Exploiting textual similarity techniques in harmonization of laws

Emilio Sulis[0000−0003−1746−3733], Llio Bryn Humphreys[0000−0002−2484−0026], Davide Audrito[0000−0002−9239−5358], and Luigi Di Caro[0000−0002−7570−637X]

Computer Science Department, University of Torino, Corso Svizzera 185, Torino, Italy
{emilio.sulis,lliobryn.humphreys,davide.audrito,luigi.dicaro}@unito.it

Abstract. This paper describes an application of textual similarity techniques in the Legal Informatics domain. In European law, a relevant interest relates to the transposition of EU directives by the Member States, which can be complete, partial, or eventually absent. As part of an European project, legal experts annotated transpositions of six directives on a per-article basis. Following an established NLP pipeline, we explore a similarity-based technique to identify correspondences between transpositions of national implementations. Early results are promising and show the role that Artificial Intelligence may play within the process of harmonization and standardization of domestic legal systems as a result of the adoption of EU legislation.

Keywords: Legal Informatics · Text Similarity · Harmonization of Laws · Natural Language Processing

1 Introduction

Computational text analysis is an important research area with many practical applications in a variety of research areas, e.g. sentiment analysis [19], marketing [16], education [1], business process management [3]. Typically, text mining techniques concern unstructured text, such as reviews, social media posts, and online comments [28,40]. The goal of these analyses mostly involves identifying patterns and extracting knowledge through supervised or unsupervised mechanisms [15]. Law represents a rapidly-growing area of application of Natural Language Processing (NLP)[30], and Legal Informatics is a particular research area which concerns the application of Information and Communication Technologies (ICT) in the legal domain [22,13]. Legislative documents are usually formally structured and contain special features such as preambles, citations, recurring phrases, and references [7].

This contribution concerns European Law, focusing on the *approximation of laws* and *harmonization*, i.e. the alignment of domestic legal frameworks in light of the EU legislation. In the EU, legislative harmonization has two important functions: first, it reduces legal differences between Member States, with a view to foster economic, social and cultural exchanges. Moreover, it aims to achieve a variety of political results, e.g. the establishment of a European single market,

the achievement of common minimum standards regarding social protection, the establishment of rules concerning the rights of suspects and accused persons in criminal proceedings. This concept finds practical application in the analysis of national implementations (NIMs) of European directives, i.e., the *transposition* of European law in each Member State legislation. In particular, we describe the results of a legal experts' effort aiming at identifying and labeling NIMs. In general, "a directive shall be binding, as to the result to be achieved, upon each Member State to which it is addressed, but shall leave to the national authorities the choice of form and methods" (Article 288(3), Treaty on the Functioning of the European Union). As such, although national legislators have a certain margin of discretion in the choice of methods and forms for implementation, a certain degree of similarity of NIMs is expected. By comparing the English versions of the different implementations, this type of legal text can be explored using computational methods to assess the similarity of legal texts [5].

As a case study, we based on a research project in which "transpositions" of six EU directives were assessed "manually" by legal experts. Two main methods have been used for transposing EU law into national law: i.*Copy-out*: implementing legislation adopts the same, or mirrors as closely as possible the original wording of the directive; ii.*Elaboration*: choosing a particular meaning according to what the draftsperson believes the provision to mean, with the aim of working a provision into something clearer (this is an UK practice). The typical method for transpositions is *copy-out* [12]. In this respect, texts of NIMs are expected to be similar.

As main objective, we investigate the impact and efficacy of standard text analysis techniques applied to NIMs, focusing on the following research questions:

i Can we compare the implementations of EU directives in different countries by using NLP techniques?
ii By focusing on "Explicitly Transposed" articles for each directive in TT, can we adopt some meaningful metrics (e.g., similarity or network measures) to compare (pairs of) NIMs? Are these metrics significant at the article's granularity level?

In this paper we describe an essential application of NLP for legal texts by taking advantage of the initial results of an ongoing EU research project, *CrossJustice*. We first introduce some related work (Section 2), and the dataset of the case study (Section 3). Then, we report a possible solution to investigate the harmonization with similarity of NIMs (Section 4). We conclude the paper in Section 5.

2 Related Work

Legal research has seen an increased focus on the use of Artificial Intelligence (AI) techniques to the law [46,17,45,10,8,21,33,9]. In a critical area of AI, machine learning techniques include similarity measures [31] as an essential analysis in a

NLP pipeline [14]. Existing methodologies for finding similar legal documents can be classified into two main categories [25,6]: (i) network-based methods, which rely on citations to prior case documents [43]; (ii) text-based methods, which use the content/textual information of the documents [24]. We explore (ii), whereas recent works on 'similarity' in legal informatics concern the comparison between the EU directive and the transposition into the national law [18,20].

Text mining and NLP techniques have been explored to assist the Commission and legal professionals in studying and evaluating the transposition of directives at a fine-grained provision level [29]. Some approaches adopted embeddings models [26] to represent legal texts in a semantic vector space, by applying the method of cosine similarity (CS) [32]. Recent work addressed the task of identifying similarities among court rulings by adopting a graph-based method, to identify prominent concepts present in a ruling by extracting representative sentences [44]. Some experiments on legal judgments [25] explored CS by considering the document vector, where each term score is calculated with Term frequency – Inverse document frequency method (Tf-Idf) [34]. They performed well by considering only legal terms in the document vector, instead of using all terms or co-citations. In previous work, a pipeline with Tf-Idf, stemming, and co-occurrence networks has been shown to be significant in the automatic analysis of legal texts [38].

Finally, a recent work has measured the similarity between two court case documents, observing how "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" [27].

Harmonization. The effective protection of fundamental rights throughout the EU is heavily affected by the highly varying legal frameworks which characterize Member States regulation on procedural rights [42,4]. Legal actors often struggle to identify which legislation and therefore which procedural rights are applicable to persons accused or suspected of a crime in specific cases, due to both language barriers and the peculiarities of different national legal systems [35,36]. This situation persists also after the introduction of the EU directives derived from the Stockholm Programme, aiming at creating a certain level of harmonized rules on the matter [23]. A directive comes into effect only after it has been transposed into national law by the Member States, via the so-called NIMs [37].

3 Case Study

3.1 CrossJustice project

The CrossJustice (CJ) project[1] on which this work is based concerns the compliance of national instruments implementing EU directives with the *acquis communautaire*, in the protection of fundamental rights for persons accused or suspected of a crime (one of the main objectives of EU policy in the field of justice).

[1] `https://www.crossjustice.eu`

Legal experts have been involved to assess the compatibility between national frameworks as a result of the implementation of six EU directives. The output concerns the creation of a web platform to support and disseminate the results.

CJ aims to tackle the issues described above by identifying critical gaps and solutions in a comparative perspective, to improve the efficiency of judicial systems and their cooperation, thanks to information and communication technology. The online platform contains advice and support on the effectiveness of procedural rights providing a free service, mainly directed to legal professionals, but accessible to law students, NGOs and all EU citizens.

The CJ platform[2] addresses information pertaining to procedural rights, by delivering: i) A free of charge and updated information and advisory service directed to legal professionals (lawyers, magistrates, and public servants), but also accessible to law students and citizens. ii) Capacity building for legal professionals and law students.

3.2 Types of annotations

The annotation process from the legal experts used the following four labels to distinguish the four types of national implementations:

1. Explicitly transposed - either via new legislation or via amendments to existing legislation.
2. De facto/indirectly implemented - transposition unnecessary because the right already existed in previous legislation.
3. No national implementation (either explicitly or de facto/indirectly) - lack of transposing national norm or non-conformity of the national norm with the requirements of the EU provision.
4. Specific transposition is not required - transposition may be unnecessary because: i) The legal provision lacks deontic or constitutive value e.g. articles 1 and 2 of directives usually only define the scope of the directive; ii) Member states may derogate from a particular provision (e.g. Article 6(3) of directive 2016/800).

3.3 Dataset overview

The six EU directives under consideration (2010/64, 2012/13, 2013/48, 2016/343, 2016/800, 2016/1919) obtain different transpositions in the laws of different Member States. Legal experts involved in CJ annotated each part (e.g, an article or a paragraph) of a directive with both the above mentioned labels and the text of transposing legal provisions with a commentary in the so-called Transposition Table (TT). The CJ platform includes 3,458 annotations in the TT – as extracted on 1st June 2021 – and the distribution is represented in Table 1. The TT contains several differences among the Member States in terms of the number of annotations. For instance, the Member State with the lowest number

[2] `https://www.crossjustice.eu/en/index.html#crossjustice-platform`

of TT annotations is Bulgaria (223), the highest is Portugal (375). In particular, the number of explicitly transposed (ET) parts of EU directives in the TT varies depending on the Member State. Croatia and the Netherlands have the highest value of "explicit" transpositions, while Portugal and Sweden have the lowest number of ETs according to the CJ table.

Table 1. Number of NIMs of the considered EU directives by Member States and by four types: Explicitly transposed (Explicit), De facto/indirectly implemented (Indirect), No national implementation (NoImpl), Specific transposition is not required (NotReq)

Member State	Explicit	Indirect	NoImpl	NotReq	Total
Bulgaria	40	151	17	15	223
Croatia	146	81	24	0	251
France	49	153	41	0	243
Germany	99	234	11	0	344
Italy	65	221	32	0	318
Netherlands	150	146	57	16	369
Poland	32	154	86	0	272
Portugal	0	353	22	0	375
Romania	85	239	50	0	374
Spain	91	135	89	0	315
Sweden	8	325	0	41	374
Total	**765**	**2,192**	**429**	**72**	**3,458**

4 Methodology

4.1 Text processing

We adopted a quite established NLP pipeline with preprocessing, stemming, and calculating n-grams. The processed data needed to be converted into a numerical format, where each text is represented by an array (vectors). In natural language processing, the assumption about vectorization is that similar texts must result in nearest-neighbor vectors (i.e., vectors derived from textual data to reflect various linguistic properties of the text).

In particular, the here proposed methodological framework includes the analysis of legal texts of the TT by using both *bag-of-ngrams* and the frequency of terms with Tf-Idf.

NIMs have been processed with the following four main phases:

- *Preprocessing and POS tagging.* We processed texts according to the following steps: lower case reduction, stop words and punctuation removal, pos-tagging (to consider only nouns, verbs, and adjectives).
- *Stemming.* Stemming further reduces the variability of the text. The root form of terms is computed according to Porter stemming algorithm [41]. Finally, we removed all the *stems* of one single char length.

- *Modeling text.* The automatic analysis of legal text requires a numerical text representation (model). A typical computational approach in NLP and IR represents text in vectors of frequency of terms (bag-of-words). Another typical approach considers the aggregations of a certain number of letters (n) which appear contiguous in the given text or speech (*n-grams*). In particular, *bigrams* are sequences of two consecutive terms, while *trigrams* are three consecutive terms. With *bag-of-ngrams* models, by considering *n-grams*, instead of individual words (*stems*), we obtain different more effective representation of the same text. Furthermore, most frequent features can be selected to reduce sparsity. Finally, the corresponding vector of numbers counts the occurrences of terms in the document.
- *Tf-Idf transform.* A typical automatic text analysis pipeline involves transforming each piece of text (i.e., legal provisions) into a vector, where each word is replaced by significant numbers. Such numbers can be mere counts or frequency of occurrences, as well as more sophisticated measures such as Tf-Idf. This scoring measure is widely used in NLP based on the complete collection of terms from the transpositions of each directive (each directive therefore has a *corpus* of variable dimension). Term frequency–inverse document frequency (Tf-Idf) is a numerical statistic for reflecting how important a word is to a document in our collection. The measure implies two parts: Term Frequent (Tf) simply describes how frequently a term (t) appears in each document (d). Inverse Document Frequency (Idf) computes the importance of the term in the complete collection.
- *Document-Term Matrix.* For each individual NIMs we obtain a vector in the corresponding Document-Term Matrix (DTM). In the resulting matrix, every row is a NIM (here, a single TT part/article) and every column is a term/stem/n-gram. The values in the matrix are the frequencies of each term in a document. As the columns are too many, we considered the application of a dimensionality reduction strategy, e.g. Principal Component Analysis (PCA) or Multidimentional Scaling (MDS) [11]. We opted to reduce the number of features with MDS, which improves similarity measure by exploiting the latent semantics of co-occurrences between words.

4.2 Similarity measure

To investigate text similarity with the above mentioned research objectives, we adopted CS as an established similarity metrics in this kind of research. Mathematically, CS represents the cosine of the angle between two vectors projected in a multi-dimensional space. In particular, CS between the vectors of two NIMs (A and B) is computed as follows:

$$CS(A, B) = \frac{V1V2}{\|V1\|\|V2\|}$$

The numerator is the dot product of the vectors V1 and V2, representing A and B respectively. The denominator is the product of their Euclidean norms,

which normalizes the similarity value. The range of values that the CS can vary is -1, 1. The CS values have been computed between the NIMs at the level of each part/article considered in the TT[3]. For instance, in an EU directive, by considering two Member States (e.g., Italy and Bulgaria), we compare the corresponding "Explicit transpositions" of Article 1 both in Italy and in Bulgaria. Finally, we obtain the most similar NIMs for each EU directive.

5 Output

5.1 Text representation

We summarize here the transformation of each text (corresponding to *Explicit transpositions*) to a fixed-length vector of integer values by describing the *bag-of-ngrams* output, and dimensionality reduction, as better detailed in the following paragraphs.

Explicit transpositions. We focused on the annotation effort of CJ's legal experts, who indicated in the TTs the parts (at the level of Article or Sub-Article) of the EU directives that were explicitly transposed in Member States legislation. For instance, the case of EU directive 2012/13 has 563 different implementations of different types, of these the ET implementations are 245. In particular, Article 2 has only 7 ET cases concerning 2 Member States, i.e. Croatia (4) and Spain (3), according to the complete database (a view in Figure 1). Next, we considered merging the contents of all implementations regarding the same part of the EU directive, for each Member State.

ID	EUdir	State	NumArt	Label	Text
2039	0013	Croatia	art_2	Exp	summon suspect must specifi suspect suspect theins...
2040	0013	Croatia	art_2	Exp	upon arrest arrest person must immedi provid writt...
2041	0013	Croatia	art_2	Exp	letter right must deliv accus person search warran...
2043	0013	Spain	art_2	Exp	ani person punish act attribut may exercis right d...
2044	0013	Spain	art_2	Exp	admiss complaint suit ani procedur action imput cr...
2045	0013	Spain	art_2	Exp	right defens shall exercis without limit expressli...
2050	0013	Croatia	art_2	Exp	prior file indict compet court bodi proceed perpet...

Fig. 1. A view of Explicitly Transposed (ET) legal provisions for each parts of Article 2 of the European directive n.2012/13.

Bag-of-words and n-grams. We considered ET of NIMs as our *corpus*, for each EU directive. With the bag-of-words technique we represented the text of each document in numbers, based on a vocabulary from all the unique *stems*. As

[3] From the *scikit-learn* python library *sklearn.metrics.pairwise* we adopted cosine_similarity method

mentioned, we obtained the *bag-of-ngrams* of our *corpus*, as a more sophisticated approach based on a vocabulary of grouped *stems* of length n (i.e., *n-grams*). We computed the *stems* for 1,714 individual parts of our ET implementations for the considered EU directives. In particular, we obtained a median value of 105 *stems*, as well as a maximum of 1,365 *stems* (for EU directive 2013/48, art_10 paragraph_3).

Dimensionality reduction. The definitive *corpus* includes the vectors for each article which has been explicitly transposed in TT, where the 'columns' are the terms (or the n-grams considered). In the case of bigrams, the number of features is 4,549. In the case of trigrams, the features are 6,213. We reduced the dimension of the problem with multidimensional scaling of different size, e.g. 100 or 200 features.

Similarity. The CS between two implementations of each pair of Member States describes the degree of similarity between the vector representations of the text. For instance, we mention here the simple case of the "Annex 1" of the 2012/13 European directive which has been explicitly implemented by three States (France, Spain, Romania). A "manual" inspection of the three corresponding NIMs describes a certain similarity only between France and Spain, and not in the other pairs. This is also true after observing the CS measures, both using top 100 more frequent terms (Vect100) or MDS method with 100 or 200 features (MDS100, MDS200), as described in Table 2.

Table 2. An example of similarity scores concerning the Article "Ann_1" of the European directive n.2012/13 for three pairs of States.

Member States		**Vect100**	**MDS100**	**MDS200**
France	**Romania**	0.376	0.055	0.056
France	**Spain**	0.581	0.206	0.205
Romania	**Spain**	0.460	0.079	0.084

5.2 Heat maps visualization

To facilitate the understanding of the results, we considered heat maps describing the degree of similarity between pairs of Member States, for each parts of NIMs. Darker colors (e.g., blue or green) imply no similarity, while lighter colors (e.g., white/yellow) indicate a certain degree of text similarity. For the sake of clarity, Figure 2 is an example of a heat map concerning a NIM of EU directive n. 2012/13. This type of visualization clearly describes how France and Sweden have more similar text (lighter color) than other States. The diagonal is null (dark color, in our case), because the relationship between the text of a State and itself is not considered. This type of visualization allows an immediate understanding of the similarity among NIMs. As part of the project, legal experts have confirmed that the heat maps are meaningful. Therefore, the tool appears

to be useful in helping analysts to detect/suggest the degree of harmonization of national laws.

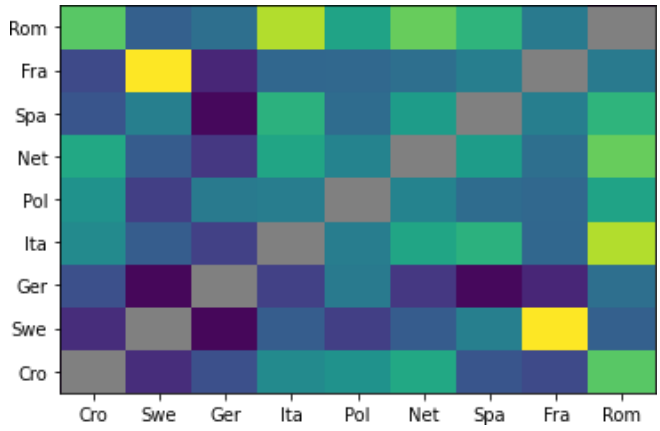

Fig. 2. A heat map representation of similarity metrics for NIMs

6 Conclusions

This paper discussed the first outcome of an ongoing research project involving NLP and Law, focusing on the key concept of Harmonization in EU Law. We investigate computational text similarity technique to the idea of making identical rules in more areas of governance. We performed similarity metrics computation, analysis of results, and visualisation to demonstrate how an established NLP pipeline for preprocessing text and similarity metrics can be applied to support legal harmonization purposes.

As a future work, we aim to explore network analysis techniques with co-occurrence of terms or *stems*. The approach already provided meaningful results [38,39] in modeling inter-relationships between norms [2]. We plan to investigate different similarity techniques and hybrid approaches, including embedding methods (e.g., Node2Vec for graph embedding approach or Word2Vec implementation). Finally, we plan to extend the evaluation with a "user study" and at the same time propose an extension of the technology used in the CJ project.

Acknowledgement

This work has been supported by the European Union's Justice Programme (Grant Agreement No. 847346) for the project "Knowledge, Advisory and Capacity Building Information Tool for Criminal Procedural Rights in Judicial Cooperation".

References

1. Amado, A., Cortez, P., Rita, P., Moro, S.: Research trends on big data in marketing: A text mining and topic modeling based literature analysis. European Research on Management and Business Economics **24**(1), 1–7 (2018)
2. Amantea, I.A., Caro, L.D., Humphreys, L., Nanda, R., Sulis, E.: Modelling norm types and their inter-relationships in EU directives. In: Ashley, K.D., Atkinson, K., Branting, L.K., Francesconi, E., Grabmair, M., Waltl, B., Walker, V.R., Wyner, A.Z. (eds.) Proceedings of the Third Workshop on Automated Semantic Analysis of Information in Legal Texts co-located with the 17th International Conference on Artificial Intelligence and Law (ICAIL 2019), Montreal, QC, Canada, June 21, 2019. CEUR Workshop Proceedings, vol. 2385. CEUR-WS.org (2019), `http://ceur-ws.org/Vol-2385/paper8.pdf`
3. Amantea, I.A., Robaldo, L., Sulis, E., Boella, G., Governatori, G.: Semi-automated checking for regulatory compliance in e-health. In: 25th International Enterprise Distributed Object Computing Workshop, EDOC Workshop 2021, Gold Coast, Australia, October 25-29, 2021. pp. 318–325. IEEE (2021). https://doi.org/10.1109/EDOCW52865.2021.00063
4. Andenas, M., Andersen, C.B.: Theory and practice of harmonisation. Edward Elgar Publishing (2012)
5. Ashley, K.D.: Artificial intelligence and legal analytics: new tools for law practice in the digital age. Cambridge University Press (2017)
6. Bhattacharya, P., Ghosh, K., Pal, A., Ghosh, S.: Methods for computing legal document similarity: A comparative study. CoRR **abs/2004.12307** (2020), `https://arxiv.org/abs/2004.12307`
7. Biasiotti, M., Francesconi, E., Palmirani, M., Sartor, G., Vitali, F.: Legal informatics and management of legislative documents. Global Center for ICT in Parliament Working Paper **2** (2008)
8. Boella, G., Di Caro, L., Humphreys, L., Robaldo, L., Rossi, P., van der Torre, L.: Eunomos, a legal document and knowledge management system for the web to provide relevant, reliable and up-to-date information on the law. Artificial Intelligence and Law **24**(3), 245–283 (2016)
9. Boella, G., Di Caro, L., Leone, V.: Semi-automatic knowledge population in a legal document management system. Artificial Intelligence and Law **27**(2), 227–251 (2019)
10. Boella, G., Di Caro, L., Rispoli, D., Robaldo, L.: A system for classifying multi-label text into eurovoc. In: Proceedings of the fourteenth international conference on artificial intelligence and law. pp. 239–240 (2013)
11. Cox, M.A., Cox, T.F.: Multidimensional scaling. In: Handbook of data visualization, pp. 315–347. Springer (2008)
12. Dimitrakopoulos, D.G.: The transposition of eu law:'post-decisional politics' and institutional autonomy. European Law Journal **7**(4), 442–458 (2001)
13. Durante, M.: Computational Power: The Impact of ICT on Law, Society and Knowledge. Routledge (2021)
14. Elekes, Á., Schäler, M., Böhm, K.: On the various semantics of similarity in word embedding models. In: 2017 ACM/IEEE Joint Conference on Digital Libraries, JCDL 2017, Toronto, ON, Canada, June 19-23, 2017. pp. 139–148. IEEE Computer Society (2017). https://doi.org/10.1109/JCDL.2017.7991568, `https://doi.org/10.1109/JCDL.2017.7991568`

15. Feldman, R., Sanger, J.: The Text Mining Handbook - Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press (2007)
16. Ferreira-Mello, R., André, M., Pinheiro, A., Costa, E., Romero, C.: Text mining in education. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **9**(6), e1332 (2019)
17. Friedrich, R., Luzzatto, M., Ash, E.: Entropy in legal language. In: Aletras, N., Androutsopoulos, I., Barrett, L., Meyers, A., Preotiuc-Pietro, D. (eds.) Proceedings of the Natural Legal Language Processing Workshop 2020 co-located with the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020), Virtual Workshop, August 24, 2020. CEUR Workshop Proceedings, vol. 2645, pp. 25–30. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2645/paper4.pdf
18. Haverland, M., Steunenberg, B., Van Waarden, F.: Sectors at different speeds: analysing transposition deficits in the european union. JCMS: Journal of Common Market Studies **49**(2), 265–291 (2011)
19. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 168–177 (2004)
20. Humphreys, L., Santos, C., Di Caro, L., Boella, G., Van Der Torre, L., Robaldo, L.: Mapping recitals to normative provisions in eu legislation to assist legal interpretation. In: JURIX. pp. 41–49 (2015)
21. John, A.K., Di Caro, L., Robaldo, L., Boella, G.: Legalbot: A deep learning-based conversational agent in the legal domain. In: International Conference on Applications of Natural Language to Information Systems. pp. 267–273. Springer (2017)
22. Katz, D.M., Dolin, R., Bommarito, M.J.: Legal Informatics. Cambridge University Press (2021)
23. Kaunert, C., Occhipinti, J.D., Léonard, S.: Introduction: supranational governance in the area of freedom, security and justice after the stockholm programme (2014)
24. Kim, M.Y., Xu, Y., Goebel, R.: Legal question answering using ranking svm and syntactic/semantic similarity. In: JSAI International Symposium on Artificial Intelligence. pp. 244–258. Springer (2014)
25. Kumar, S., Reddy, P.K., Reddy, V.B., Suri, M.: Finding similar legal judgements under common law system. In: Madaan, A., Kikuchi, S., Bhalla, S. (eds.) Databases in Networked Information Systems - 8th International Workshop, DNIS 2013, Aizu-Wakamatsu, Japan, March 25-27, 2013. Proceedings. Lecture Notes in Computer Science, vol. 7813, pp. 103–116. Springer (2013). https://doi.org/10.1007/978-3-642-37134-9_9, https://doi.org/10.1007/978-3-642-37134-9_9
26. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. Trans. Assoc. Comput. Linguistics **3**, 211–225 (2015). https://doi.org/10.1162/tacl_a_00134
27. Mandal, A., Ghosh, K., Ghosh, S., Mandal, S.: Unsupervised approaches for measuring textual similarity between legal court case reports. Artif. Intell. Law **29**(3), 417–451 (2021). https://doi.org/10.1007/s10506-020-09280-2
28. Meo, R., Sulis, E.: Processing affect in social media: A comparison of methods to distinguish emotions in tweets. ACM Trans. Internet Techn. **17**(1), 7:1–7:25 (2017). https://doi.org/10.1145/2996187, https://doi.org/10.1145/2996187
29. Nanda, R., Siragusa, G., Caro, L.D., Boella, G., Grossio, L., Gerbaudo, M., Costamagna, F.: Unsupervised and supervised text similarity systems for automated identification of national implementing measures of european directives. Artificial Intelligence and Law **27**, 199–225 (2018). https://doi.org/https://doi.org/10.1007/s10506-018-9236-y

30. Nay, J.J.: Natural Language Processing for Legal Texts, p. 99–113. Cambridge University Press (2021). https://doi.org/10.1017/9781316529683.011
31. Ontañón, S.: An overview of distance and similarity functions for structured data. Artificial Intelligence Review **53**(7), 5309–5351 (2020). https://doi.org/https://doi.org/10.1007/s10462-020-09821-w
32. Renjit, S., Idicula, S.M.: CUSAT nlp@aila-fire2019: Similarity in legal texts using document level embeddings. In: Mehta, P., Rosso, P., Majumder, P., Mitra, M. (eds.) Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation, Kolkata, India, December 12-15, 2019. CEUR Workshop Proceedings, vol. 2517, pp. 25–30. CEUR-WS.org (2019), http://ceur-ws.org/Vol-2517/T1-4.pdf
33. Robaldo, L., Villata, S., Wyner, A., Grabmair, M.: Introduction for artificial intelligence and law: special issue "natural language processing for legal texts" (2019). https://doi.org/doi10.1007/s10506-019-09251-2
34. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information processing & management **24**(5), 513–523 (1988)
35. Satzger, H.: The harmonisation of criminal sanctions in the european union - a new approach. eucrim (2019). https://doi.org/10.30709/eucrim-2019-007, https://doi.org/10.30709/eucrim-2019-007
36. Schroeder, W.: Limits to european harmonisation of criminal law. eucrim (2020). https://doi.org/10.30709/eucrim-2020-008, https://doi.org/10.30709/eucrim-2020-008
37. Steunenberg, B., Rhinard, M.: The transposition of european law in eu member states: between process and politics. European Political Science Review **2**, 495 – 520 (2010). https://doi.org/https://doi.org/10.1017/S1755773910000196
38. Sulis, E., Humphreys, L., Vernero, F., Amantea, I.A., Audrito, D., Di Caro, L.: Exploiting co-occurrence networks for classification of implicit inter-relationships in legal texts. Information Systems p. 101821 (2021). https://doi.org/https://doi.org/10.1016/j.is.2021.101821
39. Sulis, E., Humphreys, L., Vernero, F., Amantea, I.A., Caro, L.D., Audrito, D., Montaldo, S.: Exploring network analysis in a corpus-based approach to legal texts: A case study. In: Tagarelli, A., Zumpano, E., Latific, A.K., Calì, A. (eds.) Proceedings of the First International Workshop "CAiSE for Legal Documents" (COUrT 2020) co-located with the 32nd International Conference on Advanced Information Systems Engineering (CAiSE 2020), Grenoble, France, June 9, 2020. CEUR Workshop Proceedings, vol. 2690, pp. 27–38. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2690/COUrT-paper3.pdf
40. Sulis, E., Lai, M., Vinai, M., Sanguinetti, M.: Exploring sentiment in social media and official statistics: a general framework. In: Bosco, C., Cambria, E., Damiano, R., Patti, V., Rosso, P. (eds.) Proceedings of the 2nd International Workshop on Emotion and Sentiment in Social and Expressive Media: Opportunities and Challenges for Emotion-aware Multiagent Systems co-located with 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Istanbul, Turkey, May 5, 2015. CEUR Workshop Proceedings, vol. 1351, pp. 96–105. CEUR-WS.org (2015), http://ceur-ws.org/Vol-1351/paper8.pdf
41. Van Rijsbergen, C.J., Robertson, S.E., Porter, M.F.: New models in probabilistic information retrieval, vol. 5587. British Library Research and Development Department London (1980)
42. Vogenauer, S., Weatherill, S.: The harmonisation of European contract law: implications for European private laws, business and legal practice. Bloomsbury Publishing (2006). https://doi.org/https://doi.org/10.1111/j.1468-0386.2007.00376_4.x

43. Wagh, R., Anand, D.: Application of citation network analysis for improved similarity index estimation of legal case documents : A study. In: 2017 IEEE International Conference on Current Trends in Advanced Computing (ICCTAC). pp. 1–5 (2017). https://doi.org/10.1109/ICCTAC.2017.8249996
44. Wagh, R.S., Anand, D.: Legal document similarity: a multi-criteria decision-making perspective. PeerJ Comput. Sci. **6**, e262 (2020). https://doi.org/10.7717/peerj-cs.262, https://doi.org/10.7717/peerj-cs.262
45. Wyner, A., Mochales-Palau, R., Moens, M.F., Milward, D.: Approaches to text mining arguments from legal cases. In: Semantic processing of legal texts, pp. 60–79. Springer (2010)
46. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: How does NLP benefit legal system: A summary of legal artificial intelligence. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J.R. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020. pp. 5218–5230. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.466