2006), IEEE Computer Society. p. 148.

[19] Grover, A., Leskovec, J., 2016. node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, ACM. pp. 855–864.

[20] Hess, S., Pio, G., Hochstenbach, M.E., Ceci, M., 2021. BROCCOLI: overlapping and outlier-robust biclustering through proximal stochastic gradient descent. Data Min. Knowl. Discov. 35, 2542–2576.

[21] Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. Mach. Learn. 42, 177–196.

[22] Holzinger, A., 2021. The next frontier: AI we can really trust, in: Machine Learning and Principles and Practice of Knowledge Discovery in Databases - International Workshops of ECML PKDD 2021, Virtual Event, September 13-17, 2021, Proceedings, Part I, Springer. pp. 427–440.

[23] Hussain, S.F., Mushtaq, M., Halim, Z., 2014. Multi-view document clustering via ensemble method. J. Intell. Inf. Syst. 43, 81–99.

[24] Kim, Y., Amini, M., Goutte, C., Gallinari, P., 2010. Multi-view clustering of multilingual documents, in: ACM SIGIR 2010, Geneva, Switzerland, July 19-23, 2010, ACM. pp. 821–822.

[25] Kuhn, H.W., 2010. The hungarian method for the assignment problem, in: 50 Years of Integer Programming 1958-2008 - From the Early Years to the State-of-the-Art. Springer, pp. 29–47.

[26] Kumar, A., Makhija, P., Gupta, A., 2020. Noisy text data: Achilles' heel of BERT, in: Proceedings of the Sixth Workshop on Noisy User-generated Text, W-NUT@EMNLP 2020 Online, November 19, 2020, Association for Computational Linguistics. pp. 16–21.

[27] Lewis, S., 2021. Precedent and the Rule of Law. Oxford Journal of Legal Studies 41, 873–898.

[28] Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.W., Sartor, G., Torroni, P., 2019. Claudette: an automated detector of potentially unfair clauses in online terms of service. Artificial Intelligence and Law 27, 117–139.

[29] Liu, R., Krishnan, A., 2021. PecanPy: a fast, efficient and parallelized Python implementation of node2vec. Bioinformatics 37, 3377–3379.

[30] Lu, Q., Conrad, J.G., Al-Kofahi, K., Keenan, W., 2011. Legal document clustering with built-in topic segmentation, in: CIKM 2011, Glasgow, United Kingdom, October 24-28, 2011, ACM. pp. 383–392.

[31] Mandal, A., Ghosh, K., Ghosh, S., Mandal, S., 2021. Unsupervised approaches for measuring textual similarity between legal court case reports. Artificial Intelligence and Law 29, 417–451.

[32] Medvedeva, M., Dam, T., Wieling, M., Vols, M., 2021. Automatically identifying eviction cases and outcomes within case law of dutch courts of first instance, in: Legal Knowledge and Information Systems - JURIX 2021: The Thirty-fourth Annual Conference, Vilnius, Lithuania, 8-10 December 2021, IOS Press. pp. 13–22.

[33] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: NIPS 2013, Curran Associates Inc., Red Hook, NY, USA. p. 3111–3119.

[34] Mutasodirin, M.A., Prasojo, R.E., 2021. Investigating Text Shortening Strategy in BERT: Truncation vs Summarization, in: (ICACSIS) 2021, pp. 1–5.

[35] Pawlik, M., Augsten, N., 2016. Tree edit distance: Robust and memory-efficient. Inf. Syst. 56, 157–173.

[36] Pellicani, A., Pio, G., Redavid, D., Ceci, M., 2023. SAIRUS: Spatially-aware identification of risky users in social networks. Information Fusion 92, 435–449.

[37] Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: EMNLP 2014, October 25-29, 2014, Doha, Qatar, pp. 1532–1543.

[38] Qiao, Y., Zhang, B., Zhang, W., 2020. Malware classification method based on word vector of bytes and multilayer perception, in: 2020 IEEE International Conference on Communications, ICC 2020, Dublin, Ireland, June 7-11, 2020, pp. 1–6.

[39] Rabbany, R., Zaïane, O.R., 2015. Generalization of clustering agreements and distances for overlapping clusters and network communities. Data Mining and Knowledge Discovery 29, 1458–1485.

[40] Sabo, I.C., Pont, T.R.D., Wilton, P.E.V., Rover, A.J., Hübner, J.F., 2022. Clustering of brazilian legal judgments about failures in air transport service: an evaluation of different approaches. Artificial Intelligence and Law 30, 21–57.

[41] Sloan, A.E., 2018. Basic legal research: Tools and strategies. Wolters Kluwer.

[42] Song, D., Vold, A., Madan, K., Schilder, F., 2022. Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre-training. Information Systems 106, 101718.

[43] Stratos, K., Collins, M., Hsu, D.J., 2015. Model-based word embeddings from decompositions of count matrices, in: ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, pp. 1282–1291.

[44] Sun, C., Qiu, X., Xu, Y., Huang, X., 2019. How to fine-tune BERT for text classification?, in: CCL 2019, Kunming, China, October 18-20, 2019, Proceedings, Springer. pp. 194–206.

[45] Tissier, J., Gravier, C., Habrard, A., 2019. Near-lossless binarization of word embeddings, in: AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, pp. 7104–7111.

[46] Wahid, A., Gao, X., Andreae, P., 2014. Multi-view clustering of web documents using multi-objective genetic algorithm, in: CEC 2014, Beijing, China, July 6-11, 2014, IEEE. pp. 2625–2632.

[47] Whang, J.J., Dhillon, I.S., Gleich, D.F., 2015. Non-exhaustive, overlapping k-means, in: Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015, SIAM. pp. 936–944.

[48] Zamora, J., Sublime, J., 2020. A new information theory based clustering fusion method for multi-view representations of text documents, in: SCSM 2020, Held as Part of HCII 2020, Copenhagen, Denmark, July 19-24, 2020, Proceedings, Part I, Springer. pp. 156–167.

[49] Zhan, K., Shi, J., Wang, J., Tian, F., 2017. Graph-regularized concept factorization for multi-view document clustering. J. Vis. Commun. Image Represent. 48, 411–418.

[50] Zhao, Y., Karypis, G., Fayyad, U., 2005. Hierarchical clustering algorithms for document datasets. Data Min. Knowl. Discov. 10, 141–168.