








Unsupervised Factor Extraction from Pretrial Detention Decisions by Italian and Brazilian Supreme Courts

Isabela Cristina Sabo¹, Marco Billi², Francesca Lagioia^{2,3},
Giovanni Sartor^{2,3}, and Aires José Rover¹

¹ Department of Law, Federal University of Santa Catarina, Florianópolis, Brazil
isabelasabo@gmail.com

² CIRSFID - Alma AI, Alma Mater Studiorum- University of Bologna, Bologna, Italy

³ Department of Law, European University Institute, Florence, Italy

Abstract. Pretrial detention is a debated and controversial measure since it is an exception to the principle of the presumption of innocence. To determine whether and to what extent legal systems make excessive use of pretrial detention, an empirical analysis of judicial practice is needed. The paper presents some preliminary results of experimental research aimed at identifying the relevant factors on the basis of which Italian and Brazilian Supreme Courts impose the measure. To analyze and extract the relevant predictive-features, we rely on unsupervised learning approaches, in particular association and clustering methods. As a result, we found common factors between the two legal systems in terms of crime, location, grounds for appeal, and judge's reasoning.

Keywords: E-justice · Case factors extraction · Machine learning · Association rules · Clustering · Criminal law · Pretrial detention

1 Introduction

In criminal proceedings, pretrial detention is debated and controversial, since it is an exception to the fundamental principle of the presumption of innocence, by depriving defendants of their liberty at the initial stages of proceedings, before their guilt is proven. The conditions under which such a measure is legitimate include, for instance, the reasonable suspicion of the person having committed the offence, the necessity to prevent defendants from absconding or committing further offence(s), and the risk of interfering with the course of justice during pending procedures. Their occurrence is subject to a case-by-case evaluation, based on the judge's discretionary assessment. Moreover, the remand measure

This research has been supported by Brazilian Institutional Program for Internationalization (CAPES/PrInt); ADELE (Analytics for Decision of Legal Cases, EU Justice program Grant (2014–2020); COMPULAW (Computable law), ERC Advanced Grant (2019–2024); LAILA (Legal Analytics for Italian Law), MIUR PRIN Programme (2017).

© The Authors(s) 2022

R. Guizzardi and B. Neumayr (Eds.): ER 2022 Workshops, LNCS 13650, pp. 69–80, 2022.

https://doi.org/10.1007/978-3-031-22036-4_7

shall last no longer than necessary to achieve the objectives pursued by the law [7]. Unfortunately, while there have been numerous studies on the legal framework governing pretrial detention, limited research has been carried out to date into the practice of pretrial detention decision-making. In this regard, Italy and Brazil are interesting fields of investigation¹. According to the World Prison Brief latest rates², in both countries, approximately 30% of the prison population are pretrial detainees. In this context, our research is aimed at identifying the relevant factors on the basis of which Italian and Brazilian Supreme Courts impose the pretrial detention –more exactly, maintain rather than reform, decisions on this matter by lower courts-, as well as how such factors relate to each other. To this end, we built two different corpora of Italian and Brazilian judicial decisions, as detailed in Sect. 2. Section 3 describes the unsupervised learning approaches, in particular association and clustering methods, used to analyse and extract the relevant predictive-features from the documents in the corpora. Section 4 reports the experimental setup and the results, as well as delineates commonalities and differences between the two legal systems. Section 5 concludes and outlines possible future research lines. This project follows recent attempts at explaining decision-making systems through factor-based reasoning, justifying decisions on the basis of legal features of a case [9, 10]. In order to identify the legally relevant factors, described by [2] as case decision predictors, we followed recent experiments as seen in [5].

2 Datasets

We built two different datasets of Brazilian and Italian judicial decisions, as we could not find any existing data collections to help augment our own. The Brazilian corpus consists of 2,018 documents, collected from the official Brazilian Supreme Court’s website (stf.jus.br). Documents are structured in the following sections: (a) heading (lawsuit metadata), (b) summary of the judgment, (c) case report (including the grounds of appeal), (d) reasons and decision of the judge-rapporteur, (e) votes of the other judges (when they differ from the judge-rapporteur), and (f) final decision. The Italian corpus consists of 718 judicial decisions by the Italian Supreme Court, downloaded from the DeJure database. Documents are structured according to the following sections: (a) heading (lawsuit metadata), (b) summary of the judgment, (c) case report (including the grounds of appeal), (d) reasons and (e) the final decision. In this regard, the main difference between the two corpora concerns the absence of dissenting statements in Italian rulings.

¹ For more information visit “Brazil has the world’s 3rd largest prison population.” <https://www.conectas.org/en/noticias/brazil-worlds-3rd-largest-prison-population/> (2017), online; accessed 30 May 2022; and “A measure of last resort? The practice of pretrial detention decision-making in the EU.” <https://www.fairtrials.org/articles/publications/a-measure-of-last-resort-the-practice-of-pre-trial-detention-decision-making-in-the-eu/> (2016), online; accessed 30 May 2022.

² World Prison Brief. <https://www.prisonstudies.org/>, online; accessed 09 Jun 2022.

3 Methodology

In this section we briefly describe the general methodology and the unsupervised learning techniques we employed. We approach the research problem in two goals: (i) identification, aimed at extracting the relevant factors, and (ii) correlation, aimed at finding relationships between the extracted factors and judicial outcomes, i.e., whether Italian and Brazilian Supreme Courts maintain rather than reform decisions on pretrial detention. To this end, we adopted, for both the Brazilian and Italian corpora, a four-step process. First, we manually extracted some factors from judgments which we call *objective* factors, since they are clearly stated in the text. Second, we addressed the association task to find possible relationships between these *objective* factors and the decision outcomes. Third, to automatically extract further relevant features, we split each dataset into 2 subsets, on the basis of the outcome of the decisions. Finally, we applied clustering methods to each subset in order to detect what we name *subjective* factors, i.e., those that are more difficult to identify. Note that we did not apply association methods to the *subjective* factors, since the 2 corpora were already split depending on their outcome. To perform our experiments, we have relied on existing implementations and standard methods, including the open-source software Orange 3 [6] and Carrot2 [15], as detailed in Sect. 4. In Sects. 3.1 and 3.2, we briefly explain association and clustering methods.

3.1 Association

To identify relationships between factors and outcomes, we extracted association rules having the forms $x \rightarrow y$, where x is a set of factors and y is one of the two outcomes. For each rule, we determined its support and confidence, namely (a) the proportion of the cases in which both the antecedent x and outcome y are satisfied (the likelihood of finding x and y cases), as a fraction of all cases in the dataset, (b) the proportion of cases in which outcome y is satisfied, as a fraction of all cases satisfying factors x (the likelihood of x cases have outcome y).

$$s(x \rightarrow y) = \frac{\text{Frequency}(x,y)}{N} ; c(x \rightarrow y) = \frac{\text{Frequency}(x,y)}{\text{Frequency}(x)} \quad (1)$$

In particular, we applied the FP-Growth association algorithm to scan the whole data and find the rules which satisfy given support thresholds. Then the rules were represented as a conditional tree, which saves the costly dataset scans in the subsequent mining processes [8].

3.2 Clustering

Clustering is an unsupervised learning task used to uncover hidden patterns in unlabeled data [12]. Considering that documents may present common factors, we adopted the so called soft clustering approach, whereby documents can be assigned to one or more clusters. In particular, we applied Hierarchical Clustering, which builds tree structures, by merging documents, and clusters of them, depending on similarities [1]. To assess similarities we used the cosine measure

[4]. Once clusters have been generated, we ran the Latent Semantic Indexing (LSI) algorithm, which captures the underlying semantics of textual documents and computes how words relate to each other, so as to reveal the occurrences of topics within the corpora [16]. We also used the Lingo algorithm, which extracts frequent phrases from documents, under the assumption that such phrases provide informative human-readable descriptions of topics. Among the techniques on which Lingo relies, we employed the LSI, aimed at discovering any existing latent structures of diverse topics. Finally, Lingo matches the cluster description with the extracted topics and assigns each document to one or more clusters. To select the best label for each cluster, it uses a score measure, based on cosine similarity [14].

4 Experiments and Results

As explained in Sect. 3, we addressed our research questions as identification and correlation goals. In the following we detail the experimental uptake, we report the results and make some considerations.

4.1 Manually Extracted Information

Following the first step, we manually extracted 5 *objective* factors: the prisoner status, the name of the judge rapporteur, the crime category, the crime location and the judgment date. In the following, we detail each factor and the values it may assume depending on the data.

1. *Prisoner Status*, i.e., the situation of the accused after the appeal ruling. This factor may have two alternative values, i.e., *released* and *not released*. Cases in which the Court replaced pretrial detention with house arrest, were considered as released.
2. *Judge Rapporteur*, i.e., the judge who furnishes a report on the case at hand. The Italian data is characterised by a higher variance compared to the Brazilian one, due to the different number of seats in the two Supreme Criminal Courts: at least 35 in the Italian Supreme Court, regularly replaced³, versus 11 seats in the Brazilian one, where judges have a permanent position.⁴
3. *Crime*, i.e., the general category to which the committed crime belongs to, under the Brazilian and Italian criminal laws. In particular, we identified four main categories: (i) “crimes against the person”, (ii) “crimes against property”, (iii) “drug-related crimes”, and (iv) “criminal organization”.
4. *Location*, i.e., the place where the crime took place. While in Brazil it corresponds to a state, in Italy it is represented by a regional capital.
5. *Date*, i.e., when the judgment was issued. It corresponds to the ruling year.

³ Corte di Cassazione (Area Penale): <https://www.cortedicassazione.it/corte-di-cassazione/it/area-penale.page/>, online; accessed 30 May 2022.

⁴ Supremo Tribunal Federal: <https://portal.stf.jus.br/ostf/>, online; accessed 30 May 2022.

Following the second step, we run experiments by employing the FP-Growth association algorithm (see Sect. 3). Table 1 indicates the specific parameters we adopted. To generate a set of reliable rules having *Released* as a consequent, we had to lower the required support and confidence scores (given the smaller number of realises-cases being present in each dataset).

Table 1. Association setup parameters.

Technique	Tool	Consequent itemset	Parameters
FP-Growth	Orange 3	BR <i>Not released</i>	Min. Supp. 4%, Min. Conf. 70%
		IT <i>Not released</i>	Min. Supp. 4%, Min. Conf. 70%
		BR <i>Released</i>	Min. Supp. 1%, Min. Conf. 40%
		IT <i>Released</i>	Min. Supp. 1%, Min. Conf. 40%

Tables 2 and 3 show some selected results. In particular, we report the rules presenting a certain degree of similarity within the two corpora.

Table 2. Association rules in Italian dataset.

No	Antecedent	→	Consequent	Supp	Conf
1	Criminal organization, Reggio Calabria	→	Not released	6,6%	93,8%
2	Drug law crime	→	Not released	23,8%	84,0%
3	Napoli	→	Not released	14,5%	82,0%
4	2019	→	Not released	4,1%	96,8%
5	Crime against property, criminal organization	→	Not released	7,2%	82,5%
6	2013, drug law crime, Napoli	→	Released	1,1%	88,9%

Table 3. Association rules in Brazilian dataset.

No	Antecedent	→	Consequent	Supp	Conf
1	Judge rapporteur MA	→	Not released	39,8%	82,2%
2	Drug law crime	→	Not released	30,5%	73,6%
3	São Paulo	→	Not released	31,9%	73,9%
4	2019	→	Not released	22,7%	94,4%
5	Crime against property, criminal organization	→	Not released	4,1%	81,1%
6	2013, drug law crime, São Paulo	→	Released	1,0%	47,4%

As we can note from rules no. 2 and no. 5 within the Italian and Brazilian datasets, drug-related crimes as well as the combination of criminal organization and crimes against property, are factors usually related to the *not released*

outcome. The same is true for the date factor 2019, the locations São Paulo and Naples, as shown in rules no. 3 and no. 4 in the two tables. Conversely, rule no. 6 in both datasets shows a relationship between the *released* outcome and the combination of date 2013, drug-related crimes and the location, respectively Naples and São Paulo. However, it should be noted that in the Brazilian dataset the confidence of this association rule is lower compared to the Italian one. From a general perspective, results show highly reliable association rules for the *not released* outcome within the two datasets. Conversely, we did not find association rules related to the *released* outcome with high confidence. This remains true even by reducing the confidence threshold.

4.2 Automatically Extracted Information

Following the third step, we split each corpus into two subsets, containing respectively the judgements for the defendant (*Released*) and for prosecution (*Not released*): in the Italian corpus, the first subset contains 614 judgements, and the second 104; in the Brazilian corpus respectively 1,503 and 515. We applied pre-processing techniques before clustering: normalization, tokenization combined with regular expressions, stemming, filtering of stop words and n -grams with $n = 2$ [12]. To encode sentences, in an effort to make our method as general as possible, we opted for well-established approaches. For the Lingo algorithm, we used the Bag of Words (BOW) model [11, 17]. In this model, one feature is associated with each word in the vocabulary. The value of each feature is usually computed as the $TF - IDF$ score, and measures the importance of the corresponding word. For the Hierarchical algorithm, we used Word Embeddings, a popular technique for language models and deep learning applications [3, 13]. The parameters adopted for clustering are reported in Table 4, depending on the outcomes and the number of documents in each subset.

Table 4. Clustering setup parameters.

Technique	Tool	Subset	Parameters
Lingo	Carrot2	IT <i>Not released</i> and <i>Released</i>	Cluster Count Base* 15%
		BR <i>Not released</i> and <i>Released</i>	Cluster Count Base 10%
Hierarchical clustering	Orange 3	BR and IT <i>Not released</i>	Height Ratio* 30%
		BR <i>Released</i>	Height Ratio 30%
		IT <i>Released</i>	Height Ratio 60%
LSI	Orange 3	All	3 Topics

*Measures used to calculate the number of clusters based on the number of documents on input.

Following the last step, for clustering, we rely on the Lingo algorithm, Hierarchical clustering and LSI. Tables 5, 6, 7 and 8 report some results obtained by using Lingo, sorted by highest score.

Table 5. Lingo clusters and labels in Italian *Not released* subset.

No.	Label and cluster	DN	Score	Type	Outcome
1	Maggio 2013 (C26)	61	36,15	Date	Not released
2	Nullità dell’interrogatorio dell’indagato (C10)	63	35,53	Grounds	Not released
3	Termini di fase previsti dall’art 303 (C4)	79	35,47	Grounds	Not released
4	Gravità indiziaria delle esigenze cautelari (C23)	61	33,05	Reason	Not released
5	Ipotesi di cui all’art 304 (C24)	61	32,22	Grounds	Not released
6	Napoli Emessa in data (C26)	61	31,43	Location	Not released
7	Principio della presunzione (C12)	63	30,27	Grounds	Not released
8	Reato Associativo Reati Fine (C5)	78	24,65	Crime	Not released

Table 6. Lingo clusters and labels in Brazilian *Not released* subset.

No.	Label and cluster	DN	Score	Type	Outcome
1	Vítima compareceu (C27)	150	25,87	Reason	Not released
2	Excesso prazo custódia perdurar 5 meses (C13)	152	24,65	Grounds	Not released
3	Senhora Ministra C. L. Presidente Exatamente (C3)	151	24,23	Judge	Not released
4	Prática crimes tráfico drogas porte (C25)	150	22,34	Crime	Not released
5	Nulidade absoluta processo (C23)	150	20,75	Grounds	Not released
6	Prevista art 44 Lei n 11343 (C24)	150	17,22	Reason	Not released
7	Dezembro 2014 (C12)	152	16,98	Date	Not released
8	Natureza droga apreendida cocaína (C28)	149	10,06	Reason	Not released

We classified the obtained labels as follows: (a) *grounds* of appeal (i.e. elements alleged by the defendant); (b) the *reasons* of the decision (elements indicated by the judges); (c) the type of committed *crime*; (c) the *location* of the lower court; (d) the *date* of the Supreme Court judgment; (e) and the name of the *judge rapporteur*. In analysing the results, we found some difficulties since multiple labels had similar meanings, and certain documents were included in more than one cluster. From the *Not released* subset of the Italian corpus we extracted grounds of appeal such as the nullity of the defendant’s interrogation (label no. 2), the expiration of the pretrial detention term (label no. 3), and the violation of the presumption of innocence principle (label no. 7). Lingo also extracted labels referring to manually identified *objective* factors, e.g., the location (Naples, label no. 6), the date (May 2013, label no. 1) and the crime type (criminal organization, label no. 8). Among the requirements needed to apply the pretrial detention measure, the seriousness of the risks (label no. 4) is also related with maintaining the prison order. From the Brazilian *Not released* subset we extracted similar grounds of appeal, such as the expiration of the pretrial detention term (label no. 2) and the procedural nullity (label no. 5). As reasons for judgment, we listed the victim’s appearance in court (label no. 1), the impossibility of converting the prison into an alternative measure in cases of drug-related crimes (label no. 6), also depending on the nature of the drug

seized (cocaine, label no. 8). Here we also identified manually extracted labels, such as the date (December 2014, label no. 7), the crime (drug law crime) and the judge rapporteur (C. L., label no. 3).

Table 7. Lingo clusters and labels in Italian *Released* subset.

No.	Label and cluster	DN	Score	Type	Outcome
1	L'interrogatorio di garanzia ex art 294 (C5)	12	42,16	Reason	Released
2	Periodi di sospensione di cui all'art 304 (C2)	14	34,52	Reason	Released
3	Sostituzione degli arresti domiciliari (C3)	14	34,52	Grounds	Released
4	Difensore alle ore (C11)	9	29,59	Reason	Released
5	Febbraio 2009 (C6)	11	26,45	Date	Released
6	Doppio dei termini previsti dall'art 303 (C9)	10	26,03	Reason	Released
7	Caso di regressione (C8)	10	24,13	Reason	Released
8	Tribunale di Catanzaro (C12)	8	18,53	Location	Released

Table 8. Lingo clusters and labels in Brazilian *Released* subset.

No.	Label and cluster	DN	Score	Type	Outcome
1	Rio de Janeiro RJ (C2)	57	36.85	Location	Released
2	Constrangimento ilegal decorrente excesso prazo (C5)	52	36.69	Reason	Released
3	Regime inicial aberto requer (C10)	52	33.96	Grounds	Released
4	Impte Defensoria Pública (C3)	57	29.59	Reason	Released
5	Empresas investigadas (C17)	42	22.30	Reason	Released
6	Junho 2017 (C14)	50	20.06	Date	Released
7	Furto insignificante (C21)	9	18.83	Crime	Released
8	G. M. Segunda Turma Habeas Corpus 112 (C12)	51	15.25	Judge	Released

As regards the *Released* outcome, in the Italian subset we can note as related reasons the procedural nullity involving the defendant's hearing (label no. 1) as well as the suspension of the prison term-limit and its expiration (labels no. 2 and no. 6). These reasons can also be framed as grounds, as they were alleged by the defendant. We can further identify the following reasons: the issues concerning the defender (label no. 4), cases returned to the previous grade of judgement (label no. 7), and the replacing imprisonment with less restrictive measures (house-arrest, label 3). Once again, we verify factors regarding the date (February 2009, label no. 6) and the location (Catanzaro Court, label no. 8). We also found similarities in the Brazilian *Released* subset in terms of judgment reasons and grounds of appeal, such as the expiration of the prison term and unlawful constraint (label no. 2), less restrictive measures (label no. 3) and appeal proposed by the public defender (label no. 4). Cases related to investigated companies are also a factor that we classified as a reason (label no. 5). Other labels verified are when the situation involves an insignificant burglary (crime, label no. 7) and the judge-rapporteur (G. M., label no. 8).

Tables 9, 10, 11 and 12 show some selected results from Hierarchical and LSI.

Table 9. Hierarchical clusters and LSI topics in Italian *Not released* subset.

Topics and cluster	DN	Type	Outcome
(C16) 1: p, art, 2020, comma, n, sospensione, termini, d, p p, 2 2: art 304, 304, termini, p, sospensione, comma, 304 p, p comma, p p, è 3: tribunale, 3, riesame, 304, art 304, periodo, art 309, 309, 309 p, sospensione	11	Grounds/ Date	Not released
(C19) 1: p, n, art, sez, rv, p p, 3, 1, cautelare, comma 2: r, co, cautelare, sentenza, cautelari, esigenze cautelari, esigenze, associazioni, stupefacenti, dott 3: presunzione, art 275, 3, 275, r, interrogatorio, 275 p, co, comma, comma 3	27	Reason/ Crime	Not released

Table 10. Hierarchical clusters and LSI topics in Brazilian *Not released* subset.

Topics and cluster	DN	Type	Outcome
(C12) 1: hc, habeas, art, corpo, habeas corpus, ministro, min, tribun, prisão, voto 2: lei, art, pena, liberdade, tráfico, provisória, liberdade provisória, turma, crime, droga 3: pena, provisória, liberdade, prisão, liberdade provisória, 33, art 33, regime, 4°, senhor	235	Crime	Not released
(C21) 1: crime, n°, lei, ministro, voto, tribun, habeas, turma, marco, corpo 2: crime, código, criminosa, organização criminosa, lei, organização, s, art, sob código, código senha 3: habeas, habeas corpus, corpo, crime, lavagem, n°, acórdão, relat, delito, dinheiro	28	Crime	Not released

LSI returns green and red words, respectively indicating positive and negative weights. A positive weight indicates that a word is highly representative of a topic, while a negative weight indicates that a word is highly unrepresentative for that topic [6]. We tried to either lower or increase the number of topics with no real impact on the overall intelligibility of the results. Hence, the disadvantage of combining Hierarchical clustering and LSI is that we had to interpret single words rather than strings.

In the Italian *Not released* subset, we could identify factors already identified with Lingo, e.g., the suspension of the prison term and its expiration as grounds (C16 topics) on the one hand, and the seriousness of precautionary requirements, the connection between criminal organizations and drug-related crimes as a reason for applying pretrial detention (C19 topics) on the other hand. This factor can also be observed in the Brazilian *Not released* subset (C12 and 21 topics).

In the Italian *Released* outcome, we can observe similar results to those obtained with Lingo. In particular, we identified a few words referring to the hearing of the defendant, the general requirements for applying a precautionary

Table 11. Hierarchical clusters and LSI topics in Italian *Released* subset.

Topics and cluster	DN	Type	Outcome
(C7) 1: p, art, p p, n, comma, cautelare, misura, 1, 2, ordinanza 2: sentenza, appello, fase, interrogatorio, cort, misura, grado, p, pena, p p 3: misura, 2, interrogatorio, bi, pena, comma, art 275, 275, comma 2, carcer	54	Reason	Released
(C4) 1: p, art, comma, cautelare, custodia, n, 1, custodia cautelare, p p, termini 2: art, termin, 1, comma, termini, p, fase, art 1, sentenza, durata 3: misura, custodia, termin, p, sospens, 1, termini, giudic, custodia cautelare, sentenza	16	Reason	Released

Table 12. Hierarchical clusters and LSI topics in Brazilian *Released* subset.

Topics and cluster	DN	Type	Outcome
(C16) 1: prisão, min, cautelare, hc, penal, liberdade, c., m., c. m., rel 2: direito, art, prazo, prisão, cautelare, rs, excesso, preventiva, prisão preventiva, duração 3: pena, liberdade, lei, prazo, n ^o , privativa, sp, pena privativa, penal, privativa liberdade	20	Reason/ Judge	Released
(C5) 1: hc, min, prisão, turma, art, sp, habeas, corpus, habeas corpus, ministro 2: liberdade, turma, lei, art, m., c., c. m., dje, liberdade provisória, provisória 3: primeira, primeira turma, g., m., g. m., prisão, domiciliar, min g., turma, prisão domiciliar	33	Reason/ Judge	Released

measure (C7 topics), and the prison time expiration (C4 topics). In the Brazilian subset we identified a set of words referring to the time-limit of the prison (C16 topics), and house arrest as an alternative measure (C5 topics). Moreover, the algorithm extracted the name of two judges that are related to the release outcome (C16 and C5 topics).

5 Conclusion and Future Works

It is well known that the Brazilian and Italian Supreme Courts usually maintain, rather than reform, decisions on pretrial detention by lower courts. In our experiments, we aimed to go beyond this obvious observation and analyse the reasons behind such decisions. This may help us in determining whether this practice is legally correct or rather reflects the reluctance to overhaul decisions by lower courts. While our analysis does not provide a definitive answer, it shows a certain consistency in high court decisions. In both legal systems, clustering labels and topics point to factors in common, i.e., the excessive length of time spent in prison, and the time-limits established by the law are factors which support the release. On the other hand, crimes against property, drug-related crimes and involvement in criminal organizations are highly related to the maintenance of the pretrial detention measure. The same is true with regard to the locations of Naples and Sao Paulo, suggesting that in these places serious crimes are more

recurrent. In the Brazilian dataset, we found relationships between the judicial outcome and the judge rapporteur. This situation is absent in the Italian dataset. This may be due to the higher variability of judges in this Court. Concerning the experimented methods, Lingo performs better than the Hierarchical clustering combined with LSI. Labels are immediately intelligible and contain meaningful information, from both computer science and a legal perspective. Moreover, the topics resulting from LSI could not be as easily linked to any relevant legal circumstance.

Future research includes structuring a dataset based on the factors highlighted and performing classification experiments through deep and classical machine learning to predict the outcome. In this sense, we also aim to obtain explanation of the predictions through the extracted factors.

References

1. Aggarwal, C.C.: Machine Learning for Text. Springer, Cham (2018). <https://doi.org/10.1007/978-3-319-73531-3>
2. Bex, F., Prakken, H.: On the relevance of algorithmic decision predictors for judicial decision making. In: Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, pp. 175–179 (2021)
3. Bojanowski, P., et al.: Enriching word vectors with subword information. *Trans. Assoc. Comput. Linguist.* **5**, 135–146 (2017)
4. Cichosz, P.: Data Mining Algorithms: Explained Using R. Wiley, Chichester (2015)
5. Dal Pont, T.R., et al.: Classification and association rules in Brazilian supreme court judgments on pre-trial detention. In: Kö, A., Francesconi, E., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) EGOVIS 2021. LNCS, vol. 12926, pp. 131–142. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86611-2_10
6. Demšar, J., et al.: Orange: data mining toolbox in python. *J. Mach. Learn. Res.* **14**(1), 2349–2353 (2013)
7. Duff, R.: Pre-trial detention and the presumption of innocence. Oxford University Press, Minnesota Legal Studies Research Paper 12-31 (2012)
8. Han, J., Pei, J., Yin, Y.: Mining frequent patterns without candidate generation. *ACM SIGMOD Rec.* **29**(2), 1–12 (2000)
9. Horty, J.: Reasoning with dimensions and magnitudes. *Artif. Intell. Law* **27**(3), 309–345 (2019). <https://doi.org/10.1007/s10506-019-09245-0>
10. Horty, J.F., Bench-Capon, T.J.: A factor-based definition of precedential constraint. *Artif. intell. Law* **20**(2), 181–214 (2012)
11. Hu, X., Liu, H.: Text analytics in social media. In: Aggarwal, C., Zhai, C. (eds.) *Mining Text Data*, pp. 385–414. Springer, New York (2012). https://doi.org/10.1007/978-1-4614-3223-4_12
12. Kotu, V., Deshpande, B.: *Data Science*, 2nd edn. Morgan Kaufmann (Elsevier Science), Cambridge (2019)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) (2013)
14. Osiński, S., Stefanowski, J., Weiss, D.: Lingo: search results clustering algorithm based on singular value decomposition. In: Kłopotek, M.A., Wierzchoń, S.T., Trojanowski, K. (eds.) *Intelligent Information Processing and Web Mining. Advances in Soft Computing*, vol. 25, pp. 359–368. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-39985-8_37

15. Osiński, S., Weiss, D.: Carrot2 (2019)
16. Papadimitriou, C.H., et al.: Latent semantic indexing. *J. Comput. Syst. Sci.* **61**(2), 217–235 (2000)
17. Sebastiani, F.: Machine learning in automated text categorization. *ACM Comput. Surv. (CSUR)* **34**(1), 1–47 (2002)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

