vol. 9, 2022,
DOI: 10.19092/reed.v9.600

SUMARIAÇÃO DE TEXTOS COMO FERRAMENTA DE PESQUISA EMPÍRICA EM DIREITO BASEADA EM ALGORITMOS DE APRENDIZADO DE MÁQUINA

Danilo Carlotti[1]
João Eduardo Ferreira[2]

RESUMO

Este artigo propõe explorar técnicas de sumariação de textos jurídicos como forma de auxiliar a pesquisa empírica, gerando um sumário do texto tendo em vista a capacidade preditiva destas frases em relação ao resultado da ação. É utilizado um dataset de decisões de tribunais do país sobre habeas corpus que expressamente citam a pandemia de COVID como um dos seus fundamentos para solicitar a liberdade dos pacientes. É criado um modelo preditivo e expõe-se, ao final, os argumentos encontrados que tem maior correlação com o resultado das ações.

PALAVRAS-CHAVE: aprendizado de máquina; direito; sumarização de textos.

[1] Universidade de São Paulo, ORCID
[2] Universidade de São Paulo, ORCID

vol. 9, 2022
DOI: 10.19092/reed.v9.600

TEXT SUMMARIZATION AS AN EMPIRICAL LEGAL RESEARCH TOOL BASED ON MACHINE LEARNING ALGORITHMS

Danilo Carlotti
João Eduardo Ferreira

ABSTRACT

This paper aims to explore text summarization techniques as a tool for empirical legal research, creating a summary of the decisions given the phrases predictive power with regard to the decision outcome. A dataset of habeas corpus decisions prompted by innumerable courts in Brazil is used that explicitly cite the COVID pandemic as a reason for requesting the release of the patients. A predictive model is created and through this analysis we propose to find the arguments most correlated with the outcome.

KEYWORDS: machine learning; law; text summarization.

1. INTRODUCTION

Information in the legal domain is primarily made available as text. Therefore, it is important to develop tools that facilitate the exploration of large corpora of legal documents outstandingly. In the process of decision making, at least in countries with a considerable amount of precedents likewise Brazil, it may be necessary to sort through thousands of precedents to find the most suitable to define your legal strategy.

A fundamental scheme is to determine what are the main arguments in a corpus of texts and how it is possible to use these arguments in one's advantage. Finding the arguments more correlated with the outcome might also help the courts further homogenize precedents and discourage litigation.

One growing field is argument mining (Lippi & Torroni, 2015) that proposes the identification of argumentative sentences, their parsing and establishing the relationship between statements. The objective of this research, however, is not limited to identify in this phase the arguments but to predict the collection of potential texts that indicate result to some extent. It is not necessary for these phrases to be arguments, since they can be simply references to precedents or other aspects of the legal procedure that can reveal patterns of reasoning in the jurisprudence. The main purpose of the paper aims to present a technique of summarization as a tool for empirical legal research, helping a practitioner to better understand the arguments most correlated with the outcome. This is a similar purpose to that of general summarization, however, it has the outcome as the main focus, instead of a range of gold standard summaries.

Text summarization is an important field of research in natural language processing. A review of current techniques and success cases can be found in Yogan (2016). As presented in the related works section, summarization of legal documents presents important differences from general text summarization: the language is technical with a particular vocabulary, documents tend to have a common structure (with exceptions), there are logical constraints inherent to a decision and there are pieces of information, as legal precedents, important in themselves that can be extracted. Therefore, it is imperative that methods to summarize legal documents address these particularities.

We propose that text summarization is an important asset in empirical legal research, combining the predictive power of machine learning algorithms with knowledge of the legal field. The text is reduced to collections of phrases that are likely to inform what is considered important in the decision making of judges. The task of summarization is achieved by breaking the text in different logical parts and by reducing the redundancy of phrases in the summary by comparing their relative distances, after processing.

This document is organized as follows: in the related work section it is presented a revision of literature in the field; in the data and methods section we present how the data was collected and processed, leading to the distillation of the texts; in the section results and discussion the empirical and qualitative results are presented and discussed; in the conclusion we present our considerations to the impact text summarization has in the field of empirical legal research.

2. RELATED WORK

The literature has discussed why legal texts are different than other texts (Kanapala, Pal & Pamula, 2019). Decisions have logical parts that are organized in such a way that lawyers, the parties and other judges should be able to reconstruct: the problem the parties wish to solve through the lawsuit; their legal arguments and how the judge evaluates them; and finally the verdict.

The size of the legal documents tends to be longer than documents in other domains because many domains still depend on collections of abstracts rather than the full text of the documents. Legal documents show a different internal structure. They have status and administrative codes and follows hierarchical structure. The vocabulary of legal texts is different. Legal language uses a number of domain specific terminology besides the standard language. Legal text may be ambiguous as there can be multiple different meaning for the same term, phrase, statement. The same text could be interpreted differently if it occurred in high court opinion than if it occurred in the opinion issued by a district courts. Citations play a prominent role in legal domain than they do in other domains and generally they also highlight issues of the case. These differences are taken into account in the

proposed algorithm for summarization or in the techniques for information extraction.

Lloret & Palomar (2012) state that the purpose of summarization is to extract sentences of interest that inform the most about the texts with extractive and indicative techniques. There are statistical techniques that consider the frequency of words, in numerous combinations, to determine what are the most common words or documents. There are approaches that use "topics", classifying the phrases and assigning them following importance. There are graph-based approaches that model the text as relations between words, the nodes, and their possible relationships, the edges. The final approach presented is machine-learning based so that the phrases are ranked with models trained in corpus such as Wikipedia. To evaluate the summary, it is commonly used a gold standard previously annotated, and the machine summaries are compared with rouge and f-score.

Nenkova & McKeown (2012) resume the task of summarizing legal texts in the following three steps: "creating an intermediate representation of the input which captures only the key aspects of the text, scoring sentences based on that representation and selecting a summary consisting of several sentences". The selection of main phrases proposed in this paper mixes the importance of each sentence in the text with their correlation with the outcome.

Nenkova & McKeown (2012) present what they call "intermediate representation" of texts. The possible representations discussed are: "topic representation" that reduce the text to a collection of topics, tokenization of texts with simple word frequency or TF-IDF (term frequency inverse document frequency). The proposed algorithm represents the texts as a vector with TF-IDF.

Regarding the selection of phrases, Nenkova & McKeown (2012) state: "In global selection approaches, the optimal collection of sentences is selected subject to constraints that try to maximize overall importance, minimize redundancy, and, for some approaches, maximize coherence". This is done, we propose, by correlating each phrase with the outcome by a predictive model and the redundancy is minimized by selecting the most unique phrases.

The ranking of phrases has already been used to summarize legal documents in Sentences and Words from Alternating Pointer Networks (SWAP-

NET) (Jadhav, Rajan, 2018), in which they create "Extractive summaries comprising a salient subset of input sentences, often also contain important key words". The method used to explore the relationship with certain words and sentences. This can be helpful in order to summarize certain concepts, but it also needs the options of keywords to be previously selected.

Templeton, Kalita (2018) explored the idea that the phrases that are most similar to all others should be ranked higher. The problem with this approach is that the proposed algorithm chooses the most common phrases, not the most important, since there are common phrases without any importance or meaning that may convey information about: lawsuit costs, location of the lawsuit, procedural arguments, among others.

An important step in the summarization of legal texts is the segmentation of the decision in categories. This can be done in an unsupervised manner (Alguliev 2009) or with supervised learning (Teufel 1997; Hachey 2006; Yousfi-Monod 2010). This aims to split the text and contemplate the different logical parts in the summary. Hachey e Grover (2016) propose the following classes: facts, proceedings, background, framing, disposal, textual and others.

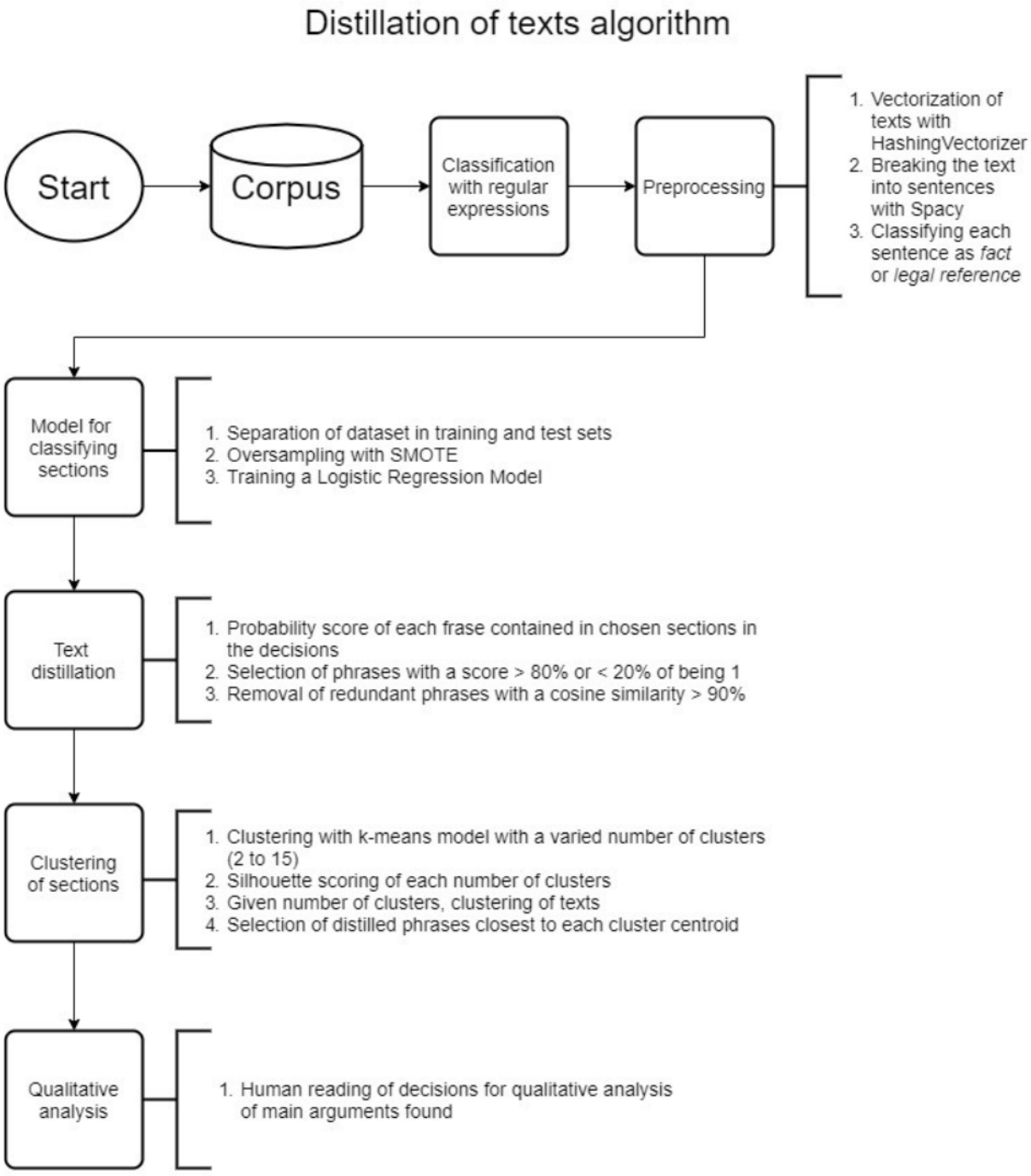
The main objection made to these papers is that their objective in the summarization process is to mimic the proposed gold standard of summaries made specifically in those datasets. We propose a more generic approach that also serves the purpose of empirical legal research since it tries to build a summary to narrate the lawsuit as it also chooses the phrases most correlated with the outcome, given a predictive model.

3. DATA AND METHODS

The diagram in Image 1 illustrates all steps proposed in this paper for the distillation of texts and selection of the best phrases for the study of the arguments most correlated with the outcome.

Image 1

Flowchart of summarization algorithm



DATA

The corpus of decisions was extracted from court publications with crawlers from the following courts: The Court of the States of São Paulo, Brazil's Capital District, Ceará, Alagoas and the Superior Court of Justice (Superior Tribunal de Justiça). We considered only decisions by these courts in habeas corpus that contained words related to the covid pandemic ("covid", "pandemia" and "coronavírus") and that questioned decisions from lower courts or judges requesting the release of the patients in the year 2020. We found 16126 decisions. It was possible to definitively ascertain the outcome in 9578 decisions, where 91.71% of decisions dismissed the request for liberty. Other decisions were not final or the suit was temporarily dismissed on a technicality.

The corpus was stored in a NoSQL database with the following fields in each document presented in Table 1.

Table 1

Database description

Name	Description
Class	The class of the decision, granting or denyingthe request
Section	Name of the section of the text
Section text	Set of phrases identified with the given section
Vector	The vector representation of the section text
Probability of section	The probability the model assigns to this sectionof belonging to the class of 1's

PREPROCESSING AND CLASSIFICATION METHODS

The decision outcome was determined using regular expressions and to each decision it was assigned a class of 0, if the request was denied, or 1, otherwise.

The corpus of texts was broken into sentences using the library spacy. Each sentence was classified as facts, legal references, request, ruling or other using a dictionary of regular expressions. The separation of decisions in these logical parts tries to reproduce the main aspects of a decision segregating information about the reasons for the litigation, facts and requests, the legal grounds by which the decision is reached, legal references, the arguments used in the reasoning of the judge or judges, that may be in the aforementioned sections or in other and finally the ruling or the decision reached.

We considered in this study only sentences of the category “facts” or “legal references”, since they are the most informative of the judges valuation of the facts and legal arguments presented. The phrases with the ruling contain, in themselves, the decision’s outcome and the request made by the patients, in these cases, are all the same.

To test for the correlation between sections of phrases and the outcome we trained a supervised classification model to predict, given a text, what is the probability of the decision, with that section, belonging to the class 0 or 1.

To train the classifier each section of texts was vectorized using the Hashing Vectorizer, implemented in the python scikit-learn library with 25000 positions. Each text was then represented as a vector and stored in the database. Of all 9578 texts, only 6734 had sections of text bigger than three sentences. These were the only texts used in this part of the study because sections too small are not sufficiently informative and may contain only noise. The predictive models of facts and legal references used Logistic Regression to determine the probability of the classes of each data point. Various machine learning models were tested and this was the model with the best performance. There are no embedding models trained with a corpus of portuguese legal texts publicly available, preventing the use of other techniques that use them as input. To train these models all data points were divided in test (25%) and training sets (75%). Since the data is unbalanced, we used SMOTE to oversample the data and balance the training dataset. SMOTE is a technique that uses k nearest neighbors to create new data points similar to the ones in the least sampled class (Chawla et al, 2002). The models were exposed only

to the training sets in the learning phase. Afterwards, the models were tested using the test dataset and the scores are presented in the next section.

After the training, the models are able to assign to each section a probability that they belong to class 0 or 1.

To select the best phrases for each text we distilled the texts in two phases. In the first, we selected only the phrases in each section of each text with a probability score of over 80%, high chance of success, or under 20%, high chance of failure. In the second phase we discarded phrases too similar, removing redundancies. The vector representation of all phrases selected in the first phase were compared by their cosine similarity, creating a distance matrix. In all cases in which two vectors had similarity greater than 0.9 the second one considered was discarded. The remaining phrases were clustered using scikit-learn's implementation of k-means. The number of clusters was varied between 2 and 15. The best number of clusters found was 9 for the class 0 and 11 for the class 1. We propose that this set of sections can best inform the decision maker about the most correlated arguments with the outcome in the dataset.

The qualitative analysis step is comprised of the reading of the selected phrases and a human evaluation of the importance and significance of the arguments presented. Patterns may be apparent to the human reader such as: an ideological orientation of judges, preference for a line of reasoning or others. The summary of the decision is then the collection of phrases from each section, after the distillation process. This subset of sentences should be sufficiently informative of why the cases are brought to the courts and the reasoning behind the judges' decisions. The outcome statistics can be obtained separately.

4. RESULTS AND DISCUSSION

MODEL SCORES

The final dataset has 6734 valid points and 91.65% of them are of class 0. The classification of the section of texts by the Logistic Regression model registered the following metrics in Table 2.

Sendo assim, eis o gráfico que representa os dados encontrados:

Table 2

Model scores

Section	Matthew's correlation	Accuracy	Recall	Specificity
Legal references	38.35%	85.57%	65.44%	87.33%
Facts	34.06%	81.65%	68.38%	82.81%

The model obtained a sufficiently high score that justifies its use for exploratory purposes.

After training, it was possible to predict the probability of each section pertaining to each class. This demonstrates the level of certainty of the algorithm regarding each section of each text. The results follow in Image 2 and 3. On the x-axis there are all sections of all decisions and the y-axis is the probability of this section belonging to the class 1.

Image 2

Scoring for legal references section of dataset

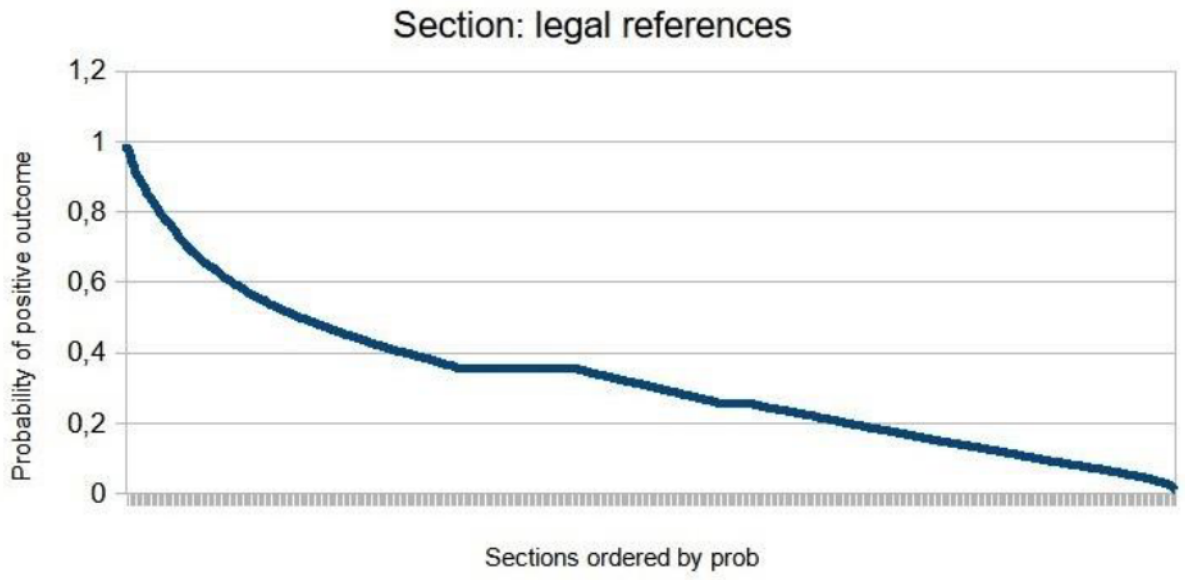
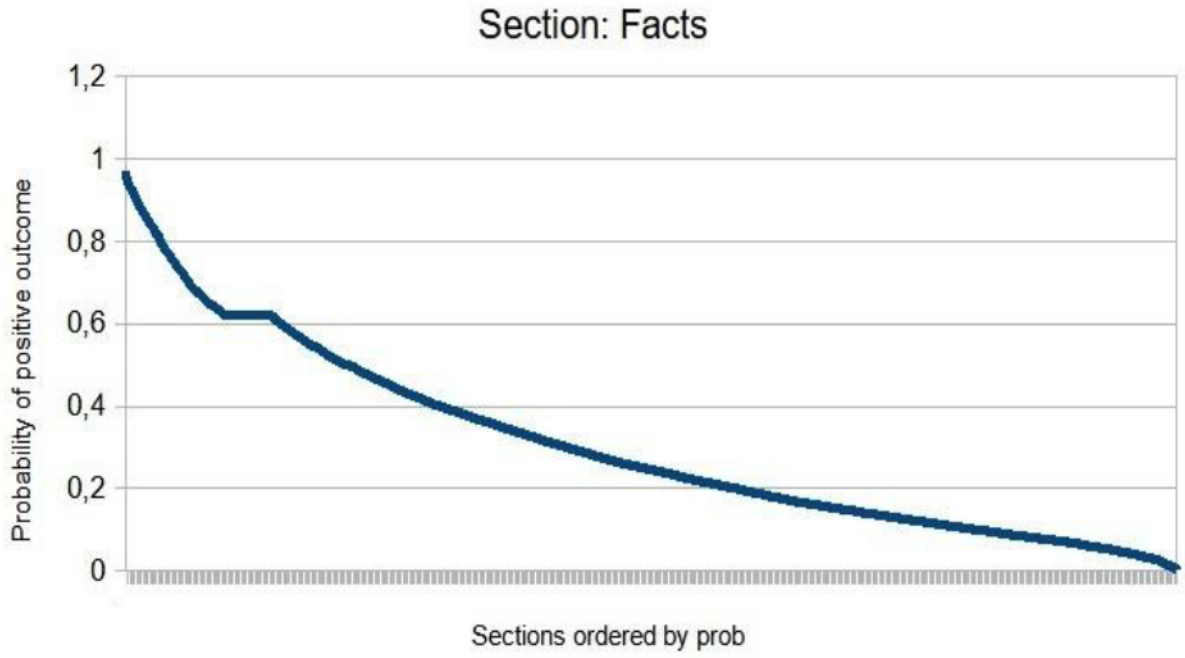


Image 3

Scoring for facts section of dataset



COMMENTARIES ON THE DISTILLATION OF TEXTS

There are two phases of the distillation process. In the second phase the algorithm removes redundant phrases. It is possible to note some improvement in the second phase, but the main difference is the fact that with a smaller dataset of phrases it is easier to analyze them manually.

In the phrases selected in the first and second phase of distillation there were two decisions which were summed in two phrases that described the public prosecutor's office of the State of São Paulo, the biggest litigator due to population size.

The other summaries were all informative and contained principles used by the judges to reject or grant the request for liberty. COVID-19 had some importance in the decisive arguments, but it was not universally accepted or rejected. It seems that COVID19 was only relevant if the patient in the habeas corpus showed decisive proof of risk and even so it may be subject by judges evaluation, since the decisions do not mention or reproduce documents presented by the plaintiffs that could automatically qualify them for release according to the courts.

These are the main arguments found in the first and second phase of text distillation:

- Generic justifications that the lower court did not commit any illegality as justification for rejecting the request, which is really common in these cases in Brazil;
- Considerations of the crime committed, drug traffic, and the danger of the person to society as justification for denying release; Considerations of the crime committed, a smaller offense, and this as the main reason for the release of the patient.

These are arguments found exclusively in the first phase:

- There is explicitly mention of the criteria needed for the patient's release: the patient is specially vulnerable to covid; the person is not able to receive treatment in the prison system; and that the danger of release is smaller than keeping the person locked up;
- Procedural justifications, specially the binding precedent number 691 from Brazil's Supreme Court, that specifies that a higher court

cannot decide on the preliminary decision of a lower judge, before the final judgment;

- In one case, the health crisis is considered as a reason for reconsideration, even though the person is accused of drug trafficking.

These are arguments found exclusively in the second phase:

- There is a case which explicitly takes in consideration the patient's age and vulnerability to the disease as a reason for granting the request;
- The lack of proof of vulnerability is considered as the main argument for denying release;

Using regular expressions, it was possible to extract the main legal references quoted in this dataset. We present the top 5 below in Table 3.

Table 3

Legal references parsed from the dataset

Legal reference	Description	Frequency
Federal law number 11.419/2006	Regulates the digitalization of law-suits in Brazil	4202
Binding precedent 691 –Supreme Court	States that the supreme court will not admit *habeas corpus* that has not been judged in final bya lower court	1426
Statute of Penal Proce-durearticle 312	Establishes the criteria to hold somebody inprovisional custody	550
Statute of Penal Proce-durearticle 319	Establishes alternative measures other thanprison	436
Federal law number 11.343/2006	Law that criminalizes drug traffick-ing	354

It is possible to extract information from these texts using supervised learning or simply searching for keywords. For example, there are 1955 precedents that explicitly mention drug trafficking ("tráfico de drogas") and the request for liberty was granted in 13,58% of them. If a criminal lawyer could read the main excerpts of the decisions that accepted the plaintiff's request it would be possible to have better insights into possible bias of the courts or the arguments with the biggest likelihood of success in these specific cases. A series of other selections could be made, with or without machine learning, to delimit even further the set of decisions of interest.

This is the biggest contribution of this method, the ability to stack layers of information extraction and text distillation to evaluate arguments and better understand, for each court, time period, region, subject or any other desired filter, what are the most iconic decisions and arguments, given the final outcome.

5. CONCLUSION

The impact for empirical legal research is to use summarization as a research tool and also as a non trivial classifier for legal decisions. This can help the legal practitioner extract the arguments or strategies most commonly correlated with the outcome.

The classification of texts by their logical helps to detect correlation of reasoning and outcomes by different courts and how they evaluate different facts to reach their final decisions. These are the arguments most correlated, according to the model, with the outcome. In this sense, these are the most important arguments that should be considered when formulating a legal strategy, even though they do not prove a causal relation between arguments and the decision outcome.

The extraction of legal references is also an important tool to assess which precedents are most discussed in the decisions. In this case, it can be noted that the criminal law most explicitly present in the decisions is drug trafficking, that is also the crime most successfully prosecuted in Brazil. But through information

extraction it is possible to create an ensemble of variables such as decision subject, leading judge, express citation of precedent A or B, express citation of federal agency J, among others.

With these tools at the disposal of lawyers and judges tendencies may emerge and help in the decision making process. The next step in the research is to further test ways of segmenting legal decisions to compare judges' preferences in different areas and eventually use information about the decision making process in one field as a tool for predicting the outcome in another field.

REFERÊNCIAS

Alguliyev, R., Aliguliyev, R. (2009). Evolutionary Algorithm for Extractive Text Summarization. *Intelligent Information Management*. 1. 128-138. doi: 10.4236/iim.2009.12019.

Chawla, N., Bowyer, K., Hall, L., Kegelmeyer, W. (2002). SMOTE: Synthetic MinorityOver-sampling Technique. *J. Artif. Intell. Res.* (JAIR). 16. 321-357. doi: 10.1613/jair.953. Hachey, Ben & Grover, Claire. (2006). Extractive summarisation of legal texts. *Artificial Intelligence and Law*. 14. 305-345. doi: 10.1007/s10506-007-9039-z.

Kanapala, A., Pal, S., Pamula, R. (2019). Text summarization from legal documents: a survey. *Artificial Intelligence Review*. 51. doi: 10.1007/s10462-017-9566-2.

Group, Lazuage & Moens, Marc. (2002). Sentence Extraction as a Classification Task. Jadhav, Aishwarya & Rajan, Vaibhav. (2018). *Extractive Summarization with SWAP-NET: Sentences and Words from Alternating Pointer Networks*. 142-151. doi:10.18653/v1/P18-1014.

Lippi, Marco & Torroni, Paolo. (2015). *Argument Mining: A Machine Learning Perspective*. 9524. 163-176. doi: 10.1007/978-3-319-28460-6_10.

Lloret, E., Palomar, M. (2012) Text summarisation in progress: a literature review. *ArtifIntell Rev* 37**,** 1–41. doi: 10.1007/s10462-011-9216-z

Nenkova, A., Kathleen, M. (2012) A survey of text summarization techniques, *Mining text data*, Springer-Verlag New York, doi: 10.1007/978-1-4614-3223-4

Templeton, A., Kalita, J. (2018) "Exploring Sentence Vector Spaces through Automatic Summarization," *17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Orlando, FL, USA, 2018, pp. 55-60, doi: 10.1109/ICMLA.2018.00016.

Teufel, S., Uk, S., Moens, M. (2001). *Sentence Extraction as a Classification Task*.

Yogan, J.K., Goh, O.S., Halizah, B., Ngo, H.C., Puspalata, C.S. (2016) A Review On Automatic Text Summarization Approaches. *Journal Of Computer Science*, 12 (4). pp. 178-190. ISSN 1549-3636

Yousfi-Monod, Mehdi & Farzindar, Atefeh & Lapalme, Guy. (2010). Supervised Machine Learning for Summarizing Legal Documents. 51-62. 10.1007/978-3-642-13059-5_8.

Danilo Carlotti: Cientista de dados, é pós-doutorando em ciência da computação no Instituto de Matemática e Estatística da Universidade de São Paulo.

João Eduardo Ferreira: Professor Titular do Departamento de Ciência da Computação no Instituto de Matemática e Estatística da Universidade de São Paulo.

Data de submissão: 19/04/2021.

Data de aprovação: 29/06/2021.