Information Systems 106 (2022) 101821

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/is

# Exploiting co-occurrence networks for classification of implicit inter-relationships in legal texts

Emilio Sulis *, Llio Humphreys, Fabiana Vernero, Ilaria Angela Amantea, Davide Audrito, Luigi Di Caro

*University of Turin - Computer Science Department, Corso Svizzera 185, 10149 Torino, Italy*

ARTICLE INFO

*Article history:*
Received 2 November 2020
Received in revised form 16 March 2021
Accepted 4 May 2021
Available online 6 June 2021
Recommended by Andrea Tagarelli

*Keywords:*
Information extraction
Legal databases
Text mining
Network analysis
Card sorting

ABSTRACT

The interpretation of any legal norm typically requires consideration of relationships between parts within the same piece of legislation. This work describes a general framework for the development of a system to identify and classify implicit inter-relationships between parts of a legal text. In particular, our approach demonstrates the usefulness of co-occurrence networks of terms, in a practical experimental setting based on an EU Regulation. First, a manual annotation task identify instances of different kinds of implicit links in the norm. In addition to a typical NLP pipeline, our framework includes a technique from Information Architecture, i.e. card sorting. Second, we construct co-occurrence networks of the law terms to derive graph metrics. Third, binary classification experiments identify the existence (and the type) of inter-relationships by using a Bag-of-Ngrams model integrated with network analysis features. The results demonstrate how the adoption of co-occurrence network features improves the identification of links, for all the classifiers here considered. This is encouraging toward a wider adoption of this kind of network analysis technique in legal informatics.



## 1. Introduction

Legal interpretation is required for adapting norms to unforeseen situations [1]. The interpretation of any legal norm typically requires consideration of a broader context than the norm itself, not least related norms within the same piece of legislation. Moreover, the relationships between different norms and within the same legal text are relevant in the process of construction of legal arguments.

In recent years there has been an increase of studies in legal informatics, also in legal relations domain [2] as well as legal citation networks. Computational approaches include the advances in Natural Language Processing (NLP) and Machine Learning (ML), along with the increased attention to graph analysis. In particular, two kinds of relations are typically considered in a legal text: explicit and implicit links. The majority of researches are concerned with explicit citations to legislation or case law in their entirety, or parts thereof. This work focuses on the discovery of *implicit inter-relationships*, i.e., the references between parts of a norm without explicit references. This is recognized as a difficult task for both humans and, even more so, machines.

Nevertheless, automatic systems are increasingly helpful in providing a support to the interpretation process of legal practitioners [3,4]. Several instruments have typically been equipped with Artificial Intelligence (AI) systems that extract information from the norms [5,6]. This work proposes a general framework towards an automated identification and classification of the above-mentioned implicit inter-relationships between parts of a legal text. The methodology combines a pipeline including NLP, ML, and a graph representation of legal text.

In particular, a specific contribution of the proposed approach is the adoption of a *Co-occurrence Network (CN)* analysis to improve the identification of implicit inter-relationships between parts of a legal text. Text CN is a particular type of network between items (from the given text) know as unigrams (e.g., the words), of which it is possible to compute graph measures related to the role and importance of vertices and edges. We demonstrate how such graph metrics can improve the accuracy of results. There have been various previous research addressing the detection, resolution, and labeling of citations in the legal domain. But to the best of our knowledge, there has not been any systematic approach exploiting CNs to improve a legal classification effort.

The identification of types of inter-relationships between the parts of a norm neither is a trivial task nor do annotated dataset exist. Therefore, a necessary requirement is to have a manually annotated training dataset. Our general framework aims to address a system for discovering *implicitly* related norms, firstly by human annotators, and then by automated means. We applied the methodological steps here presented in a practical use case involving a European Union regulation.

* Corresponding author.
*E-mail address:* emilio.sulis@unito.it (E. Sulis).

https://doi.org/10.1016/j.is.2021.101821
0306-4379/

Finally, this study addresses in a practical manner the following research questions:

- RQ1: how well do interrelationship types succeed in identifying implicit links between recitals and (sub-)articles of a Regulation?
- RQ2: can we detect such links using classification algorithms?
- RQ3: can co-occurrence network analysis metrics improve the performance of classification algorithms?

The paper is organized as follows. Section 2 describes the background with some related work and introduces the case study. Section 3 details the three parts of the methodological framework. Section 4 focuses on the annotation effort, while Section 5 describes CNs and the graph analysis. Finally, Section 6 summarize the classification experiments. We conclude the paper with some remarks and future work in Section 7.

## 2. Background

### 2.1. Legal citation network studies

Legal citation networks is by now a mature research area involving a range of different approaches and specific sub-problems. First, there is the task of identifying and extracting the citations themselves, challenging when references to different entities are in a range of formats [7] or where the citations are anaphoric or imprecise [8]. Moreover, citations can differ in the level of specificity, from the legislation in its entirety to an Arabic number within a sub-article [9]. Approaches include gazetteers and concept markers [8], regular expressions [9,10], Conditional Random Fields and BiLSTM neural networks [11], with the latter two achieving similar performance in an evaluation by [7]. In [12], a combined named entity recognition approach which engages rules and supervised learning is used to identify citations, with approximate string matching step employed to resolve typo issues and imperfect entities.

Second, some studies concern the task of ranking legal sources. In [13], a network analysis of citations in French legal codes is based on undirected edges between codes, where the weights are the number of citations between the two codes. Case law are often ranked in terms of authority (often-cited) and 'well-founded in law' hubs (citing many important decisions) or a combination of these measures. In [14], it is argued that in-degree centrality is a poor measure of importance of legal cases because a rarely cited judgment can be important if it is cited in an important judgment. Some domain-specific methodologies take into account aspects such as the competence level of the court, whether a news item was published on the courts' website, publication in jurisprudence magazines, and age of the judgments [15,16]. Similar analysis was conducted in [17] on a network of French legal codes.

Third, few works addressed the task of labeling citations. In [18], the authors seek to help legal practitioners with the task of gathering citations to case law to support their argumentation. Starting with a seed case, they undertake a recursive process of forward chaining to find cases citing by the current case, and backward chaining to find cases cited by the current case. It is a selective process, since cases are cited for numerous reasons. The authors use the text area around the citation, the Text of Interest (TOI), to identify the "Reason for Citing". A simple term-based vector comparison is used to measure the similarity between TOIs. With their Semantics-Based Legal Citation Network Viewer, the user can obtain a network of explicitly linked cases focussed on a particular legal issue. The citation network of legislation and case law in [19] is a complex multi-relationship model. One key distinction is between ready-only references and references that modify the text or life cycle of other legislation. Read-only references are classified as legal basis, instrument cited, affected by case, and other. This labeling allows different sub-networks to be identified and traversed: sub-networks of jurisprudence or treaties, a sub-network of secondary legislation that implement EU policy directly on member states, a sub-network of instruments cited resembling traditional citation networks, and a sub-network of legal basis to identify the internal hierarchy of the legislation corpus. Network analysis in legislation is addressed in [20]. They also refer to the words surrounding the citation, but in this case, they seek predicates, words that express the relationship without reference to the subject-matter in question and which would make sense if applied to any other provision. They classify such edges with 9 labels: legal basis, authority, definition, example, exception, criterion, limitation, procedure and amendment. Conditional Random Fields are used to find the predicates, and k-means classification with word embeddings to automatically classify the edges.

A previous work addressed the task to automatically identify semantic relations existing among concepts based on a pattern-matching approach [21]. This analysis finally identify five kinds of semantic relations among legal terms that we considered to build our typology of relations, better described in Section 2.3. Another similar work investigates explicit cross-references to external legal texts in order to identify a taxonomy of edges in a citation graph [22]. Compared to their work, we investigate internal cross-references to portions of the same legal text. While most literature takes into account explicit citations, [23] explored implicit citations building a network of individual text units or paragraphs. Their work used a text similarity technique to draw a more complete picture of implicit citations, while in our work we focused on binary classification tasks.

### 2.2. Text analysis and machine learning

Several NLP techniques have been applied to the analysis of legal texts [24], including systems capable of making semantic connections. The growing interest of AI in Law [5] in such techniques is evidenced by recent summary monographs [25]. As a matter of fact, several practical applications typically exploit a NLP pipeline to address machine learning experiments [26,27], e.g., by performing the classification of judgment norms [28–30]. Text mining techniques have proven to be useful in extracting structured information from legal sources [31]. More recently, deep neural networks are gradually used in legal informatics to explore text classification, information extraction, and information retrieval [32]. While these kind of studies generally performs well, is recognized how standard NLP approaches may still work better than recent neural-based methods, as a very recent study states: "It is observed that the more traditional methods (such as the TF–IDF and LDA) that rely on a bag-of-words representation performs better than the more advanced context-aware methods (like BERT and Law2Vec) for computing document-level similarity" [33]. Therefore, as this work focused the attention on the role of co-occurrence networks, we preferred to adopt a standard NLP pipeline. Although there are no previous work directly similar to ours on implicit links between parts of a rule, there are several works that have addressed the automatic classification of legal texts. For instance, the system of [34] correctly forecasts 69.7% of Case Outcomes and 70.9% of Justice Level Vote Outcomes of the supreme court of the united states over the sixty year period, but they did not exploit any network features. Another recent work demonstrates the potential of machine learning approaches in the legal domain in predicting the violation of 9 articles of the European Convention on Human Rights with an average accuracy of 0.75. In particular, they exploited a normalized frequency vectorization method with words

(unigrams) or sequence of terms (bigrams or trigrams) by obtaining a prediction accuracy for various binary classification tasks (F-measure in a range from 0.62 to 0.85) [35]. Similar results are obtained in the automated classification of legal norms in German statutes with regard to their semantic type, whereas machine-learning based approaches obtained an accuracy of 0.83 [30]. Finally, although applied in relatively different tasks, we can consider these accuracy values to compare the results of our classification experiments.

### 2.3. Norms type identification

Some recent work [36,37] uses text similarity techniques for the automatic identification of implicitly linked norms, specifically between conceptually similar norms, the former between recitals and (sub-)articles in EU legislation, and the latter between (sub-)articles in EU Directives and National Implementing Measures. In [38], we consider the degree of similarity between recitals and (sub-)articles in EU legislation, (sub-)articles in EU Directives and National Implementing Measures, and EU recitals and (sub-)articles in National Implementing Measures, as well as a more extensive evaluation of different similarity measures for this task.

The classification scheme described in [39] investigates implicit links to include eight different types, based on a detailed study of a Directive. We moved from that theoretical work towards the creation of a gold standard corpus and the development of a system to automatically identify such links, by considering in the current work the following eight different types of implicit links:

**Conceptually Similar (CS)** whether using the same or different wording

**Constitutive (Co)** linking norms containing definitions of legal terms to norms containing those terms

**Motivation (Mo)** where one norm provides the principle or goal that motivates another norm

**Impact (Im)** in terms of conflicting goals that may restrict one or both of the norms

**Procedural (Pr)** linking a norm describing a procedure by an EU institution to support the goal of another norm

**Contextual (Cx)** linking deontic or other norms to norms that provide contextual information such as jurisdiction and entry into force

**Indirect Internal (II)** where norms A and C are linked indirectly (by another internal norm B)

**Via Other Law (VOL)** involving a norm that is related to another norm and cannot be understood without reference to that law

### 2.4. The case study: an EU regulation

As a practical case study, we focus on arecent EU Regulation.[1] A first motivation is that the EU legislation is a reference by a wide audience. Moreover, this legal text is interesting as it concerns healthcare, although interesting economic issues are also present. Eleven years after its adoption, it has been amended on only a few occasions and is still almost entirely in force. In addition, this Regulation was chosen because of its comprehensibility and length, i.e. long enough to provide a variety of different implicit links but short enough to be annotated by two annotators.

The here considered legal text includes 11 recitals and 11 articles. We further divided the articles into paragraphs obtaining 38 (sub)-articles. Every parts of the law (both recitals and articles) have to be related each others. Therefore, the full combination of inter-relationships between recitals and (sub)-articles explored in this work is 418. Each of them can be annotated with one label out of a total of eight, accordingly to the above-mentioned typology.

We conclude this section on the role of recitals in the EU law, by providing a practical example about the importance of identifying the relationships between parts of a norm for legal interpretation. According to [40], recitals play two main functions: first, explain the reasons for the adoption of the act; second, as an interpretative legal tool. This second function has been developed by the case law of the European Court of Justice (ECJ). Authors in [41] state that "evidence, albeit indirect, of the importance of recitals in transposition can be found in every case in which the ECJ strikes down a local provision in a transposed rule due to the influence of a recital upon the scope of the transposed operative provision". Recitals are also used in the interpretation of cases brought by individual claimants. They also assess that "recitals in EC law are not considered to have an independent legal value, but they can expand an ambiguous provision's scope". More precisely, recitals are not able to "restrict an unambiguous provision's scope, but they can be used to determine the nature of a provision, and this can have a restrictive effect". The ECJ makes use of three main methods of interpretation [42]. The first is literal interpretation, which is focused entirely on the wording of the law. The second is systematic interpretation, which refers to the context of the norm, i.e. its historical background, the legal source where it is codified, its placing within the law, etc. The third is teleological interpretation, which concerns the most suitable way to realize the purpose of the norm.

We refer here to an example from [36] of the use of recitals in systematic interpretation. Cases *Case C-162/97* and *C-344/04*, referred to recitals 14 and 15 of Regulation 261/2004 (regarding compensation to air passengers) to define the term "extraordinary circumstances" in article 5 of the Regulation. Article 5 stated that the operating air carrier shall not be obliged to pay compensation if it can prove that the cancellation is caused by extraordinary circumstances which could not have been avoided even if all reasonable measures had been taken. Examples of events that may be regarded as extraordinary circumstances in the relevant recitals included political instability, meteorological conditions, security risks and strikes. These examples were used as analogies by the Court to help determine the extent to which air carriers are exempted from paying compensation.

## 3. Methodological framework

Our methodological framework includes three analytical steps. A first part focuses on the identification of norm types for links between parts of a legal text (Section 3.1). This is a difficult task involving experts in the legal domain. We presented the methodology of the card sorting exercise and the manual annotation of a corpora to build a training set. The degree of accordance between annotators will provide us an answer to RQ1 about the goodness of the types of links between norms proposed here. A second part (Section 3.2) concerns the graph-driven feature extraction process to integrate a typical NLP feature representation of legal text. In a third part (Section 3.3) several binary classification

[1] The Regulation No 141/2000 on orphan medicinal products, where the text of the Italian version we adopted is: https://eur-lex.europa.eu/legal-content/IT/TXT/HTML/?uri=CELEX:32000R0141&from=IT.

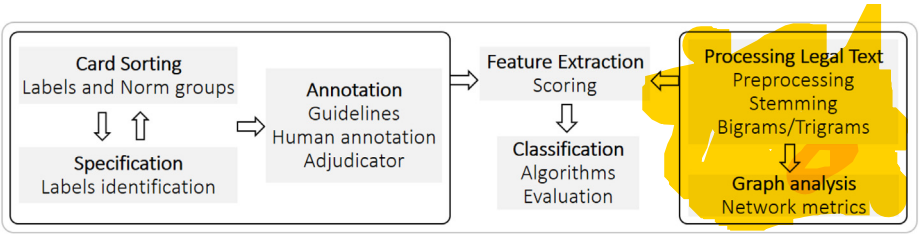


**Fig. 1.** The general methodological framework to address experiments on inter-relationships between norm types.

tasks explore the role of lexical features and network metrics in the identification of inter-relationships in a legal document. The evaluation of this automatic task will provide us an answer to RQ2 (i.e., the ability to classify these relationships) and RQ3 (i.e., the specific utility of co-occurrence network measures). Fig. 1 describes the main steps of the framework, which is better detailed in the following paragraphs.

### 3.1. Identification of norm types and the annotation process

#### 3.1.1. Training annotators: guidelines and card sorting

The creation of a corpus of annotated relationships in legal texts is a prerequisite to perform a classification task. This step relies on the effort of experts in the specific domain of interest. The annotation process involves the following steps: identification of classes or labels; definition of detailed guidelines; manual annotation of the corpus; and finally computation of inter-annotator agreement. We also carry out a card sorting activity, with the primary goal of understanding how easily identifiable different norm types are, and the secondary goal of exploring the adoption of an alternative annotation method, in a user-centered perspective. Commonly used by information architects, card sorting is a popular method aimed at the identification of patterns among data [43]. Participants, who are asked to work on their own, group physical or digital cards, each displaying a piece of information, based on their own mental model of the information domain. More specifically, while in closed card sorting participants are provided with a set of initial groups, in open card sorting they can define the groups which they feel are the most appropriate and then they have to describe each group with a label. In the case of legal text, cards can be used to display single paragraphs, the process output consisting in labeled sets of norms. Aiming at assessing the annotators' experience and, in particular, the usefulness and ease of use of card sorting in comparison with the currently adopted annotation method , we used a survey with both open-ended and 5-point Likert scale questions.

#### 3.1.2. Annotation process

The annotation process includes a scheme's definition and its application to the legal document. The aim of the scheme was to clearly define the kind of information which must be annotated. This phase includes also the inventory of labels to be used, as well as the annotation's granularity. In the practical implementation, it is very important that each annotator worked independently. This way we had evidence of any situations of disagreement, to be resolved later. The annotation task was presented in a spreadsheet file in order to facilitate the work of the annotators. Inter-annotation agreement is measured using the Cohen's kappa metric [44]. Two others annotators with a background in legal informatics were asked to solve disagreement cases. The outcome of the adjudication process was considered as a gold standard corpus obtained from the collaborative annotation process.

### 3.2. Co-occurrence networks

The analysis concerning the existence of inter-relationships between two parts of a legal text can benefit by two graph representations. In particular, we represent relations between the parts of the law connected by a common stem (i.e., the root form of a term), as well as the stems co-occurring in the same part of text. By focusing on stems, instead of words, we gain more advantage of more existing co-occurrences. We computed network metrics which can describe the role of the vertex in the graph with respect to the relationships with other vertices. Therefore, we focused the attention on the two following different kinds of undirected graphs (G1, G2).

*A graph with parts of text as vertices.* A first graph (G1) considers every parts of the law (i.e., recitals and (sub-)articles) as vertices, which are (eventually) connected by an edge if there is at least one stem in both parts. The edge weight corresponds to the number of co-occurring stems. If a stem is detected in different parts (e.g., recital 2 and article 6) of a legal text, then an edge will connect the two parts.

*A graph with stems as vertices.* A second graph (G2) explores the role of stems in the document by focusing on the relations between stems (as vertices), where the weight of the edge is the number of parts in which the two stems co-occur. For instance, an edge between Stem X and Stem Y weighted by 3 indicates that Stem X and Stem Y co-occur in a same part of the document three times.

To improve the understanding of the co-occurrence networks adopted in the study, we added a description of two sub-networks in the following paragraphs.

#### 3.2.1. Two samples of the produced graphs

Fig. 2 describes two "toy examples" of small subgraphs for the two different analysis considered in the current work, G1 and G2.

In the G1 subgraph (Fig. 2, [a]) we have four vertices well connected each others, i.e. the recitals *R5*, *R10*, *R11*, and the *Article 8.2*. The weight of the corresponding edges is the number of co-occurrent stems. We focus here on the edge from *Article 8.2* to *Recital 10*. To clarify the steps of graph construction, the following paragraphs highlight in bold the co-occurrent stems (in Italian) of both *Article 8.2* and *Recital 10*. Stems of recital 10 are:

> programm specif biomed quart programm quadr ricerc svilupp tecnolog 1994–1998 sovvenzion ricerc terap malatt rar metodolog istitu programm celer svilupp medicinal orfan inventar medicinal orfan **dispon** europ tal fond intes promuov collabor internazional mater ricerc bas ricerc clinic malatt rar comun continu attribu ricerc malatt rar import prioritar previst **quint** programm quadr ricerc svilupp tecnolog 1998–2002 present regol defin quadr giurid consent tempest effett applic **risult** tal ricerc

Stems of article 8.2 are:

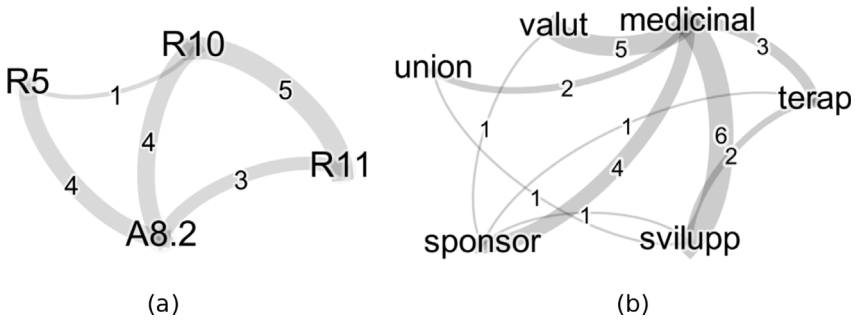


**Fig. 2.** Two sub-graphs for the co-occurrence networks of stems. On the left an example of G1 (identified with the letter 'a') and an example for G2 on the right (i.e., 'b').

tal period può tuttav esser ridott anni scadenz **quint** anno **risult** medicinal question conform criter articol **risult** fra altro bas dat **dispon** rend tal giustific manten esclus merc tal fin stat membr inform agenz criter bas concess esclus merc potrebb esser rispett segu ciò agenz avvi procedur defin articol sponsor forn agenz inform necessar riguard

We clearly notice how the two parts of the law here considered have three co-occurring stems: *risult*, *quint*, and *dispon*. We also notice that *risult* appears twice. As we compute the weight of the edge as the sum of all the co-occurrences, the final weight between *A8.2* and *R10* is *4*.

In the G2 subgraph (Fig. 2, [b]), we focus on the link between "medicinal" and "sponsor". In our pipeline, we started from the text of the Law (in Italian) to turn it into the corresponding stems.

To have a clear insight, we report here the stems of article 5.2, to finally look for the co-occurring stems. The above-mentioned pair of stems is here highlighted in bold:

**sponsor** domand dev esser corred inform document seguent nom ragion social indirizz permanent **sponsor** princip attiv **medicinal** indic terapeut propost giustif relat osserv criter articol paragraf nonc descrizion stat svilupp compres indic previst

We perform the same operation to all the parts of the law (i.e., all recitals and all articles), summing up the number of co-occurrences of both *medicinal* and *sponsor*. We notice that this pair both occur four times in the following parts of the norm: R9, A5.2, A5.12, A6.1. Accordingly, the edge between the vertices *medicinal* and *sponsor* has a weight of 4.

### 3.3. Classification

The classification problem typically involves statistical text categorization to learn automatic rules based on human-labeled training documents. We consider distributed words representation [45], as a standard technique to represent each part of the legal text using vectors. A convenient text representation for each part (i.e., recitals and (sub-)articles) in a legal document $d$ encodes the presence of words (unigrams) or sequence of words (n-grams) as $x_{(i)}$. The document as a whole is represented as a feature vector of length $p$, $x = (x_{(1)}, \ldots, x_{(p)})$. The BoN approach is a standard method typically used with the Term Frequency–Inverse Document Frequency (TF–IDF) weighting scheme [46], where TF($i$, $d$) (term frequency) is the number of times a term $i$ occurs in document $d$. To reduce dimensionality we selected the features having the highest document frequency as a feature selection method. Our classification concerns both a binary problem, i.e. the existence or non-existence of a relation between two parts (a pair) of a legal document, and the identification of the specific class type (with eight classes). The last "multi-class" problem is a difficult task for machine-learning approaches, therefore we performed binary classification tasks for each category. In binary classification tasks the cases of existence of a relationship between two parts of the legal text are considered as positive examples. We adopted a training sample of manually classified pairs $z$ denoted as $S = (z_1, y_1), \ldots, (z_n, y_n)$ where $n$ is the number of training pairs, and $y_i \in \{1, 0\}$ indicates the class label. Finally, our binary classification problem adopts the training sample in a supervised learning experimental setting. The goal is the identification of a classification rule, i.e., a function mapping from the $p$-dimensional feature space to the one-dimensional class label.

#### 3.3.1. Feature representation of legal text

This phase involves the vectorization of each part of the legal text by automatically extracting a set of features to be used in machine learning experiments. Documents are preprocessed including conversion to lowercase characters, the removal of punctuation marks and stop-words, and tokenisation (to separate terms into items). Then we opted for a Bag-of-Ngrams (BoN) model by considering stems, which is more informative than Bag-of-Words, as they capture more context around each item. For instance, it will be possible to compare singular and plural forms of the same term occurring in different parts of the legal text. To provide more importance to the rare stems that were more prominent in the text under consideration than other texts, we performed TF–IDF. In this way, we counted the stem ngrams in that document scaled down by the count of the documents that have that stem ngram. Then we considered both bigrams and trigrams of stems, created from pairs and triplets of stems appearing in sequence. Finally, we considered BoN containing information on the more important stems. In our experiments, we adopted the top 200 ones for bigrams and trigrams (henceforth, we will refer to Bigr200 and Trig200 features sets). In a preliminary exploration, we also considered a larger number of stems (e.g., top 1000 for bigrams), but the results did not improve, so we opted to focus on Bigr200 and Trig200.

#### 3.3.2. Network metrics

We adopted several network metrics summarized in Table 1 with a brief definition and the corresponding acronym, finally grouped in the following three sets (NM1, NM2, NM3). Two sets of metrics are based on G1. A first set (NM1) includes network metrics at the vertex level, i.e., Degr, DegC, BetC, CloC, EigC, LoaC, ClCo, Cons. A second set (NM2) concerns specifically edge-related metrics of the same graph, i.e., a boolean feature *isLink* to check an edge between two vertices (1 in case of a link, otherwise 0), *Weig* (the weight of the edge), in addition to EdBC, CFlo, Effi, Conn, Cliq, kCli, Jacc. A third set of metrics (NM3) is based on G2. We computed for each part of the legal text the average values of the corresponding stem-related graph-based features.

**Table 1**
Description of network metrics and their respective acronyms in brackets.

| Name | Description |
|---|---|
| Degree | The number of edges that are incident to a vertex (Degr) |
| Clustering coefficient | The fraction of possible triangles through that vertex (Clus) [47] |
| Centrality | Centrality metrics [48] address the position of the vertex in a graph by considering different perspectives, e.g. the neighbors of the vertex: Degree centrality (DegC); the importance of the position in the graph: Betweenness centrality (BetC); the distances from other vertices: Closeness centrality (CloC); the influence of the vertex in the network: Eigenvector centrality (EigC); Current-Flow betweenness centrality (CFlo) [49]. By focusing on edges, we consider also edge betweenness centrality (EdBe). |
| Constraint | A measure of the extent to which a vertex is invested in those vertices that are themselves invested in the neighbors of the vertex (Cons) [50] |
| Clique | A clique is a subset of vertices of an undirected graph such that every two distinct vertices are adjacent. We distinguish vertices and edges of the maximal clique (Cliq), as well as k-clique communities (kCli) [51] |
| Jaccard coefficient | Compute the Jaccard similarity index (Jacc) between all pairs of vertices [52] |
| Efficiency | The efficiency of a pair of vertices is the multiplicative inverse of the shortest path distance between the vertices (Effi) [53] |
| Connectivity | Returns local edge connectivity as the minimum number of edges that must be removed to disconnect them (Conn) [54]. |

**Table 2**
Sets of network metrics (NM1, NM2, NM3) from two kinds of graphs G1, G2.

| Name | Graph | Network metrics |
|---|---|---|
| NM1 | G1 | Degr, DegC, BetC, CloC, EigC, LoaC, ClCo, Cons |
| NM2 | G1 | isLink, Weig, EdBC, CFlo, Effi, Conn, Cliq, kCli, Jacc |
| NM3 | G2 | avDegC, avBetC, avCloC, avEigC, avLoaC, avClCo, avCons |

Therefore, we considered the averaged centrality metrics for all the stems included in the same part of the law, i.e., Degree (avDegC), Betweenness (avBetC), Closeness (avCloC), Eigenvector (avEigC), and Load centrality (avLoaC). Finally, we included also the average value of two relevant vertices properties such as Clustering Coefficient (avClCo), and Constraint (avCons). Table 2 summarizes the three sets of metrics from the two graphs.

*3.3.3. Experimental settings*

The combination of the above-mentioned methodological steps allows us to perform a set of supervised binary classification experiments. First, we are interested in considering the existence of at least one relationship between two parts of the legal text. Second, we focus on the prediction of the existence of an individual label. Each part of our legal text is consequently represented as a concatenation of features obtained by both a BoN model and a graph-based analysis. Several classification algorithms are explored, i.e. Logistic Regression (LR), Decision Tree (DT), Support Vector Machines (SVM), k-Nearest Neighbors (kNN) [55]. In addition, we explore Naive Bayes (NB), as well as a Dummy (DU) classifier that makes predictions using simple rules.

*Validation and evaluation*. We opted to consider standard k-fold cross-validation to test the effectiveness of our models. This is a well-known re-sampling procedure used to evaluate a model, by keeping aside a portion of the data which is not used to train the model. As a standard way of the predicting error rate of a classifier, we performed ten-fold cross-validation. In particular, the process repeats ten times the following steps: (i) break the training data into 10 equally-sized partitions; (ii) apply the classification algorithm on nine parts, while testing on the remaining folds. The final measure is the performance average of the ten parts. The classifiers can be evaluated by computing Accuracy as well as F-measure, which provides information by combining the ratio between precision and recall measures. Accuracy is the ratio of correctly predicted observation to the total observations. Precision is the ratio of correctly predicted positive observations to the total predicted positive observations. Recall is the ratio of correctly predicted positive observations to all observations in the actual class (i.e., the existence of a relationship). As we had an uneven class distribution, we mainly considered F-measure, i.e., the weighted average of precision and recall measures.

The computation performed in this work used the Python programming language and related libraries, e.g. NLTK [56], Scikit-learn [57], and NetworkX [58]. We have also performed a graph analysis using the open-source software Gephi [59].

## 4. The annotation process

This section describes the process of annotating links between pairs of norms and norm groups, as well as an analysis of the annotation results and post-discussion feedback.

### *4.1. Annotation results*

Table 3 summarizes the main results from the annotation phase. By looking at raw values, the agreement seems to be quite low, mostly due to annotators' attitudes: Annotator 2 opted to label 617 relationships, compared to only 260 by Annotator 1. Consequently, there are significant numeric differences in some classes. Nevertheless, a similar proportion is observed for each individual class, especially for some of the most frequent classes: Motivation, Via Other Law and Procedural.

The cases of initial agreement are quite encouraging: at least, the two annotators agree on the existence of 141 relationships, and the absence of a relationship in 2608 cases. The remaining cases of disagreement (595) were resolved by two other adjudicators. The final annotated corpus after resolution of disagreement contains 464 links encompassing all the eight classes.

There are: 139 cases where the annotators agree on the existence of a relationship; 99 cases where both annotators agree on the absence of a relationship. On the 180 cases of disagreement, in 22 cases only the first annotator considered that a relationship exists; and in 6 cases only the second annotator considers that a relationship exists. This provides a percentage of agreement of 57%, and a Cohen's kappa agreement of 0.2 on the existence of a relationship.

On the identification of a particular kind of link, however, the percentage of agreement is generally higher, ranging from 46.6% to 99.0%. This is because of high agreement on pairs of norms that do not have a link of a particular class. The Cohen's kappa values, on the other hand, never rises above moderate.

The data shows that agreement is highest for classes with clear objective criteria and identification involving formulaic language — Contextual and Constitutional. While the Via Other Law class also involves identification of citations via formulaic language, the lower percentage of agreement is explained by the fact that one annotator applied this class only when it was truly necessary to refer to the cited law in order to understand one of the norms under consideration, while the other applied the class systematically whenever one of the norms contained a reference to another law.

**Table 3**
Annotation results from two initial annotators (Annotator 1 and Annotator 2), with cases of initial agreement and disagreement; the final annotated corpus after the resolution of disagreement (by an adjudicator).

| | CS | Co | Mo | Im | Pr | Cx | II | VOL | Total |
|---|---|---|---|---|---|---|---|---|---|
| Annotator 1 | 15 | 16 | 81 | 38 | 53 | 11 | 11 | 35 | 260 |
| Annotator 2 | 12 | 23 | 254 | 6 | 122 | 13 | 56 | 131 | 617 |
| %Annotator 1 | 5.8 | 6.2 | 31.2 | 14.6 | 20.4 | 4.2 | 4.2 | 13.5 | 100 |
| %Annotator 2 | 1.9 | 3.7 | 41.2 | 1.0 | 19.8 | 2.1 | 9.1 | 21.2 | 100 |
| Agreement YES | 3 | 10 | 56 | 0 | 30 | 10 | 6 | 26 | 141 |
| Agreement NO | 394 | 389 | 139 | 374 | 273 | 404 | 357 | 278 | 2608 |
| Disagreement YES | 21 | 19 | 223 | 44 | 115 | 4 | 55 | 114 | 595 |
| YES — Annotator 1 only | 12 | 6 | 25 | 38 | 23 | 1 | 5 | 9 | 119 |
| YES — Annotator 2 only | 9 | 13 | 198 | 6 | 92 | 3 | 50 | 105 | 476 |
| Percentage of agreement | 95.0 | 95.5 | 46.6 | 89.5 | 72.5 | 99.0 | 86.8 | 72.7 | 82.2 |
| Cohen's kappa | 0.2 | 0.49 | 0.06 | n/a | 0.20 | 0.82 | 0.14 | 0.20 | 0.24 |
| Gold standard corpus | 9 | 18 | 238 | 3 | 57 | 11 | 55 | 73 | 464 |

We have relatively modest results for the Conceptually Similar and Motivation class, which are clearly defined classes but more 'subjective' in nature.

On the other hand, the poorer results for Impact and Procedural were largely due to differences between the directive on which the classification scheme was modeled and the regulation that was annotated. While both legislation were from the health domain, norms for Directive 2004/23/EC mention a range of actors including member states, their competent authorities, donors, tissue establishments and their personnel, whereas Regulation No 141/2000 mainly focused on the required activities of the European Agency for the Evaluation of Medicinal Products and Committee for Orphan Medicinal Products. In both legislation, norms involving EU institutions such as the European Commission, Parliament, Council, and specialist Committees were clearly subject to Procedural, and not Deontic norms (i.e., such norms were descriptions of what the aforementioned institutions will do, rather than prescriptions of what they should do), and therefore subject to Procedural links with related norms. On the other hand, post-annotation discussion with the annotators revealed that while the European Agency for the Evaluation of Medicinal Products and Committee for Orphan Medicinal Products were considered by both to be part of the EU rather than autonomous bodies, the norms involving them could be considered to be Deontic, Procedural or both, depending on the content of the norm itself rather than purely on its addressee. This means that the Procedural class as applied to the Regulation was subject to greater degree of subjectivity than expected.

The worst results are for the Impact class where the annotators did not agree on a single pair. Here, the difference between Directive 2004/23/EC and Regulation No 141/2000 resulted also in different interpretations on the meaning of the class. While the Directive contained sub-goals with the potential to conflict (transparency versus confidentiality), thus requiring balancing or defeasibility, such potential conflicts were not so obvious in the Regulation annotated for this article. The other type of Impact – linking Deontic norms to norms specifying enforceability measures – were not present in the same way in the Regulation. One annotator extended the Impact category to include the impact of planned future guidelines, in consultation with member states and other parties, on the interpretation and efficacy of the stated goals of the regulation. However, this interpretation was not shared by the other annotator.

From all this, we have a clear indication of where greater effort is required in fine-tuning or even redefining certain classes, and the need for annotation of a greater range of legislation by more annotators.

*Annotation results discussion.* Post-annotation discussion revealed that both annotators found the use of the term Procedural for both a norm type and link type to be confusing, and suggested renaming one of these classes. Otherwise, they viewed the definition of link types provided in the guidelines as highly coherent, with a post-annotation questionnaire providing a rating of 4.5 for this measure. However, one annotator stated that while the definitions were pretty exhaustive, indecision could arise because the content of the norm could be interpreted in more than one way. The annotators had different views on whether a training session could be useful before annotation in the future, with one providing a rating of 3 out of 5 and the other 5 out of 5. The first annotator was of the view that any training seminar should not include precise explanations of the classes, as it is essential that each annotator carries out the activity based on his/her understanding of the class. The other annotator stated that the definitions are often clear in theory, but questions arose when reading the norms, and so a training seminar with an intermediary featuring practical exercises with sample norms could be useful to ensure that the definitions of classes has been interpreted by all in the same way.

### 4.2. Manual identification and classification of norm groups

A preliminary card sorting activity was carried out by the two legal annotators. They were assigned two tasks: firstly, they were asked to group norms into six categories corresponding to:

- Objective (Ob): outlines the purpose behind the directive as a whole, or some parts of it
- Constitutive (Cns): containing definitions of technical concepts
- Deontic (De): specifying types of behavior to be expected or permitted. May be further classified as permission, obligation, prohibition etc.
- Scope (Sc): outlines the extent of applicability or non-applicability of norms (or entire legislation)
- Procedural (meta-norm) (Pr): refers to step-by-step processes for implementing legislation e.g. get signatures, agreement from the Committee, further signatures.
- Contextual (meta-norm) (Cnt): refers to time, space, addressee and hierarchy of norms

Secondly, annotators were asked to sort norms into as many groups as they preferred and to identify a suitable label for each of them, using the *norm group* concept discussed in [39] as a guidance: "A link between norms that are connected due to being part of the same general requirement. The links between the norms may be conjunction, disjunction or sequence. Such norms may be paragraphs of the same article, or may occur in different provisions". It is envisaged that the identification of such norm

**Table 4**
Closed card sorting results from two annotators (Annotator 1 and Annotator 2), with cases of agreement and disagreement.

| | Ob | Cns | De | Sc | Pr | Cnt | Total |
|---|---|---|---|---|---|---|---|
| Annotator 1 | 5 | 11 | 31 | 2 | 0 | 1 | 50 |
| Annotator 2 | 11 | 3 | 16 | 0 | 34 | 1 | 65 |
| Both | 16 | 14 | 47 | 2 | 34 | 2 | 115 |
| %Annotator 1 | 10 | 22 | 62 | 4 | 0 | 2 | 100 |
| %Annotator 2 | 16.92 | 4.62 | 24.62 | 0 | 52.31 | 1.54 | 100 |
| %Both | 13.91 | 12.17 | 40.87 | 1.74 | 29.57 | 1.74 | 100 |
| Agreement YES | 5 | 2 | 12 | 0 | 0 | 1 | 20 |
| Agreement NO | 38 | 37 | 14 | 47 | 15 | 48 | 199 |
| Disagreement YES | 6 | 10 | 23 | 2 | 34 | 0 | 75 |
| YES — Annotator 1 only | 0 | 9 | 19 | 2 | 0 | 0 | 30 |
| YES — Annotator 2 only | 6 | 1 | 4 | 0 | 34 | 0 | 45 |
| Percentage of agreement | 87.76 | 79.59 | 53.06 | 95.92 | 30.61 | 100 | 74.49 |
| Cohen's kappa | 0.56 | 0.21 | 0.14 | n/a | n/a | 1 | 0.19 |

groups may not only be useful in itself, but may also help to improve the identification and classification of links between norms.

The methodology employed for this exercise was card sorting (first task: closed; second task: open [43]). Each annotator was provided with a set of physical cards, where each card was used to display a single paragraph, and multiple copies of the same card were available in order to allow for the allocation of cards to multiple groups. Forty-nine different cards, corresponding to 11 recitals and 38 articles, were included in each set.

Notice that, while this activity is normally carried out with 15–20 participants [60], we decided to conduct a preliminary card sorting activity with only two annotators in order to obtain some feedback on the difficulty of the task and the viability of card sorting as an alternative annotation method when labels must be applied to sets of items.

#### 4.2.1. Card sorting results

*Closed card sorting.* Table 4 summarizes the main results from the closed card sorting activity. One hundred and fifteen card-category (or recital/article-norm type) matches were identified, with 16 cards out of 49 (33%) assigned to multiple categories by the same annotator. Annotators agreed on 20 matches: of these, 8 referred to cards which were placed in a single category by both participants and 2 to a card which was placed in the same two categories, meaning that perfect agreement could be observed for 9 cards out of 49 (18%). On the whole, Cohen's kappa (0.19) indicates slight agreement.

The percentage of cards assigned to each category was highly variable and differed for the two annotators. The highest levels of agreement were observed for two categories, *meta-norms: contextual*, which had exactly the same composition for both participants (Cohen's kappa: 1), and *objective*, where annotators agreed on about half the cards assigned to the group (Cohen's kappa: 0.56). On the contrary, no overlap could be found for *scope* and *meta-norms: procedural* categories (Cohen's kappa: n.a.), coherently with our link pairs annotation results (see Section 4.2.1).

*Open card sorting.* Table 5 summarizes the main results from the open card sorting activity. Annotator identified 6 and 16 groups, with an average of 10 and 7 cards per group, respectively. The grouping structure and proposed labels reflected quite different mental models: in fact, only in two cases the annotators used the same or very similar labels (namely, *committee*, Cohen's kappa: 0.40, and *market exclusivity/monopoly*, Cohen's kappa: 0.55, see Table 6). On the whole, annotators agreed only on 7 card-norm group (or recital/article-norm group) matches.

Forty-one cards out of 49 (84%) were assigned to multiple categories by the same annotator, thus suggesting a higher overlap between spontaneously identified norm groups in comparison with norm types. One of the annotators stated that she purposely identified various groups and subgroups, due to her hierarchical mental model of norm groups.

**Table 5**
Overall open card sorting results from two annotators (Annotator 1 and Annotator 2).

| | Annotator 1 | Annotator 2 | %Annotator 1 | %Annotator 2 |
|---|---|---|---|---|
| #1 | 2 | n/a | 3.51 | n/a |
| #2 | 13 | 4 | 22.81 | 3.45 |
| #3 | 7 | n/a | 12.28 | n/a |
| #4 | 5 | 5 | 8.77 | 4.31 |
| #5 | 18 | n/a | 31.58 | n/a |
| #6 | 12 | n/a | 21.05 | n/a |
| #7 | n/a | 17 | n/a | 14.66 |
| #8 | n/a | 3 | n/a | 2.59 |
| #9 | n/a | 6 | n/a | 5.17 |
| #10 | n/a | 3 | n/a | 2.59 |
| #11 | n/a | 6 | n/a | 5.17 |
| #12 | n/a | 2 | n/a | 1.72 |
| #13 | n/a | 13 | n/a | 11.21 |
| #14 | n/a | 9 | n/a | 7.76 |
| #15 | n/a | 20 | n/a | 17.24 |
| #16 | n/a | 2 | n/a | 1.72 |
| #17 | n/a | 11 | n/a | 9.48 |
| #18 | n/a | 4 | n/a | 3.45 |
| #19 | n/a | 9 | n/a | 7.76 |
| #20 | n/a | 2 | n/a | 1.72 |
| Total | 57 | 116 | 100 | 100 |

**Table 6**
Open card sorting results for common groups from two annotators (Annotator 1 and Annotator 2), with cases of agreement and disagreement.

| | #2: committee | #4: market exclusivity/monopoly | Total |
|---|---|---|---|
| Annotator 1 | 13 | 5 | 18 |
| Annotator 2 | 4 | 5 | 9 |
| Both | 17 | 10 | 27 |
| %Annotator 1 | 22.81 | 8.77 | 31.58 |
| %Annotator 2 | 3.45 | 4.31 | 7.76 |
| %Both | 9.83 | 5.78 | 15,61 |
| Agreement YES | 4 | 3 | 7 |
| Agreement NO | 36 | 42 | n/a |
| Disagreement YES | 9 | 4 | n/a |
| YES — Annotator 1 only | 9 | 2 | 11 |
| YES — Annotator 2 only | 0 | 2 | 2 |
| Percentage of agreement | 81.63 | 91.84 | n/a |
| Cohen's kappa | 0.40 | 0.55 | n/a |

*Card sorting results discussion.* With regards to our goal of assessing how easily identifiable norm types are, the closed card sorting activity highlighted that some concepts, most notably the *scope* and *meta-norms: procedural* norm types, are understood differently by the annotators. If confirmed by more annotators, these results might lead to an update in our guidelines or list of norm types.

Secondly, open card sorting proved to be suitable for annotation tasks where a hierarchical structure among the identified concepts can be expected, and possible labels cannot be defined *apriori*. However, if there is a need to define a "gold standard", it should be noted that handling disagreement between participants can be more complex, as not only item–group, but also group–group relationships have to be taken into account.

Finally, as far as user experience is concerned, our survey shows[2] that card sorting was judged slightly less *useful* and *easy to use* in comparison with the spreadsheet-based annotation

[2] We are aware that, due to the small number of respondents, our results have no statistical significance. However, they can still provide some interesting insights on possible usability and methodological issues.

**Table 7**
Network metrics of G1 and G2 graphs.

| Name | #V | #E | AvDegr | Dia | AvPaLe | Den | Tra | AvClCo |
|---|---|---|---|---|---|---|---|---|
| G1 | 49 | 320 | 13.1 | 3 | 1.729 | 0.272 | 0.394 | 0.645 |
| G2 | 548 | 1279 | 4.7 | 7 | 3.248 | 0.009 | 0.069 | 0.132 |

method (3.5 vs. 4.5 out of 5 for both measures). In particular, one of the annotators pointed out that card sorting can be more chaotic. However, she also appreciated the ease of working with physical cards, and suggested to provide annotators with a printed copy of the overall legal text, in order for them to have an overview of the content they have to work with.

Based on our results, we envisage that card sorting can provide several benefits, if included in our framework:

- It can be used as a preliminary step to assess whether there is a common understanding of labels to be used in subsequent annotation activities (corrective actions should be applied when needed).
- If applied to the definition of norm groups, it can be used as a preliminary "training" step to help annotators familiarize with the legal text and improve their ability to identify and classify links between norms.
- Provided that the overall user experience is improved, it can be used as an alternative annotation method.

## 5. Graph analysis

The graph representation allows to investigate structural links in the document conveyed by lexical links. To describe the two different graphs exploited in this work, we use several network metrics. In addition to the number of vertices (#V) and edges (#E), we also consider the following metrics: average degree (AvDegr), diameter (Dia), average path length (AvPaLe), density (Den), transitivity (Tra), average clustering coefficient (AvClCo). Table 7 summarizes the main characteristics of both G1 and G2. Below we introduce the specific set of network features extracted by the two graphs.

*Graph of Recitals and Articles*. To provide an idea about the resulting graphs, we propose a representation of G1 in Fig. 3, describing the parts of the text (recitals and (sub-)articles) as vertices, connected to each other where they share at least a common stem. The sizes of vertices are proportional to their degree, while each vertex class size depends on their betweenness centrality. Edges are proportional to their weight, i.e. the number of stems co-occurring between the two corresponding vertices. These values can represent the strength of the relationship between the two parts of the law. We perform community detection to further investigate the existence of more cohesive groups by applying the Louvain modularity algorithm [61]. This algorithm identifies four groups, having different colors in the graph[3] of 18, 14, 12 and 5 vertices. Further analysis is outside the scope of this work, but could lead to interesting considerations about the type of relationships.

The network metrics provide some insights about the graph topologies. G1 is a graph that is relatively compact, having a quite fair density (about 0.3). In fact, it is quite easy to reach most other vertices in a limited number of connections: the average shortest path length is 1.7, and the diameter of the whole network is 3. Coherently, the average degree appears of a certain significance: each part of text is connected to 13 other vertices on average.

[3] We provide a link to a github repository with the code and the colored graph: https://github.com/sulem76/LegNet.

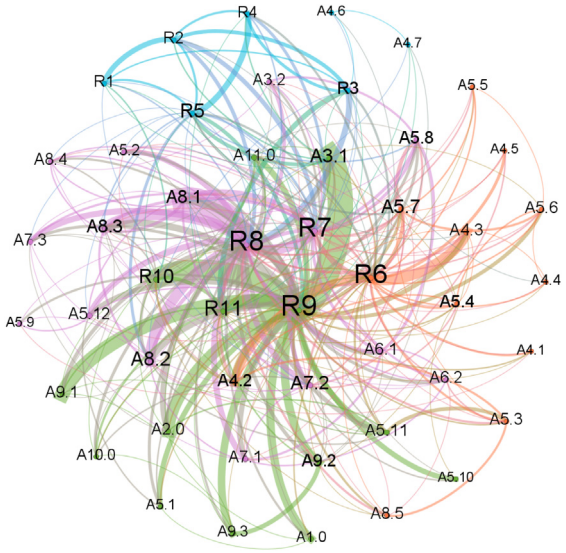


**Fig. 3.** A representation of graph G1 concerning relations between recitals and (sub-)articles. The vertices are recitals (R) and (sub-)articles (A), where edges represent a link if two vertices have at least one common stem (the weight sum of all the co-occurring stems in the two parts of the document). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Values of transitivity and clustering coefficient indicates an interesting connectivity, i.e. a certain local neighborhood, where a region is connected to its neighbors.

*Graph of Stems*. G2 is quite a sparse graph, as expected concerning the links between stems. Density is low, and the average degree indicates how each of the 548 stems (corresponding to the vertices of our graph) is connected on average to less than 5 other stems. Nevertheless, the graph is not so wide, having a diameter of 7, and the average path length is slightly higher than 3. While G1 exhibits the small-world network property (having both a high clustering coefficient and low average path length), G2 metrics indicate how there are few connections in the stems neighborhood. By considering G2, each vertex (i.e., stem) metric can be used to create a feature related to the corresponding parts of the legal text (e.g., articles or recitals). In fact, the average value of vertices' metrics (e.g. degree, betweenness centrality, closeness centrality) computed for each stem can be exploited in the classification step.

## 6. Classification output

The existence of a relationship between two parts of the legal document is explored as a supervised machine learning experiment carried out by models trained on the basis of the annotation results. A first set of experiments (S1) focuses on the existence of at least one type of inter-relationship (no matter of what kind) between two parts of the legal text. In a second set of experiments, we consider the ability to predict distinct individual labels (S2). Finally, a third set of experiments investigates the different configurations of the network-based features (S3) considered here, as our interest mainly concerns exploring the role of the network links between parts of text.

### 6.1. S1: Predicting inter-relationships

The results for the first experiments indicate that our full set of features is able to predict with a certain degree of accuracy (F-measure is about 0.84 with both LR and SVM) the existence

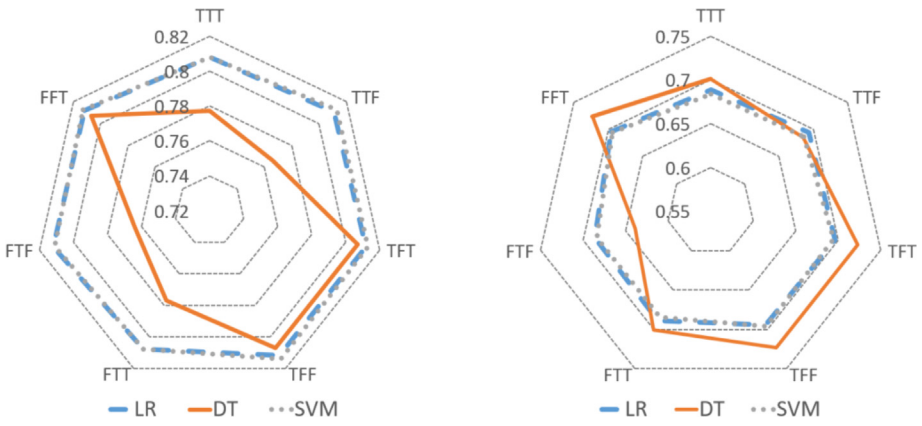


**Fig. 4.** F-measure (radar-plot on the left) and Accuracy (right) for Logistic Regression, Decision Tree and Support Vector Machine classification of network metrics configurations.

**Table 8**
Classification results for the existence of inter-relationship between two norm parts in each features set (Bigr200 and Trig200).

| Set | Performance | LR | DT | SVM | kNN | NB | DU |
|---|---|---|---|---|---|---|---|
| Bigr200 | F-measure | **0.838** | 0.792 | **0.838** | 0.752 | 0.488 | 0.656 |
| | Precision | 0.814 | 0.806 | 0.775 | 0.825 | **0.921** | 0.666 |
| | Recall | 0.869 | 0.781 | **0.914** | 0.695 | 0.334 | 0.709 |
| | Accuracy | **0.768** | 0.720 | 0.758 | 0.687 | 0.522 | 0.579 |
| Trig200 | F-measure | 0.840 | 0.775 | **0.843** | 0.783 | 0.505 | 0.699 |
| | Precision | 0.806 | 0.783 | 0.777 | 0.819 | **0.949** | 0.680 |
| | Recall | 0.879 | 0.770 | **0.924** | 0.731 | 0.349 | 0.652 |
| | Accuracy | **0.768** | 0.696 | 0.765 | 0.701 | 0.538 | 0.57 |

**Table 9**
Classification results: F-measure values (average and standard deviation) on the existence of an inter-relationship for each link norm type by adopting Bigr200.

| Norm types | LR | DT | SVM |
|---|---|---|---|
| MO | 0.753 (0.057) | 0.715 (0.079) | **0.775 (0.04)** |
| II | **0.789 (0.159)** | 0.760 (0.175) | 0.550 (0.146) |
| VOL | 0.753 (0.094) | **0.755 (0.139)** | 0.686 (0.118) |
| PR | **0.564 (0.124)** | 0.535 (0.143) | 0.522 (0.168) |

of a relationship between two parts of legal text. Our two feature sets (Bigr200 and Trig200) obtained very similar results (trigrams slightly better than bigrams in 4 classifiers out of 5). Table 8 describes the output of different classification algorithms in terms of F-measure and Accuracy. The results are quite satisfactory if we consider both the complexity of the task and the output from the Dummy classifier.

### 6.2. S2: Predicting link norm types

The prediction of individual classes is a very difficult task for several reasons. First, link norm types convey a more precise meaning, not easy to be captured by lexical features or relationships. Second, some link norm types are quite rare, so will require significant annotation effort to derive a meaningful golden standard corpus to obtain results of interest. Table 9 describes binary classification results concerning a subset of the most promising algorithms (LR, DT, SVM) by adopting bigrams (Bigr200), as the output with trigrams is very similar. We consider F-measure of the four more frequent norm link types in our annotated corpus: Mo, II, VOL, and Pr. The results show a certain possibility for classifiers to explore the difficult topic of identifying relationships of a specific type.

**Table 10**
Classification results to predict the existence of inter-relationship between two norm parts only with Bag-of-bigrams model (left column) and only with network features (right column).

| Performance | Only Bag-of-bigrams | | | | Only network features | | | |
|---|---|---|---|---|---|---|---|---|
| | LR | DT | SVM | kNN | LR | DT | SVM | kNN |
| F-measure | 0.828 | 0.828 | **0.832** | 0.726 | **0.808** | 0.777 | **0.808** | 0.757 |
| Precision | 0.809 | 0.821 | 0.776 | **0.840** | 0.698 | **0.798** | 0.691 | 0.71 |
| Recall | 0.852 | 0.839 | **0.900** | 0.650 | 0.962 | 0.760 | **0.979** | 0.819 |
| Accuracy | 0.756 | **0.763** | 0.751 | 0.670 | 0.689 | **0.701** | 0.684 | 0.644 |

### 6.3. S3: The role of network features

This subsection investigates whether the adoption of network metrics alone can be helpful, without the addition of a BoN representation. The results described in Table 10 indicate how network metrics alone obtain an F-measure of about 0.81 from two classifiers. These measures are lower but not too far from the values obtained by only adopting ngrams-based classification (about 0.83). Moreover, the results show how helpful it is to add the network metrics to a BoN model to improve the classification results (about 0.84, as detailed in Section 6.1). Although the improvement is quite slight, it seems significant that it occurs in almost all the classifiers considered in this study. This practical evidence prompts an investigation of this type of analysis.

To shed some light on network features, we explored the classification results by focusing on different sub-groups of network metrics. In particular, we considered the inclusion or the exclusion of some network metrics presented in Table 2 to assess the specific impact on the classification task. In this sort of feature ablation experiment, the considered sub-groups are NM1, NM2, and NM3, as previously introduced in Table 2. Every configuration may include a sub-group (henceforth T, which stands for "true") or its absence (F). For instance, TTT means that all the three groups are considered, TTF means that only NM1 and NM2 are considered, TFT includes only NM1 and NM3, and so on. Finally, we present the output for Bigr200, as the results for Trig200 are similar. The results concerning this specific task are described in Fig. 4. In particular, we notice two notably regularities: i. Best F-measure performances regard LR and SVM, while best Accuracy performance involves DT; ii. the configurations providing the best results are FFT, TFT, and TFF. Finally, this analysis suggests we should focus on the edge-based metrics (NM2), which is always present in the best performance groups.

## 7. Conclusions and future work

This work describes a general NLP framework to automatically classify implicit links between parts of a norm. The links for each type of relationship were obtained by a complete annotation process carried out by two legal annotators and resolution of disagreement by two researchers in legal informatics. Firstly, the annotators identified and classified one-to-one links between parts of a law in an European Directive. The results of the link annotation exercise were reviewed in order to arrive at a small 'gold standard' corpus, which was used to develop an automatic identification and classification system. We answered to RQ1 by including in the pipeline the 'card sorting' method, where the annotators grouped norms according to type and semantic relatedness. The exercise was useful to gain some insight on the link pairs annotation, as well as to refine both norm and link type classes. As discussed, while the adopted link classes were viewed by both annotators as coherent and well-described in the annotation guidelines, some differences in interpretation have emerged, e.g., some difficulties in the interpretation of the Impact and Procedural link classes. Otherwise, even for well-understood classes such as Conceptually Similar and Motivation, doubts and disagreement occurred due to the subjectivity involved in applying these classes to norms, partly because the norms themselves contained a variety of content, and partly because the norms could be interpreted in different ways.

After the creation of a labeled corpus, we developed a machine learning experiment using a graph-based NLP approach. The classification results are promising and suggest that implicit links as well as link classes can be automatically identified using typical algorithms, by answering our RQ2. In addition to a BoN model based on the root form of terms, we demonstrated how text features derived by CNs metrics can help improve the identification of some types of implicit links between norms. Responding to RQ3, in term of accuracy we noticed a slight accuracy improvement in almost all the classifiers considered here. For instance, our best classifier (SVM) improves in terms of F-measure by considering the full set of features (0.838), instead of 0.832 with BoN alone or network features alone (0.808). Moreover, we registered the promising role of the edge-related features (NM2), to be further investigated in future work.

Finally, we noticed how our classification results largely aligns with related works, as well as the results for inter-annotator disagreement to reflect the difficulty of the task. Nevertheless, in a future work we aim to improve the classification scheme which should will result in improvements in the results of classification algorithms. While the subjective nature of even well-defined and represented classes will always leave room for disagreement, we believe that even an imperfect automated system would provide significant productivity gains in the legal practitioner's search for implicitly related norms.

Another interesting perspective is to extend the work described with a qualitative analysis. While this article brings some insights into the feasibility of identifying different kinds of links, it is also important to know which kinds of links are most important to lawyers, in order to determine where to prioritize the effort. We will seek such feedback from practising legal professionals, e.g., via a seminar and detailed survey. This will be followed by refinement of the link type model in the light of the issues raised during this study and feedback from legal professionals. It seems also relevant to involve more annotators in the annotation of different types of laws, as well as links between (sub-)articles, by studying the extent to which drafting rules in documents such as [62] can indicate the type of norm.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


[1] T.F. Gordon, The role of exceptions in models of the law, Formalisierung Recht Ansätze Juristischer Expertensyst. (1986) 52–59.

[2] W. Peters, A. Wyner, Legal text interpretation: Identifying hohfeldian relations from text, in: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 2016, pp. 379–384.

[3] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, C. Soria, Automatic semantics extraction in law documents, in: Proceedings of the 10th International Conference on Artificial Intelligence and Law, in: ICAIL '05, Association for Computing Machinery, New York, NY, USA, 2005, pp. 133–140.

[4] M.-F. Moens, E. Boiy, R.M. Palau, C. Reed, Automatic detection of arguments in legal texts, in: Proceedings of the 11th International Conference on Artificial Intelligence and Law, in: ICAIL '07, Association for Computing Machinery, New York, NY, USA, 2007, pp. 225–230.

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

[6] H. Surden, Machine learning and law, Wash. Law Rev. 89 (2014) 87.

[7] D. Langone, A. Fulloni, D. Wonsever, A citations network for legal decisions, 2020, http://research.nii.ac.jp/~ksatoh/jurisin2020/, Jurisin conference.

[8] M. Adedjouma, M. Sabetzadeh, L.C. Briand, Automated detection and resolution of legal cross references: Approach and a study of luxembourg's legislation, in: 2014 IEEE 22nd International Requirements Engineering Conference (RE), IEEE, 2014, pp. 63–72.

[9] D. Bourcier, P. Mazzega, Codification, law article and graphs, Front. Artif. Intell. Appl. 165 (2007) 29.

[10] A. Sadeghian, L. Sundaram, D.Z. Wang, W.F. Hamilton, K. Branting, C. Pfeifer, Automatic semantic edge labeling over legal citation graphs, Artif. Intell. Law 26 (2) (2018) 127–144.

[11] E. Leitner, G. Rehm, J. Moreno-Schneider, Fine-grained named entity recognition in legal documents, in: International Conference on Semantic Systems, Springer, 2019, pp. 272–287.

[12] N. Sakhaee, M.C. Wilson, Information extraction framework to build legislation network, Artif. Intell. Law 29 (1) (2021) 35–58.

[13] R. Boulet, P. Mazzega, D. Bourcier, Network approach to the french system of legal codes part II: the role of the weights in a network, Artif. Intell. Law 26 (1) (2018) 23–47.

[14] M. Derlén, J. Lindholm, Goodbye van g end en l oos, hello b osman? Using network analysis to measure the importance of individual CJEU judgments, Eur. Law J. 20 (5) (2014) 667–687.

[15] M. van Opijnen, Citation analysis and beyond: in search of indicators measuring case law importance., in: JURIX, 250, 2012, pp. 95–104.

[16] M. van Opijnen, A model for automated rating of case law, in: Proceedings of the Fourteenth International Conference on Artificial Intelligence and Law, 2013, pp. 140–149.

[17] P. Mazzega, D. Bourcier, R. Boulet, The network of French legal codes, in: Proceedings of the 12th International Conference on Artificial Intelligence and Law, 2009, pp. 236–237.

[18] P. Zhang, L. Koppaka, Semantics-based legal citation network, in: Proceedings of the 11th International Conference on Artificial Intelligence and Law, in: ICAIL '07, Association for Computing Machinery, New York, NY, USA, 2007, pp. 123–130.

[19] M. Koniaris, I. Anagnostopoulos, Y. Vassiliou, Network analysis in the legal domain: A complex model for European union legal sources, J. Complex Netw. 6 (2) (2018) 243–268.

[20] A. Sadeghian, L. Sundaram, D. Wang, W. Hamilton, K. Branting, C. Pfeifer, Semantic edge labeling over legal citation graphs, in: Proceedings of the Workshop on Legal Text, Document, and Corpus Analytics (LTDCA-2016), 2016, pp. 70–75.

[21] G. Lame, Using NLP techniques to identify legal ontology components: Concepts and relations, Artif. Intell. Law 12 (4) (2004) 379–396.

[22] J.C. Maxwell, A.I. Antón, P. Swire, M. Riaz, C.M. McCraw, A legal cross-references taxonomy for reasoning about compliance requirements, Requir. Eng. 17 (2) (2012) 99–115.

[23] Y. Panagis, U. Sadl, F. Tarissan, Giving every case its (legal) due the contribution of citation networks and text similarity techniques to legal studies of european union law, in: 30th International Conference on Legal Knowledge and Information Systems (JURIX'17), 302, IOS Press, 2017, pp. 59–68.

[24] J.G. Conrad, L.K. Branting, Introduction to the special issue on legal text analytics, Artif. Intell. Law 26 (2) (2018) 99–102.

[25] M. Hildebrandt, The meaning and the mining of legal texts, in: Understanding Digital Humanities, Springer, 2012, pp. 145–160.
[26] L. Robaldo, S. Villata, A. Wyner, M. Grabmair, Introduction for artificial intelligence and law: special issue "natural language processing for legal texts", Artif. Intell. Law 27 (2) (2019) 113–115.
[27] I. Glaser, E. Scepankova, F. Matthes, Classifying semantic types of legal sentences: Portability of machine learning models., in: JURIX, 2018, pp. 61–70.
[28] J. Soh, H.K. Lim, I.E. Chai, Legal area classification: A comparative study of text classifiers on Singapore supreme court judgments, in: Proceedings of the Natural Legal Language Processing Workshop 2019, Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 67–77.
[29] B. Waltl, J. Muhr, I. Glaser, G. Bonczek, E. Scepankova, F. Matthes, Classifying legal norms with active machine learning., in: JURIX, 2017, pp. 11–20.
[30] B. Waltl, G. Bonczek, E. Scepankova, F. Matthes, Semantic types of legal norms in german laws: classification and analysis using local linear explanations, Artif. Intell. Law 27 (1) (2019) 43–71.
[31] T. Cheng, J.L. Cua, M.D. Tan, K.G. Yao, R.E. Roxas, Information extraction from legal documents, in: 2009 Eighth International Symposium on Natural Language Processing, 2009, pp. 157–162.
[32] I. Chalkidis, D. Kampas, Deep learning in law: early adaptation and legal word embeddings trained on large corpora, Artif. Intell. Law 27 (2) (2019) 171–198.
[33] A. Mandal, K. Ghosh, S. Ghosh, S. Mandal, Unsupervised approaches for measuring textual similarity between legal court case reports, Artif. Intell. Law (2021) 1–35.
[34] D.M. Katz, M.J. Bommarito II, J. Blackman, Predicting the behavior of the supreme court of the united states: A general approach, 2014, arxiv preprint arxiv:1407.6333.
[35] M. Medvedeva, M. Vols, M. Wieling, Using machine learning to predict decisions of the European court of human rights, Artif. Intell. Law 28 (2) (2020) 237–266.
[36] L. Humphreys, C. Santos, L. Di Caro, G. Boella, L. Van Der Torre, L. Robaldo, Mapping recitals to normative provisions in EU legislation to assist legal interpretation, in: JURIX, 2015, pp. 41–49.
[37] R. Nanda, L.D. Caro, G. Boella, H. Konstantinov, T. Tyankov, D. Traykov, H. Hristov, F. Costamagna, L. Humphreys, L. Robaldo, M. Romano, A unifying similarity measure for automated identification of national implementations of european union directives, in: Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law, ICAIL 2017, London, United Kingdom, June 12-16, 2017, 2017, pp. 149–158.
[38] R. Nanda, L. Humphreys, L. Grossio, A.K. John, Multilingual legal information retrieval system for mapping recitals and normative provisions, JURI SAYS (2020) 123.
[39] I.A. Amantea, L.D. Caro, L. Humphreys, R. Nanda, E. Sulis, Modelling norm types and their inter-relationships in EU directives, in: Proc. of the Third Workshop ASAIL Co-Located At ICAIL, Montreal, CEUR-WS.org, 2019, pp. 1–10.
[40] M. den Heijer, T.v.O.v.d. Abeelen, A. Maslyka, On the use and misuse of recitals in European union law, SSRN Electron. J. 1 (2019–31) (2019).
[41] T. Klimas, J. Vaiciukaite, The law of recitals in European community legislation, ILSA J. Int. Comp. Law 15 (6) (2008) 61–93.
[42] K. Lenaerts, Interpretation and the court of justice: A basis for comparative reflection, Int'L Law. 41 (2007) 1011.
[43] D. Spencer, Card Sorting: Designing Usable Categories, first ed., Rosenfeld Media, 2009.
[44] R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, Comput. Linguist. 34 (4) (2008) 555–596.
[45] A. Clark, C. Fox, S. Lappin, The Handbook of Computational Linguistics and Natural Language Processing, John Wiley & Sons, 2013.
[46] K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. (1972).
[47] J. Saramäki, M. Kivelä, J.-P. Onnela, K. Kaski, J. Kertesz, Generalizations of the clustering coefficient to weighted complex networks, Phys. Rev. E 75 (2) (2007) 027105.
[48] L.C. Freeman, A set of measures of centrality based on betweenness, Sociometry 40 (1) (1977) 35–41.
[49] U. Brandes, D. Fleischer, Centrality measures based on current flow, in: Annual Symposium on Theoretical Aspects of Computer Science, Springer, 2005, pp. 533–544.
[50] R.S. Burt, Structural holes and good ideas, Am. J. Sociol. 110 (2) (2004) 349–399.
[51] G. Palla, I. Derényi, I. Farkas, T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, Nature 435 (7043) (2005) 814–818.
[52] D. Liben-Nowell, J. Kleinberg, The link prediction problem for social networks, in: Proceedings of the Twelfth International Conference on Information and Knowledge Management, in: CIKM '03, Association for Computing Machinery, New York, NY, USA, 2003, pp. 556–559.
[53] V. Latora, M. Marchiori, Efficient behavior of small-world networks, Phys. Rev. Lett. 87 (19) (2001) 198701.
[54] A.-H. Esfahanian, Connectivity algorithms, in: Topics in Structural Graph Theory, Cambridge University Press, 2013, pp. 268–281.
[55] A.V. Joshi, Machine Learning and Artificial Intelligence, Springer, 2020.
[56] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, O'Reilly Media, 2009.
[57] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[58] A. Hagberg, P. Swart, D. S Chult, Exploring Network Structure, Dynamics, and Function Using NetworkX, Tech. Rep., Los Alamos National Lab.(LANL), Los Alamos, NM (United States), 2008.
[59] M. Bastian, S. Heymann, M. Jacomy, Gephi: An open source software for exploring and manipulating networks, in: Third International AAAI Conference on Weblogs and Social Media, The AAAI Press, 2009, pp. 361–362.
[60] T. Tullis, L. Wood, How many users are enough for a card-sorting study?, in: Learning, 2004, pp. 1–10.
[61] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. Theory Exp. 2008 (10) (2008) P10008.
[62] E. Commission, Joint practical guide of the European parliament, the council and the commission for persons involved in the drafting of European union legislation, 2016.