











## Article

# Development and Evaluation of an Intelligence and Learning System in Jurisprudence Text Mining in the Field of Competition Defense

Edna Dias Canedo <sup>1,2,†</sup>, Valério Aymoré Martins <sup>2,†</sup>, Vanessa Coelho Ribeiro <sup>2,†</sup>, Vinicius Eloy dos Reis <sup>3,†</sup>,  
Lucas Alexandre Carvalho Chaves <sup>2,†</sup>, Rogério Machado Gravina <sup>2,†</sup>, Felipe Alberto Moreira Dias <sup>3,†</sup>,  
Fábio Lúcio Lopes de Mendonça <sup>2,†</sup>, Ana Lucila Sandoval Orozco <sup>2,4,\*,†</sup>, Remis Balaniuk <sup>5,†</sup>  
and Rafael T. de Sousa, Jr. <sup>2,†</sup>

- <sup>1</sup> Department of Computer Science, University of Brasília (UnB), P.O. Box 4466, Brasília 70910-900, Brazil; ednacanedo@unb.br
- <sup>2</sup> Electrical Engineering Department, National Science and Technology Institute on Cyber Security, University of Brasília (UnB), P.O. Box 4466, Brasília 70910-900, Brazil; valeriomartins@unb.br (V.A.M.); vanessa.ribeiro@redes.unb.br (V.C.R.); lucas.alex@gmail.com (L.A.C.C.); rogerio.gravina@redes.unb.br (R.M.G.); fabio.mendonca@redes.unb.br (F.L.L.d.M.); desousa@unb.br (R.T.d.S.J.)
- <sup>3</sup> General Coordination of Information Technology (CGTI), Administrative Council for Economic Defense (CADE), Brasília 70770-504, Brazil; vinicius.reis@cade.gov.br (V.E.d.R.); felipe.dias@cade.gov.br (F.A.M.D.)
- <sup>4</sup> Group of Analysis, Security and Systems (GASS), Department of Software Engineering and Artificial Intelligence (DISIA), Faculty of Computer Science and Engineering, Office 431, Universidad Complutense de Madrid (UCM), Calle Profesor José García Santesmases, 9, Ciudad Universitaria, 28040 Madrid, Spain
- <sup>5</sup> Graduate Program in Governance, Technology and Innovation, Universidade Católica de Brasília (UCB), Taguatinga, Brasília 71966-700, Brazil; remis@p.ucb.br
- \* Correspondence: asandoval@redes.unb.br or asandoval@fdi.ucm.es; Tel.: +55-61-98114-0478
- † These authors contributed equally to this work.



**Citation:** Dias Canedo, E.; Aymoré Martins, V.; Coelho Ribeiro, V.; dos Reis, V.E.; Carvalho Chaves, L.A.; Machado Gravina, R.; Alberto Moreira Dias, F.; Lopes de Mendonça, F.L.; Orozco, A.L.S.; Balaniuk, R.; et al. Development and Evaluation of an Intelligence and Learning System in Jurisprudence Text Mining in the Field of Competition Defense. *Appl. Sci.* **2021**, *11*, 11365. <https://doi.org/10.3390/app112311365>

Academic Editor: Arcangelo Castiglione

Received: 16 October 2021  
Accepted: 25 November 2021  
Published: 1 December 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Abstract:** A jurisprudence search system is a solution that makes available to its users a set of decisions made by public bodies on the recurring understanding as a way of understanding the law. In the similarity of legal decisions, jurisprudence seeks subsidies that provide stability, uniformity, and some predictability in the analysis of a case decided. This paper presents a proposed solution architecture for the jurisprudence search system of the Brazilian Administrative Council for Economic Defense (CADE), with a view to building and expanding the knowledge generated regarding the economic defense of competition to support the agency's final procedural business activities. We conducted a literature review and a survey to investigate the characteristics and functionalities of the jurisprudence search systems used by Brazilian public administration agencies. Our findings revealed that the prevailing technologies of Brazilian agencies in developing jurisdictional search systems are Java programming language and Apache Solr as the main indexing engine. Around 87% of the jurisprudence search systems use machine learning classification. On the other hand, the systems do not use too many artificial intelligence and morphological construction techniques. No agency participating in the survey claimed to use ontology to treat structured and unstructured data from different sources and formats.

**Keywords:** jurisprudence search system; public administration; indexing; artificial intelligence; machine learning; ontology

## 1. Introduction

Jurisprudence is a set of interpretations of laws and decisions made by courts regarding similar cases. Thus, it indicates a common and preminent understanding of the judges of a particular court about a set of events, assisting them in making decisions in similar future

cases. In addition, access to the jurisprudence of the courts is also essential for lawyers to appropriately target their arguments and defenses [1].

A majority of the decisions are available for consultation through the Internet. However, a large amount of jurisprudence is published annually in a way that makes it difficult to search for decisions that have already been handed down and that may affect the judicial system [1]. Therefore, it is common to use tools capable of performing jurisprudential searches, which search for similar situations in a source of judicial decisions, resulting in a set of similar processes that can serve as a basis for various legal activities [2].

Due to the high volume of decisions and the lack of a unified system, the agencies of the Brazilian public administration use their search tools to retrieve information and allow the return of correlated documents with the criteria used for the search. However, the results are obtained based on the presence of exact terms (not retrieving content with equivalent terms) and not sorted by semantic relevance criteria [1]. As a result, the used search engines return extensive results, containing many documents and sometimes unrelated to what the users expected. Therefore, the search systems must be improved according to the needs of the end-users so that we can obtain more effective results.

The Brazilian public agencies need to define a standard for jurisprudence and seek tools and technologies that allow greater precision in the argumentation of decisions and judgments. Constâncio [1] used ontologies to develop the jurisprudence system of the Brazilian public agencies. The author described the construction of an ontology of semantic search engine development, which identifies concepts and ideas that constitute judicial decisions, called *OntoLegis*. The ontology was built based on different existing ontologies, such as the *JurisTJPR* of this Court and the *Legal Thesaurus* of the Superior Court of Justice, and is composed of more than 10,000 classes characterized by more than 12,000 labels.

Bourget and Costa [3] proposed a legal ontology for jurisprudence construction by analyzing ontologies from similar domains, redesigning a possible controlled vocabulary and a set of axioms to cover the repositories. The authors presented a computational and legal ontology model made for state jurisprudence (Paraná State) called *JurisTJPR* and, from its analysis, designed an ontology for federal jurisprudence.

Calheiros and Monteiro [4] identified the new search systems implemented by the bodies of the Regional Labor Court of the 23rd Region (TRF23), the Mato Grosso State Court of Justice (TJ/MT), the National Council of Justice (CNJ), and the Federal Regional Court of the 2nd Region (TRF2). The search tools allowed it to better document organization and more efficiently retrieve them, which is meaningful for defending individual rights and guarantees. The courts developed their tools, except for the TRF23, which adapted the search system implemented and made available for use by the Federal Court of Accounts (TCU) to optimize its jurisprudential search tool.

Due to the diversity of the existing systems in this scenario, it is essential to identify the existing Jurisprudence systems and what technological solutions they use to perform document classification and selection. Therefore, the goal of this study is to identify which Brazilian agencies use a Jurisprudence search system and to map if they developed it by themselves or the usage of third-party software. In addition, to collect the public agencies' perception of their Jurisprudence Search system and the advantages and challenges of the used systems. To address the identified research problem we have conducted a literature review and a survey with Brazilian public administration agencies to investigate the usage of a textual information retrieval system. Additionally, the identification of Artificial Intelligence (AI) techniques in jurisprudence, documents, and legislation used by the Brazilian administration and understanding the functionalities of these systems and if they use AI in their solutions.

As a result, we hope that our efforts can support IT professionals working in system development in the context of Jurisprudence in identifying the solutions used in the industry to perform the classification and grouping of documents according to a specific interest. Based on the proposed classification, it will be possible to select Jurisprudence

documents in similar situations, thus reducing the effort of development teams in the elaboration of new legal documents.

Our main findings were: (1) Most agencies use a textual database processing system; (2) the indexing engine most used is Apache Solr. In addition, the Java programming language is the most used in developing textual database processing systems in Brazil; (3) most Brazilian agencies do not use Artificial Intelligence in their solutions, and (4) less than 15% of the public administration agencies in Brazil comply with the Brazilian General Data Protection Law (LGPD).

We organized the rest of the work as follows: Section 2 introduces the background, related works compared to this study, and the case study. Section 3 presents the methods we have employed to conduct the research. Section 4 provides answers to each research question and discusses some of our findings and the implications of this research. Section 5 discusses some limitations and threats to validity. Finally, Section 6 concludes our work and presents directions for future works.

## 2. Background

A Jurisprudence system is a solution that makes available to its users a set of collegiate or court decisions, i.e., the recurring understanding of [5] decisions, as a way of understanding the law. Jurisprudence consists of the similarity of legal decisions that provide stability, uniformity, and some predictability of the analysis of a decided case. A Jurisprudence system enables the search of documents related to the topic in reference collections and databases internal and external to a given organization. Generally, resources and technologies are used in the development of a search system, such as: Facets [6], indexing [7], ontology [8], Text Mining [9,10] and natural language processing (NLP) [10].

Ontology is a taxonomy-based knowledge representation model used to present, describe and express a specific domain. Collecting the terms of a domain, as well as specifying its structure, is of great importance and one of the essential parts of an ontology [11]. Ontologies organize and structure information that describes a domain to make it understandable by all interested parties. Ontologies can establish interconnections between Information Systems when they share or make available parts or all of it for any purpose [12].

For purposes of definition and conceptual limits in the scope of knowledge management, Martins [13] defined a taxonomy as a structuring and hierarchical element, which classifies and characterizes the classes and subclasses used in the constructions of an ontology. Thus, taxonomies work towards organize the information, while ontologies seek to establish semantic relationships between concepts (classes), which assigns characteristics (properties) to the terms (attributes). The essential components of an ontology are [13]:

- Classes: Sets, collections, concepts, programmable classes, types of objects or things, organized in a taxonomy;
- Relationships: Represent the type of interaction between concepts or describe adjectives or qualities of classes;
- Axioms: Used to model always true sentences (constraints);
- Instances or individuals: Used to represent specific elements of the classes, that is, the data itself.

An ontology supports knowledge sharing and reuse by proposing its semantics for the various subject areas. Due to the structural and formal support of domain schema representations, ontologies enable the automation of structured and unstructured data processing [8], therefore is thus the core of the Semantic Web. Ontologies are considered as an alternative to solve data heterogeneity problems.

Indexing is the process of searching content from the selection of keywords and concepts for document retrieval. In indexing systems, the automation of this process uses methods that perform word or n-gram extraction as an alternative to keyword indexing, where the index formed points to the documents that contain them [7]. For conceptual

purposes, n-grams are fragments of selected words that bring good search results when used in indexes [7].

For cases of indexing jurisprudence search systems, the indexing and data search tool uses the concept of formation of classification and textually indexed knowledge bases, which allows promoting the consolidation, in the same platform, of the processes judged and decisions taken by the courts, as well as other document collections of interest. This consolidation can support the formation of knowledge of specific jurisprudence and favor the homogeneity and predominance of trends in decision-making in processes with the same content.

Information extraction automatically deriving structured and unstructured data from text, using techniques such as facets. Facets are terms classified and selected from a previously indexed text, in order to facilitate the search process, capable of covering different ranges of values and reflecting some identity of the document [14], i.e., they are textual elements classified to build composite subjects. Therefore, faceted search presented itself as an efficient technique that can significantly reduce the information overload [6] to the user.

Faceted search allows the user to explore a data collection by applying filters in an arbitrary order [15], where the information elements are organized by a classification system using facets and enables the user to elaborate their search progressively, in a refined way, presenting the different choices options and with accurate results [16].

Artificial Intelligence techniques for information retrieval are an essential component in legal science [17]. Artificial Intelligence in such a system is done using Text Mining, and Machine Learning techniques [9]. Machine Learning is a field of study that gives computers the ability to learn without being explicitly programmed. First, training of the model is done by constantly feeding it data, and after that process cross-validation, which allows to estimate the training error and validate the selected data set in the test. In addition, machine learning can be used to extract the parts of a legal document, identify the correlations and generate a document structure file based on a legal ontology [10].

Text Mining is a resource for organizing and structuring data extracted from collections or discovering textual knowledge in databases by natural language processing (NLP) tasks. It generally refers to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents [10]. In the document, information is extracted and converted into structured data, and then knowledge is extracted by parts or fragments of text by combining patterns. Textual structure in NLP is a directional relationship between text fragments, which methods handle to recognize, generate or extract parts of textual expressions and infer the relationship of the parts to the whole [17].

### *2.1. Related Works*

Barros et al. [18] presented a study in which supervised machine learning techniques classified documents related to judicial decisions in order to ascertain the opinion trends that Brazilian courts have. The authors applied a methodology to process the judicial decisions from the Regional Labor Court (TRT) of the 3rd Region, located in the Brazilian state of Minas Gerais, for data mining to extract and process the information present in the judicial documents and made use of natural language to perform the automatic classification of the documents, reaching more than 90% accuracy when indicating the tendency of each judge in a sentence.

Gomes and Ladeira [19] presented a study related to the use of a text search tool of the Superior Court of Justice (STJ) and presented the performance of searches based on Boolean queries with logical and proximity operators. The authors concluded that the court's system could be improved to facilitate the search for decisions already made by the STJ, optimize access to jurisprudence, and follow the evolution of the court's understanding on several themes. The improvement was possible because the Text Retrieval Conference (TREC) technique compares textual similarities. In addition, the authors found that the Best Match 25 (BM25) and Term Frequency Inverse Document Frequency (TF-IDF) models

enabled an improvement in search performance, obtaining better results than semantic models based on prediction such as Word2Vec and Bidirectional Encoder Representations from Transformers (BERT).

Bueno et al. [20] used Artificial Intelligence to assist legal professionals in searching for jurisprudences in quality databases on judicial decisions. The authors' textbase was powered by relevant legal cases and identification of appropriate jurisprudence for retention, with automatic extraction of information from the document into the database, integrated with a thesaurus based on standard legal terms and with retrieval based on similar terms.

Ordoñez et al. [21] presented the PROJLAW application with support for Natural Language Processing (NLP) to analyze the texts that make up a court judgment. NLP and linked used the data for document identification, indexing, and recommendation. After seeking validation of the system through user experience, the application produced answers for the searches performed efficiently and with keyword insertion during the search. The authors concluded that the more keywords used, the greater the search accuracy.

Aletras et al. [22] addressed the use of Artificial Intelligence with natural language processing for the analysis of judicial decisions in building predictive models that reveal patterns that guide judicial decisions in order to be able to predict possible future decisions. The authors proposed to build a tool to predict patterns from the European Convention on Human Rights (ECHR) using the supervised machine learning (SVM) algorithm [23].

Silva et al. [24] presented their research and development project, called *VICTOR*, aimed at solving recognition pattern problems in texts from court cases belonging to the Supreme Court (STF). Differently from previous researches, in this work, the authors proposed a solution to speed up the analysis of judicial decisions directed to the STF and identify which cases are linked to particular subjects of general repercussions, such as competition, price taking, etc., using Convolutional Neural Network (CNN) [24] and Natural Processing Language.

The main difference between existing Jurisprudence Search Systems and the system proposed in this research is that the developed system applied evaluation techniques and iterative redefinitions in the verification and validation of all the functionalities of the proposed solution and used the accessibility and usability guidelines proposed in the literature during the system development process. Thus, we can infer that the developed system follows the best practices used in existing Jurisprudence Search systems. Moreover, one of the differentials of the developed Jurisprudence Search system was that it was submitted to a usability evaluation by four experienced usability experts [5]. Canedo et al. [5] performed the usability heuristic evaluation of the Jurisprudence Search system, using a set of 13 usability heuristics and their respective sub-heuristics, considering the system user, the context of use, the task, and the cognitive load as usability factors [25]. Finally, the Jurisprudence Search system development team added all the suggestions for improvement suggested by the usability experts in the final version of the system made available to the end-users.

Regarding the technological aspects, in the collection and loading processes of structured data, modern resources of early data processing were used, implemented concerning existing documentary resources and external data environments. We can highlight the introduction of statistical concepts in the inference of natural language understanding and discourse analysis to form a supplementary knowledge base about the methodologies and techniques used. In addition, we performed data preprocessing, transformation, and cleaning [26].

Weber et al. [27] defined the concept of Intelligent Jurisprudence Research (IJR) as the activity of performing a jurisprudence search using a computational tool with Case-Based Reasoning (CBR) systems. According to the authors, data retrieval systems that use statistical methods have low accuracy. Thus, the authors consider that the knowledge-based indexing process is more efficient by applying case-based reasoning, an artificial intelligence technique that models aspects of human cognition to solve expert problems. Court cases

are described in natural language, and this makes systematic reading difficult. Therefore, it requires case engineering efforts. The model proposed by the authors converts textual decisions into cases by defining the attributes comprising the issues that best represent the experiences described in the judicial decisions and employing mining methods to extract values for the attributes automatically.

Giacalone et al. [28] proposed a statistical model for text mining on a web database to verify the duration of a trial, the solution adopted by the judge, and its correspondence with other stored decisions. The model was based on a knowledge base and used a hybrid approach to search for text similarities and semantic relations between two concepts. The authors tested the proposed model on a repository containing more than 100 sentences.

Houy et al. [29] developed a system called ARGUMENTUM to search for arguments, justifications, and refutations of statements to analyze judicial decisions. The authors' used techniques from argumentation mining, Support Vector Machines (SVM), Argument Markup Language (AML), and Natural Language Processing (NLP). Pasquale and Cashwell [30] made a critique of the indiscriminate use of prediction techniques in the judicial system and their impact on civil law, questioning the social utility of prediction models when applied to the legal system. The authors stated that using algorithms to perform predictive analysis in judicial contexts is an emergent jurisprudence of behaviorism since it relies on a fundamentally mental process model as a black box of transforming inputs into outputs. Furthermore, in dealing with a system created by humans, the authors stated that predictive analytics could be biased instead of performing informed decision-making since the people affected by automated classification and categorization cannot understand the reason for the decisions that affected them.

Concerning the text mining features we used the Named Entity Recognition (NER) techniques [31], Sentence Breaking and N-Gram treatment [26] performed by removing stopwords [26] and breaking N-Grams [32]. The ontology used morphology analysis [33], ontology formation [34] and ontology analysis [35]. Feature extraction used morphological analysis [33], frequency distribution [36] and prioritization techniques [33]. Finally, we performed clustering analysis [37], topic modeling analysis [38], analysis by classifications [37], and analysis by regressions [39].

The selection of these techniques was to build and expand the knowledge generated regarding the economic defense of competition to support the procedural business activities of the organization. As sub-products of this process, we have the automated revision of controlled vocabularies and resources for structuring semantic and ontological databases.

## 2.2. Case Study

In order to assist the legal process managers of the Administrative Council for Economic Defense (CADE) developed to retrieve information stored in the Electronic Information System (SEI) and other databases, the technical solution named Jurisprudence Search System, which can index and search the information requested by the user within the scope of processes already defined by CADE. Furthermore, the data indexing and search system use the concept of forming textually indexed, classificatory knowledge bases, which allows the consolidation, on a single platform, of judged cases and decisions made by CADE, as well as other collections of documents of interest, in the consolidation of the search to support the formation of specific jurisprudence knowledge. As a result, the system favors the homogeneity and predominance of trends in decision-making by CADE's Commissioners and supports managers in competition matters. During the development of the Jurisprudence System, the project researchers analyzed and determined five important evolutionary axes in the understanding and treatment of the research problem of research and development (R&D), which are:

- Infrastructure, APIS, and Interfaces: It deals with the infrastructure requirements (deployment and configuration), interface, navigation, and Apache Solr [40] handling in its searches;

- Collection, Retrieval, and Indexing: Deals with the specific configuration for the “Jurisprudence” collection;
- Information Structure: Development of a suitable ontology for the formation of shared or unshared classificatory data, in order to incorporate new data collections in the short and medium-term, to increase the investigative capacity of the intended system;
- Analysis and Morphology: To inform CADE of the results of the statistical analysis of the incorporated documents;
- Machine Learning, Research and Investigation: Support to models, mechanisms, and techniques of adaptive and evolutionary analysis of the classifications and researches, by the automated use of the results obtained in the treatment of Analysis and Morphology.

We used machine learning classification because we considered it adequate for: (a) Performing more restrictive filters through more qualified aggregations than simple text search; (b) formation of clusters in the identification of groups of documents by interest from a given group of reserved words; (c) expansion of visual interpretation capabilities through document relation graphs and word cloud formation techniques; (d) application of advanced automatic summarization techniques; (e) interpretation of named entities and their relations across several documents, and (f) support for the formation of controlled vocabularies through n-gram validation.

We chose as system development platform, Apache Solr [15,40,41]. Solr allows for indexing and scalable searches and facets for managing searches, occurrence highlighting, and advanced analysis capabilities. Through this tool, the search system can provide advanced search filters, which can be conditional (where it adds specific fields to return an exact answer from the system), search with specific characters/terms, by proximity or Boolean operators, search by relevance, phonetic search with spell checker and auto-suggestion. As a search result, the system features word highlighting, pagination and sorting, controlled vocabulary synonyms, the definition of term-stopwords, and document standardization. In addition, the Jurisprudence system allows the indexing of various file extensions, such as PDF with OCR, Word documents, and Excel Spreadsheets.

Figure 1 shows an example of a search using the developed system. Collections can perform the search in the system’s database, i.e., judgments from the Federal Audit Court, guides and publications, jurisprudence, legislation, news, and technical opinions, selected according to the end user’s needs. The search results show all documents with the word searched in highlight (“Cartel”). In addition, the company names returned after the search was protected and named with <blind name>. The system returns ten search results per page, and for each of them, we have the options of process data, related documents, summaries (summary of a decision), verbatim (sequence of key words, or expressions indicating the subject discussed in the text), device (rule resulting from the judgment), and conclusion (final decision), depending on the document type searched and the collection it belongs. Moreover, the user can add the search to a knowledge basket, available for future searches. Algorithm 1 presents the code for this search.

The screenshot displays the 'Jurisprudence Search System' interface. At the top, there is a navigation bar with links for 'CORONAVIRUS DISEASE (COVID-19) PANDEMIC', 'INFORMATION ACCESS', 'PARTICIPATE', 'LEGISLATION', and 'GOVERNMENT ENTITIES'. Below this is the CADE logo and the system title. The main content area shows 'Search Results' with a search bar containing 'Search for ...' and a dropdown menu with options 'and', 'or', 'not', 'prox', and '\*'. Below the search bar, it indicates '2639 found documents (10 per page)' and a pagination control showing page 1 of 8. A 'PDF' and 'CSV' button are visible. A 'Recent Documents' dropdown and a 'Highlight on' toggle are also present. The search results list includes a document titled 'Process Number: 0911522 - Administrative Process Vote - 31/05/2021' with a '+ Basket' button. The document snippet discusses the existence of a cartel and its impact on the market.

Figure 1. Results of a search performed in the Jurisprudence Search system.

### Algorithm 1: Search Algorithm

**Result:** Write here the result

```
procedure sendGetDocuments(RESTAPISERVER, TEXT)
```

```
    TEXT ← adjustSearchText(TEXT);
```

```
    TEXT ← addFilterTwo(TEXT);
```

```
    FIELD ← "content";
```

```
    FQS ← createFQs();
```

```
    ROWS ← getTotalRowsQty();
```

```
    SORT ← getSortModel();
```

```
    if getCurrentCollection() = "jurisprudence" OR "technicalReport" then
```

```
        if isProcessNumber(TEXT) then
```

```
            FIELD ← "processNumber";
```

```
            TEXT ← clearProcessNumber(TEXT);
```

```
        else
```

```
            TEXT ← switchSpecials(TEXT);
```

```
        end
```

```
    else
```

```
        TEXT ← switchOperators(TEXT);
```

```
    end
```

```
    PROX ← onVerifyProx(TEXT);
```

```
    if SEARCH <- getCurrentCollection() = "all" then
```

```
        FQS ← "fq = collection : " + "" + getCurrentCollection() + "" + FQS;
```

```
    else
```

```
    end
```

```
    FQS ← onVerifyProcessParts(FQS);
```

```
    if FIELD = "EMENTA" then
```

```
        CONTENT ← "q = documentVerbation : " + PROX +
```

```
        "ORdocumentDisposal : " + PROX + "ORdocumentConclusion : " + PROX;
```

```
    else
```

```
        CONTENT ← "q = " + FIELD + PROX;
```

```
    end
```

```
    QRY ← RESTAPISERVER + "?" + CONTENT + ROWS + FQS + SORT;
```

```
    saveLastRequest(FIELD, TEXT, QRY);
```

```
    JSON ← solrRequest(QRY);
```

```
    return JSON;
```

```
end procedure
```



The Jurisprudence Search System allows the integration with CADE's internal databases, being structured data or not, and the indexed data from CADE's SEI, CADE in Numbers, Portal, Intranet, audio transcripts, among others. In building the intelligence of the system's data, we use machine learning, and to ensure accurate indexing, the jurisprudence system is composed of an architectural model in three stages: (i) The first called indexing stage; (ii) the second is called machine learning, and (iii) the third is classification, in order to increase the Artificial Intelligence layer gradually and concisely.

In the indexing stage, we define the essential terms, create the term vectors, and apply the TF-IDF and the relevance of the terms, considering the synonym treatment performed previously, as shown in search Algorithm 1. Then, the machine learning step consists of indexing and classifying the data. We also use dictionary technologies enriched in the previous stages, allowing a model with faceted and word search with more significant support for knowledge formation. The end-user consumes a structure based on facets and pivots during the classification phase, following the selected preferences and its query routines. In the machine learning step, we first performed the model training using the Learning to Rank (LTR) technique [42], and the Support Vector Machine (SVM) algorithm [43]. Next, we perform cross-validation, that is, a process that estimates the training error and validates the dataset selected for training.

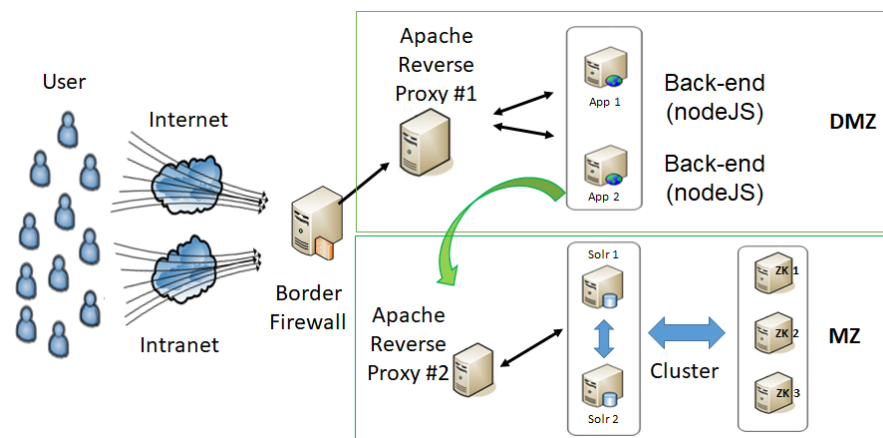
In general, the jurisprudence system allows the user to enter a set of keywords and retrieve documents related to that set, also considering synonyms relevant to the search. The system interface provides a search on classified data in which the returned results are later related laterally as a more detailed filter, either by subject categorization or related tags, composing a search filter. When consolidating the search, it should be possible to treat the results by relevance, most accessed, or referenced documents. The architecture of the proposed solution was developed using the client-server model. Figure 2 presents the architecture of the jurisprudence system developed.

At each cycle of re-evaluation of classification from other unsupervised models performed by the proposed architecture, the model itself will perform feedback using new training bases. For example, suppose we train the model for three documents (words), and the documents are being evaluated with two new documents (words). At the end of this process, the training base will be fed back with five documents (words), updating the training base with five documents and calibrating the model for further training. This process adjusts the classifications (or any data from the unsupervised treatment), and the proposed model in the architecture (Figure 2) can treat a larger supervised (trained) database.

The proposed architecture (Figure 2) has public and private access to the client interface code domain held in a demilitarized zone (DMZ) but with security guarantees (HTTPS, attack treatment, and others.) performed by an edge firewall. In this context, to ensure high availability in the access to the client interface codes, there is an Apache Reverse Proxy (#1) acting after the edge firewall that filters the results and, mainly, performs the load balancing between the two servers that provide high availability (fault tolerance and round-robin load balancing). It is important to emphasize that the data domain and the properly authenticated Solr search APIs run on Cade's militarized network (MZ). Thus, the codes in the client interface to access the data domain through an adequately secured call, and the second layer Apache Reverse Proxy (#2), maintains high availability. Only the Apache Reverse Proxy (#2) has specific access directives, using header elements of each call to the Solr APIs that ensure the authenticity of the requesting user, in this case, the Proxy itself.

The high availability of the Solr environment is guaranteed by a balanced model of servers configured adequately through a process using the Zookeeper model, according to the rules defined by Apache regarding Apache Solr instances. This way, 2 Apache Solr servers and 3 Zookeepers servers ( $2n + 1$  of the "n" Apache Solr servers) were positioned both in serving the requests coming from the NodeJS interface layer (through Apache Reverse Proxy #2) and the internal data loading/downloading processes for making indexed textual data available in support of the model (Figure 2). Furthermore, the channel

issues width, memory, and disk, specific to the model, were measured and applied as much as possible, according to the recommendations of each element/layer.



**Figure 2.** Architecture of the developed system.

On the client-side, we use the AngularJS framework [15,44], Bootstrap 4 [45], the HTML5 [46,47] and the CSS 3 [46], for building the front-end. The application consumes a REST type API [48], built by means of the NodeJS framework [49], which makes the requests and does all the processing of the information stored in the databases used. Thus, the front-end of the jurisprudence system will be responsible for rendering on the screen all the functionalities that will be available to the users of the application (Figure 2).

Commonly, each interaction performed by the user in the application results in a request to the controller [15,50], between the front-end and back-end modules. This request can be anything from a page change (where new information must be loaded) to a new query to the jurisprudence system database. It is important to note that this request exchange interaction between the different modules (back-end to front-end and vice-versa) uses the HTTP protocol [51] through asynchronous requests (Figure 2).

On the back-end of the jurisprudence system, we use Node.js technology [49,52], as the execution environment and in this environment two modules were implemented: (1) Solr API [40] together with MySQL Client Driver [53] to communicate with the SEI database using the MySQL DB database management system [17,54,55]; (2) the back-end application is a REST API [48], which must interact with the Solr API on “/api/select” calls. The Solr API is responsible for accessing the Lucene data persistence kernel. The communication between the Angular client and Solr via API serves as a proxy that controls its access. In the back-end API, we use a module to communicate with the Solr environment (Figure 2).

### 3. Method

In this paper, we performed a literature review to investigate the characteristics of existing Jurisprudence Search systems to identify some challenges and functionalities that we could incorporate in its development for a Brazilian Federal Public Administration agency. In addition, we surveyed to identify the Jurisprudence Search systems used by Brazilian agencies and their features and functionalities.

#### Research Questions

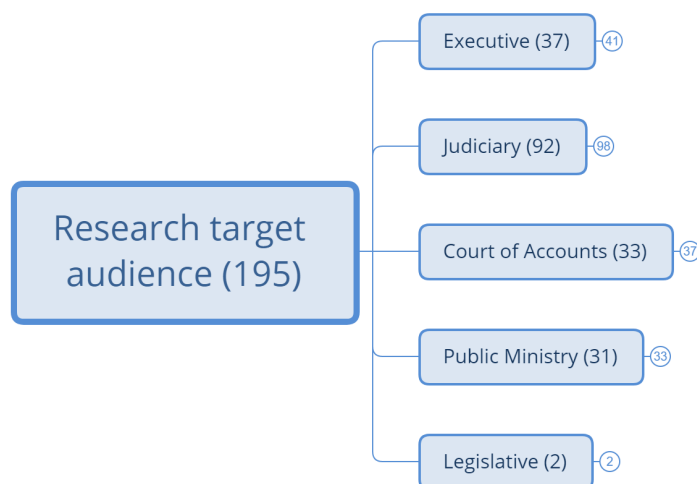
In order to achieve the main goal of this research, the following research questions have been defined to help achieve this main objective:

RQ.1. Which agencies of the Brazilian Public Administration use a Jurisprudence Search System, and what are the characteristics of these systems?

RQ.2 What functionalities offer the Jurisprudence Search Systems?

To answer the research questions, we conducted a literature review and a survey of Brazilian agencies. The survey was composed of 52 questions and addressed to all agencies

of the Brazilian Public Administration. In total, there are 195 agencies, divided among the executive, judiciary, court of accounts, public ministry, and legislative branches, as shown in Figure 3. In addition, we contacted the 195 agencies through institutional e-mails. The questionnaire was applied between June 2021 and July 2021 and had 107 agencies (55% of the total), and the average response time was 12 min. Table 1 contains all questions that the agencies in the survey answered.



**Figure 3.** Research target audience.

**Table 1.** Survey Questions.

ID	Question
Q1	What is the name of your agency?
Q2	What is the power rating of your agency?
Q3	What is the sphere classification of your agency?
Q4	Does your agency have a textual database processing system?
Q5	Was the system developed in-house or contracted out?
Q6	What database is used in the textual database processing system?
Q7	What is the “other” database option?
Q8	What programming language is used in the textual database processing system?
Q9	Describe the “other” database option.
Q10	If the system is public, what is the link to the textual database processing system?
Q11	Does the system have a user’s manual?
Q12	If the manual is public, what is the link to the user manual?
Q13	What indexing engine is used by the textbase processing system?
Q14	Describe the “other” option of the indexing engine question.
Q15	What are the document formats of the textual bases processing system?
Q16	Describe the “other” option of the document format question.
Q17	In the textual base, do you have digitized documents (obtained from paper scanning)?
Q18	What is the ratio of digitized documents to native digital documents?
Q19	Is the system compliant with LGPD requirements?
Q20	Has a usability analysis of the textual basis processing system been performed?
Q21	Does the textual base’s processing system has filters by categories (date of issue, type of process, units, areas of interest, subjects, among others)?
Q22	Does the textual base’s processing system use logical operators (and, or, not, and others)?

**Table 1.** *Cont.*

ID	Question
Q23	Does the textbase processing system offers the option of exporting the results (pdf, CSV, etc.)?
Q24	Does the textual database processing system index the contents of other agencies?
Q25	Does the textual processing system index various documents (PDF, Word, Excel, other)?
Q26	Does the textual processing system using any method to define the relevance of documents?
Q27	If it does, please describe the method used to define the relevance of the documents.
Q28	Does the textbase processing system uses a Controlled Vocabulary?
Q29	If the vocabulary is public, please put the link in the field below.
Q30	Does the textbase processing system uses an ontology?
Q31	If it does, please describe the ontology used.
Q32	Does the textbase processing system using any multimedia data extraction process? (For example, deduplication of audio and video files).
Q33	If a multimedia data extraction process exists, describe it.
Q34	Does your agency use statistical methods in the textual base processing system?
Q35	If yes, what methods are used?
Q36	Does your agency use any of these text-mining techniques in the textbase processing system?
Q37	Describe “others” of the text mining techniques?
Q38	Does your agency use supervised machine learning techniques for text processing?
Q39	If used, describe the supervised machine learning technique.
Q40	Does your agency use unsupervised machine learning techniques for text processing?
Q41	If you use it, describe the unsupervised machine learning technique.
Q42	Is there a technique or model for extracting specific parts of documents, e.g., identification, menu, conclusion?
Q43	If there is, please describe the technique or model for extracting specific parts of documents.
Q44	Is any natural language processing (NLP) technique used in the textual base processing system?
Q45	If there is, please describe the NLP techniques used.
Q46	Is there any study/publication on the use of artificial intelligence in the agency’s textual base processing system?
Q47	Is there any study/publication on the use of artificial intelligence in the agency’s textual bases processing system?
Q48	Share the link to the study/publication or describe them.
Q49	What functionalities do you think a textual bases processing system should have?
Q50	What are your suggestions for improvements to your agency’s textbase processing system?
Q51	What other projects related to the use of Machine Learning, Artificial Intelligence, and Text Mining techniques are your agency
Q52	How can Machine Learning, Artificial Intelligence, and Text Mining techniques improve your agency’s activities?

#### 4. Results and Discussions

##### 4.1. RQ.1 Which Agencies of the Brazilian Public Administration Use a Jurisprudence Search System and What Are the Characteristics of These Systems?

Among the 107 agencies that participated in the survey, 55 are part of the Judiciary. The Executive had 28, 14 from the Auditors’ Court, 8 from the Public Ministry, and 2 from the Legislative. Concerning the classification of the sphere of the body, 66 were from the Federal Public Administration, 38 were in the provincial level, and 3 were municipal, as presented in Figure 4.

Figure 5a presents that 42 agencies reported that they do not use a textual basis processing system, and 56 stated that they do. In addition, 44 agencies stated that the textual bases processing system was developed internally by them. Figure 5b shows that 11 agencies contracted the system.

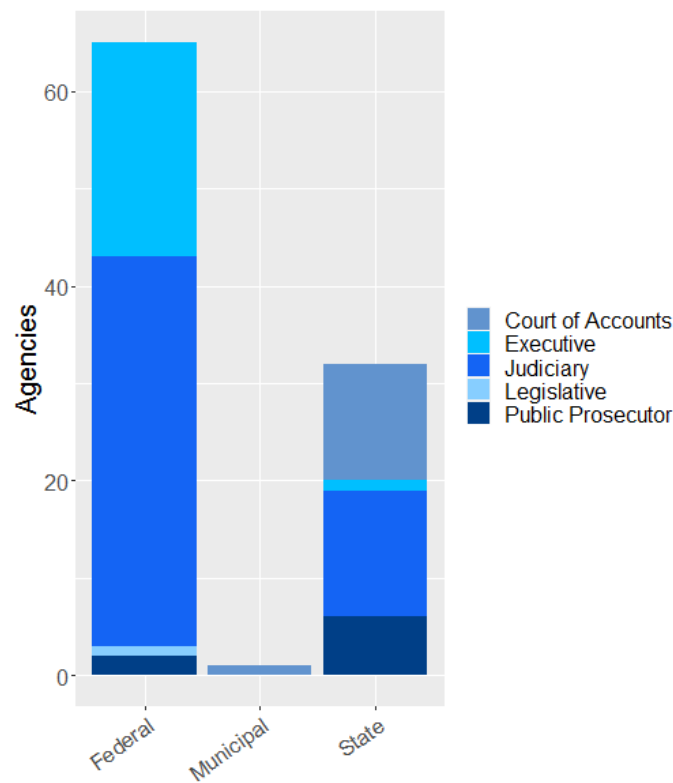


Figure 4. Classification of power and agency sphere.

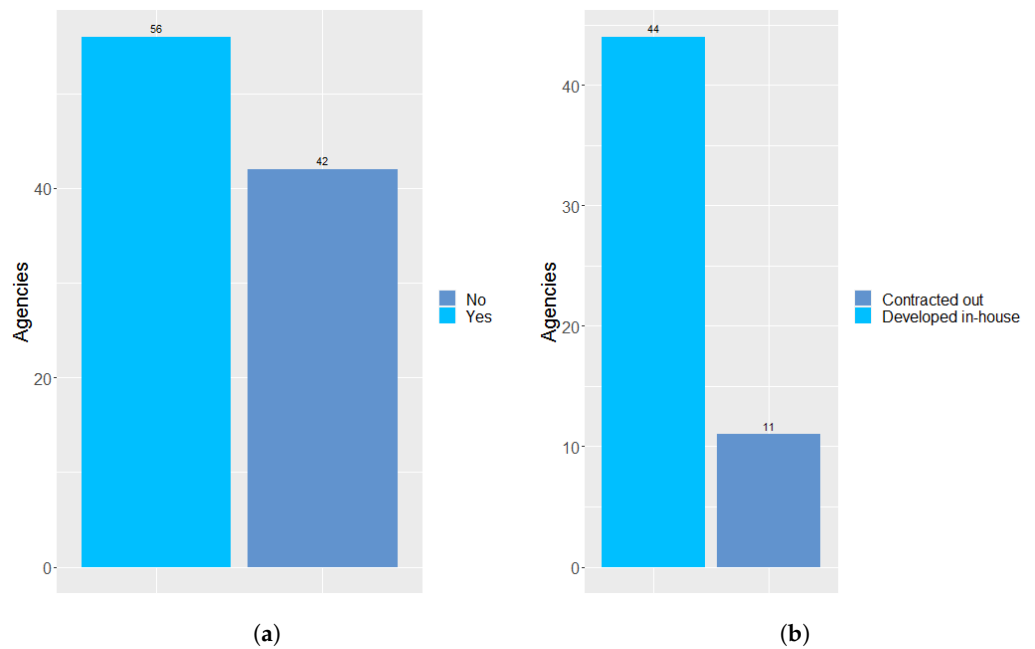
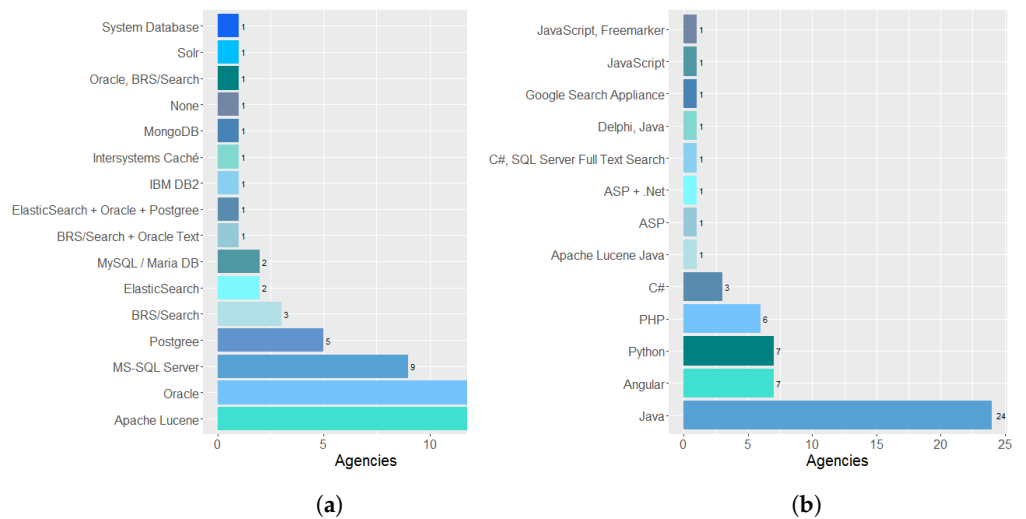


Figure 5. (a) shows if there is a textual database processing system in the agency or not, while (b) shows if the agency’s textual database processing system was developed in-house or contracted out.

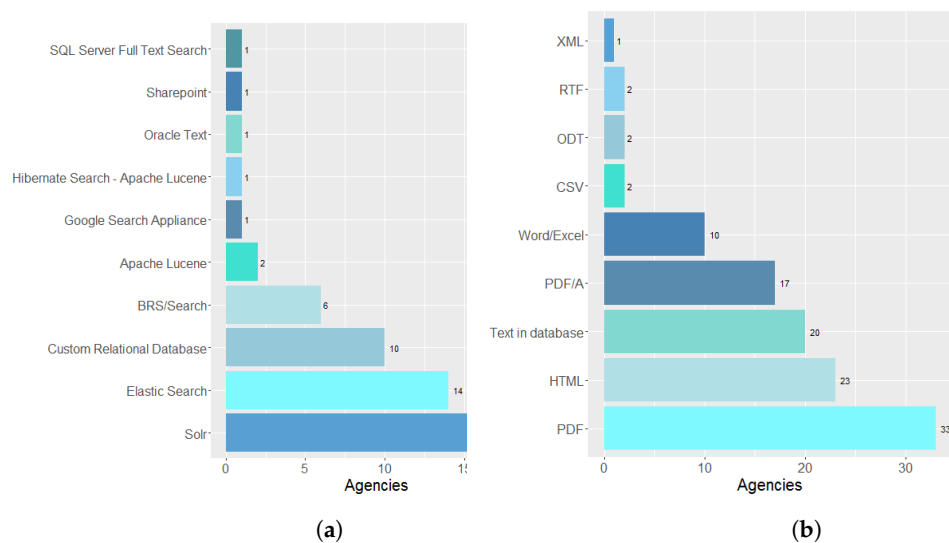
Regarding the database used in the textual database processing system, 14 agencies reported using Apache Lucene to perform indexing and retrieval of textual data. Additionally, 12 agencies use Oracle. 9 agencies use MS-SQL Server. 5 agencies use Postgree. 3 agencies use BRS/Search. Finally, 2 agencies use MySQL/Maria DB. Moreover, only one

agency uses ElasticSearch, IBM DB2, Oracle/BRS Search, or Solr. In addition, one agency uses the System Database, as presented in Figure 6a. Twenty-four agencies used the Java programming language to develop the textual base processing system, seven agencies used the Angular language and Python, six agencies used PHP, three agencies developed the system in the C# language. Finally, one agency from the survey use the Apache Lucene Java, ASP, ASP + .Net, C#, SQL Server Full Text Search, Delphi, Java, Google Search Appliance, JavaScript, and Freemarker languages, respectively, as presented in Figure 6b.



**Figure 6.** (a) shows the database used in the textual database processing system, while (b) shows the programming language used by the systems.

Thirty-six agencies stated that the developed system does not have a user’s manual, and twenty-one stated it does. Regarding the indexing engine used by the textual database processing system, 19 agencies reported that they use Solr, 14 agencies use Elastic Search, 10 agencies use a custom relational database, 6 agencies use BRS/Search, and 2 agencies use Apache Lucene. Figure 7a shows the databases: Google Search Appliance, Hibernate Search, Apache Lucene, Sharepoint by one agency, and SQL Server Full-Text Search only one agency uses them.

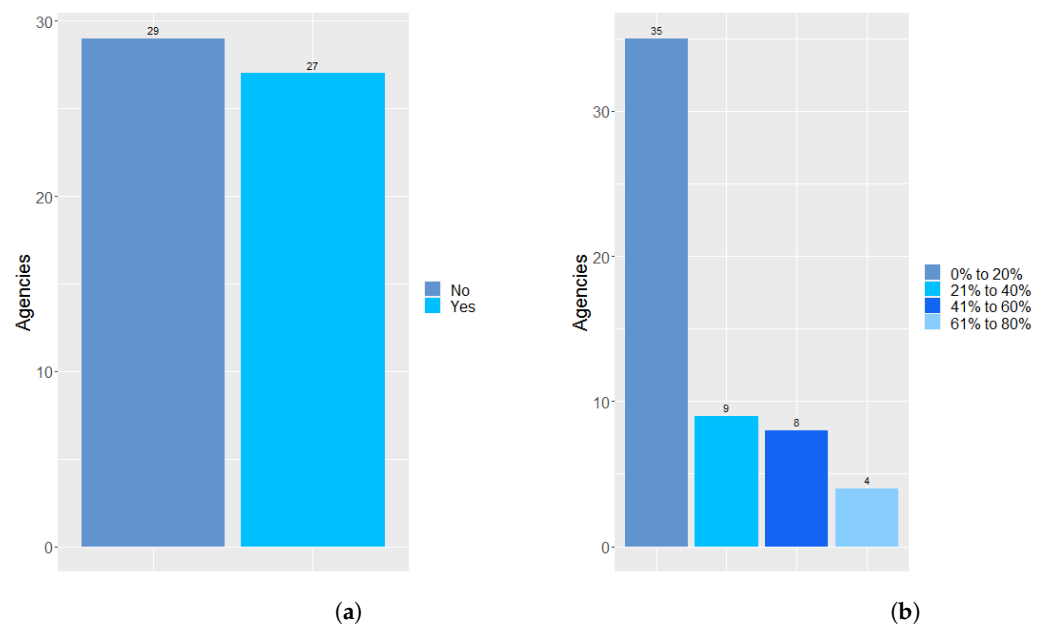


**Figure 7.** (a) shows the indexing engine used in the textual base processing system, while (b) shows the format of the documents in the textual base processing system.

Regarding the formats of the documents in the textual base processing system, 33 agencies informed that they use the pdf format, 23 in HTML, 20 in Text in the database, 17 in PDF/A, 10 in Word/Excel format, two in CSV format, two agencies use ODT and RTF format. One agency claimed to use XML format, as shown in Figure 7b.

Figure 8a shows that twenty-seven agencies informed that the textual base had digitized documents, and they obtain the documents from scanning paper documents, and 29 informed that they do not have.

Figure 8b has the proportion of digitized documents concerning already digital documents, 35 agencies stated that they have from 0% to 20%, nine agencies between 21% to 40%, eight agencies between 41% to 60%, four agencies between 61% to 80% and only one agency reported having between 81% to 100% of their digitized documents.



**Figure 8.** (a) shows if the textual database has digitized documents, while (b) shows the ratio of digitized documents to born-digital documents.

Regarding whether the Jurisprudence Search system developed by the agencies that answered the survey is in compliance with the principles of the General Law on Personal Data Protection (LGPD), only 15% of them said yes, 52% were neutral, and 33% disagreed that the system complies, as presented in Figure 9 (Q19). This result is a preoccupying factor since all systems developed by Brazilian agencies must comply with the LGPD. In this sense, the system developed in this research meets this requirement, i.e., the Jurisprudence Search system developed for CADE is compliant with the LGPD. Furthermore, 39% of the agencies stated that they perform a usability analysis of the textual database processing system, 45% were neutral, and 16% stated that there is no usability analysis, as presented in Figure 9 (Q20).

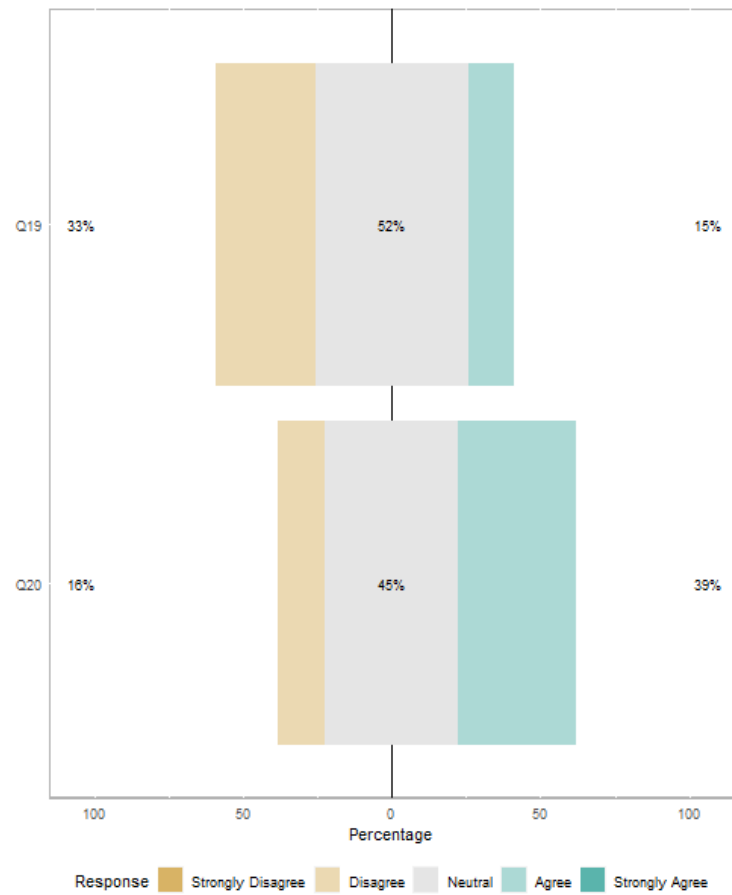


Figure 9. LGPD compliance and usability best practices.

4.2. RQ.2 What Functionalities Are Offered by the Jurisprudence Search Systems?

Concerning the functionalities of the textual bases processing system, 47% of the agencies participating in the survey informed that the textual bases processing system has filters by categories, such as date of issuance, type of process, units, areas of interest, subjects, among others. On the other hand, 41% of the agencies were neutral, and 12% of agencies stated that the system developed does not have this functionality, as presented in Figure 10 (Q21).

Figure 10 (Q22) shows 69% of the agencies strongly agree and agree that the textbase processing system uses logical operators (and, or, not, among others), 15% of the agencies were neutral, and 15% strongly disagree and disagree. 40% of the agencies strongly agree and agree that the textual base processing system offers the possibility to export the search result to pdf, CSV format, among others, 35% of the agencies were neutral, and 25% strongly disagree and disagree (Figure 10 (Q23)). Regarding whether the textual base processing system indexes content from other agencies, only 38% strongly agree and agree, 50% were neutral, and 12% strongly disagree and disagree (Figure 10 (Q24)).



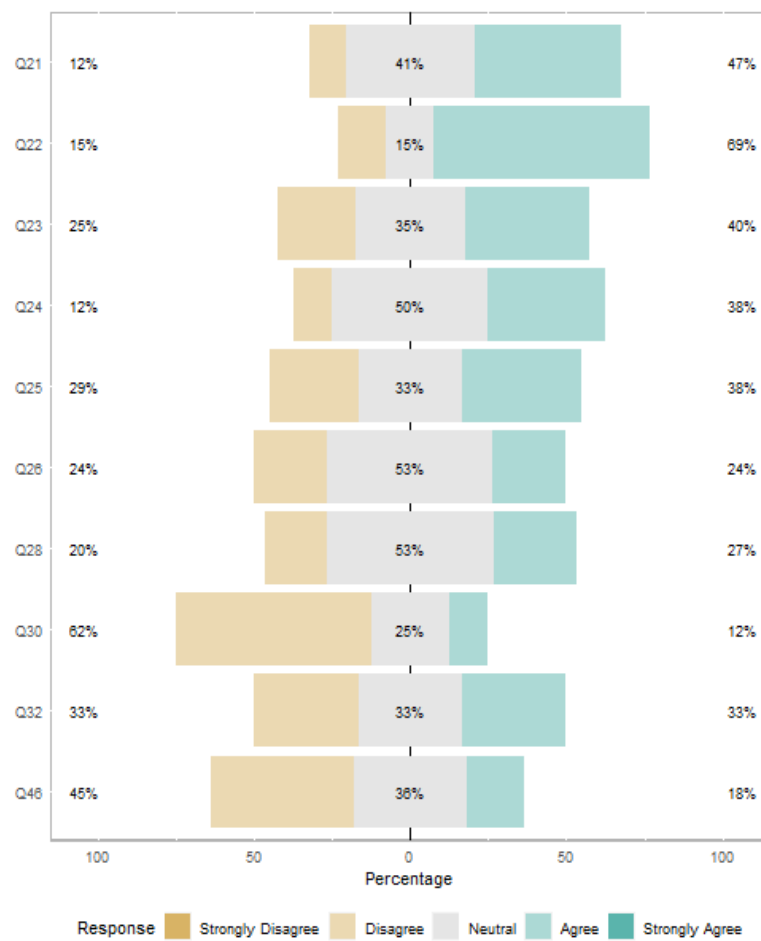


Figure 10. Jurisprudence Search System Features.

Hence, 38% of the agencies stated that the textual base processing system indexes various documents, such as digital, PDF, Word, Excel, etc. However, 33% of the agencies were neutral, and 29% of the agencies strongly disagree and disagree (Figure 10 (Q25)). In addition, 24% of the agencies stated that the textbase processing system uses some method to define the relevance of documents, 53% were neutral, and 24% stated that the agency does not use any method (Figure 10 (Q26)). Among the methods used to define document relevance, some agencies stated:

*“Lucene’s standard relevance calculation is used, based on term count, term frequency, inverted document frequency, and field size.”*

*“Relevance by publication date.”*

*“The Ranking features offered by SQL Server Full Text Search are used, for sorting the results.”*

*“The ElasticSearch database has a method of gauging document relevance from the queried term.”*

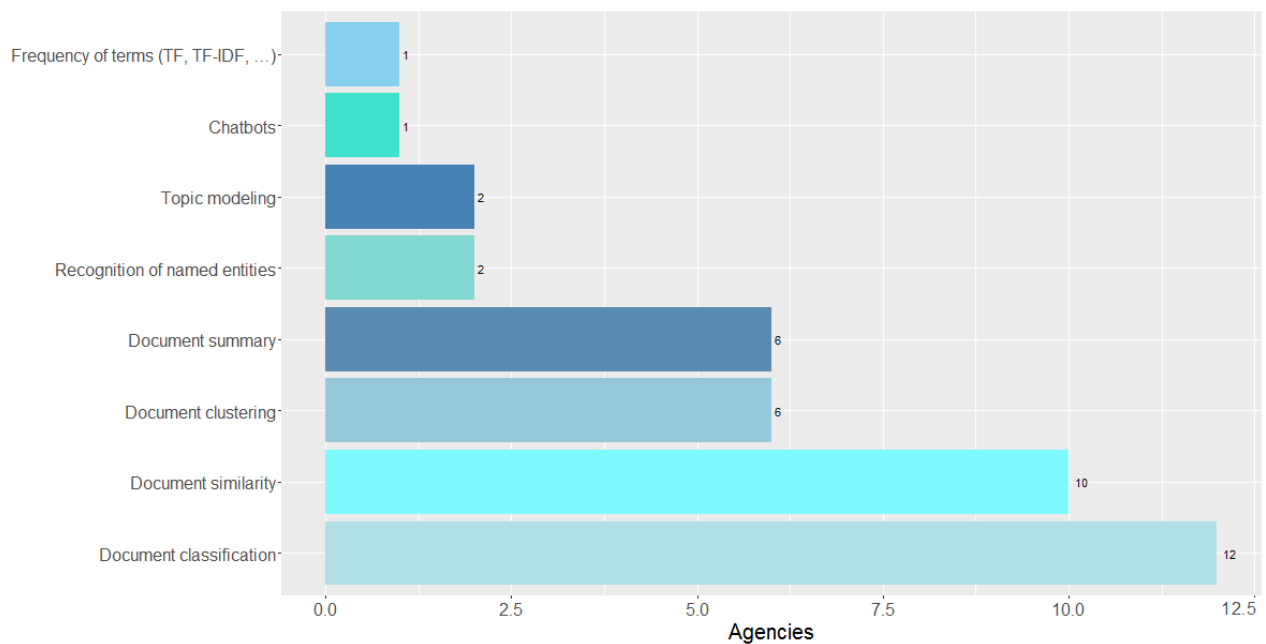
Figure 10 (Q28) shows that 27% of the agencies stated that their textual base processing system uses a controlled vocabulary. 53% of the agencies were neutral, and 20% stated that the system used by the agency does not have a controlled vocabulary. Toward, 12% of the agencies’ textbase processing systems that participated in the survey reported using some Ontology in their solutions. 25% of the agencies were neutral, and 62% reported that they do not use any Ontology (Figure 10 (Q30)).

This finding reveals that the Jurisprudence Search system developed in the context of this research has as one of the main differentials compared to the systems developed by other agencies in its development, the use of Ontologies. Only one agency participating in the survey stated that the textual base processing system uses multimedia data extraction, such as deduplication of audio and video files. This finding also reveals a differentiation from our system Figure 10 (Q32).

Just 18% of the agencies reported that there is some study/publication on the use of Artificial Intelligence in the agency's textual base processing system, 36% were neutral and 45% reported that there is no study for the use of AI, as presented in Figure 10 (Q46).

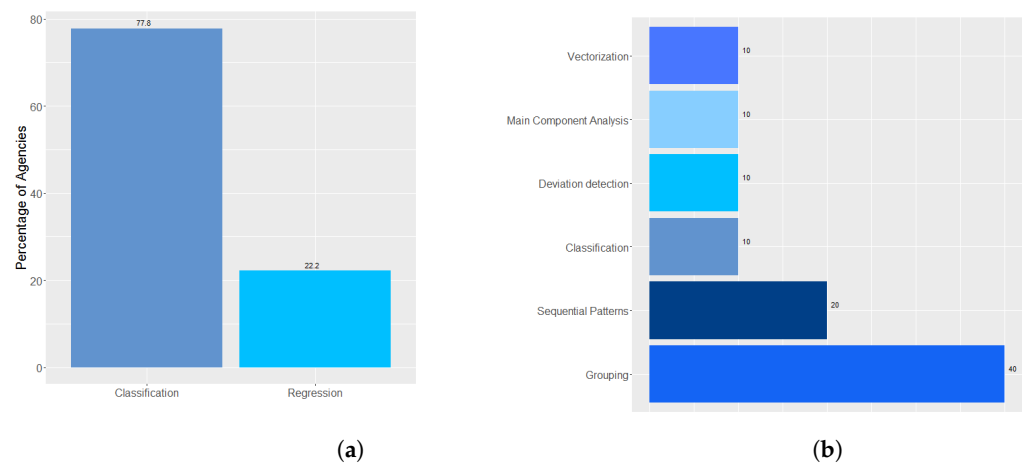
Forty-nine agencies reported that their agency does not use statistical methods in the textual bases processing system. Eight agencies stated that they used supervised classification, document classification, similarity, and document clustering.

About text mining techniques the agency uses in the textual bases processing system, 12 agencies reported document classification, and ten agencies use document similarity. Six agencies reported that they use document clustering, and six agencies reported that they use document summary. 2 agencies reported that they use recognition of named entities and two agencies use topic modeling, one agency reported that it uses chatbots, 01 agency uses the frequency of terms, as presented in Figure 11.



**Figure 11.** Text mining techniques used in textual database processing system.

Concerning the supervised Machine Learning techniques used for text treatment, 70% of the agencies informed that they use the classification technique, 20% use the regression technique and 10% use the sequential patterns, as presented in Figure 12a. Forty agencies reported that they use the clustering technique as an unsupervised Machine Learning technique. Twenty agencies use sequential patterns. Besides, ten agencies use deviation detection, ten agencies principal component analysis, and ten agencies reported that they use vectorization (Figure 12b).



**Figure 12.** (a) shows the supervised Machine Learning techniques used to text processing, while (b) shows the unsupervised Machine Learning techniques.

Thirty-seven agencies reported no extraction technique or template for specific parts of documents in the system they use, such as identification, comments, and conclusion. Only 16 agencies informed that the system they use has some extraction technique or model. Among them were textual search by terms, search by menu and structured abstract, models based on Machine Learning, and models based on Grammars and Characteristics.

Forty-four agencies use Natural Language Processing (NLP) technique in the textbase processing system, and nine agencies said yes. Among those mentioned are (a) pre-processing and vectorization of the content of the case records; (b) data collection, raw text extraction, sentence division, tokenization, normalization (systemization, lemmatization), removal of empty words and part-of-speech tagging; (c) part-of-speech tagging, machine learning (classification, clustering, named entity recognition), chunk regular expression, N-gram parsing, feature-based grammars; and (d) tokenization, stop words, stemming, thesaurus, vectors.

We also investigated the perception of Brazilian public administration agencies regarding which functionalities a textual base processing system should have. Some of the answers were:

*“Systems should offer similar content identification.”*

*“Systems should also perform the search for the content in full, not just the descriptors.”*

*“The systems must allow searching by relevance through advanced filters, such as document type, unit, subject, signatory and dates. In addition they should perform synonym handling and stemming.”*

*“Systems should perform keyword searches, Boolean operators for jurisprudence search, semantic similarity search, cluster analysis and abstract generation.”*

Regarding suggestions for improvements to the agency’s existing textual base processing system, some responses were:

*“Insertion of other databases, identification of a “paradigm decision” in the results of the jurisprudence search, identification of citations to decision contents with binding effects in the decisions resulting from the search. Improve response time and use machine learning techniques to improve result ranking.”*

*“Legislative reference and search using fuzzy logic and Artificial Intelligence. In addition, an improvement in document indexing, user interface, and user experience design need to be incorporated.”*

Concerning how the use of Artificial Intelligence, Machine Learning and Text Mining techniques can improve the agency's finalistic activities, some answers were:

*"Grouping similar processes and offering a document template to treat each group, offering greater efficiency in audit actions in the selection of objects of greater relevance, risk and materiality. In addition, assisting in the decision making of the subject areas when preparing the annual inspection plan."*

*"Through the optimization and automation of manual and repetitive work, allowing greater agility in the analysis of processes. Moreover, in the classification and recognition of textual patterns, it is possible to search for procedural pieces and opinions that can help and speed up the construction of new opinions. Thus, the use of these techniques are promising in the sense of enabling the sharing and dissemination of knowledge."*

*"Automation of routine tasks, allowing the team to focus on more strategic activities; Greater assertiveness and speed in performing activities; Quick analysis of large volumes of data, providing better subsidies for decision making; Analysis of historical and current facts to make predictions about future events, enabling, for example, better planning in inspection activities and behavioral analysis of the regulated entities aiming to evolve the Agency's regulation, always seeking to improve the return to the population."*

#### 4.3. Discussion

The systems for processing textual bases are present in most Public Administration agencies, but we can observe that the solution architecture model for implementing these systems is not consensual and does not indicate paths of best practices and the standards adopted. However, the quantitative analysis presented in the survey shows the technologies and techniques used by these systems that are in line with the model proposed by CADE, such as the predominance of Apache Lucene as a text search library, since this technology allows high-performance searches in large volumes of information.

In the Jurisprudence Search System, the proposed solution architecture uses Apache Solr to process data from the SEI database for data indexing purposes, responsible for accessing Apache Lucene resources. The Java language appears as predominant in the survey results and is used in CADE's system because it is the native language of Apache Lucene.

Concerning the availability of search resources applied to the surveyed systems, most of them have similarities with the functionalities implemented in their search systems. For example, many of the respondents cited the use of filters, logical operators, PDF file treatment, and export. Therefore, the solution proposed by CADE, besides implementing these functionalities and resources, presents the differential of basket resources, search history, and highlights.

In the development of CADE's Jurisprudence Search system, we used Artificial Intelligence techniques in conjunction with statistical techniques to perform natural language processing and discourse analysis techniques to form a supplementary knowledge base, text mining, and machine learning. Unfortunately, most of the agencies participating in the survey did not use any Artificial Intelligence techniques. However, some agencies mentioned the use of Machine Learning techniques for data classification and clustering. Thus, we can infer that CADE's Jurisprudence Search system differs from other systems used by Brazilian public administration agencies by using Artificial Intelligence and Ontology in the proposed solution.

It is noteworthy that Machine Learning techniques can be categorized into supervised and unsupervised and applied alone or combined, depending on the needs and according to the defined database. Therefore, it is necessary to perform a preliminary analysis to identify the appropriate techniques for each scenario. The technologies used in this research were Artificial Intelligence using Machine Learning and Text mining. The ML methods and techniques used in this work are for information retrieval, such as extraction (using

facets), classification (clustering, summarization, named entities), indexing (by extracting n-grams), and natural language processing.

The open questions of the survey brought important information and perceptions about the opinions of the experts of each organization that use search systems, such as (i) improve usability, accessibility, user experience in the use of search systems; (ii) improve text indexing; (iii) index other types of content such as audio and video of plenary sessions; (iv) perform a search by relevance, keywords, metadata, advanced filters, treatment of synonyms; (v) export the documents and the search results; (vi) recommend other documents; (vii) identify documents with similar content, and (viii) incorporate other textual bases. These insights were essential for the continuous improvement of the Jurisprudence Search systems.

### 5. Limitations and Threats to Validity

As in any research that investigates users' perceptions concerning a given scenario, we have some threats to validity. Regarding the fidelity of the participants' answers, we cannot guarantee that all of them answered according to the actual scenario of the Brazilian agencies and if the information represents all the technologies and techniques applied in the development of the Jurisprudence Search systems used by them. To mitigate this threat, we did not make the information of all survey participants public. In addition, the results of the quantitative data analysis would not impact the evaluation of the agencies by the controlling bodies.

Regarding the system developed in the actual case study of this research, the Jurisprudence Search system currently developed has some limitations, which are: (a) There is a current inability of the solution to apply the statistical and stochastic processes using Artificial Intelligence techniques in interpreting the indexed terms that exist in the database, posing a challenge in transforming the processes that use simple natural language processing to the intended final understanding, which supports the marking of summaries and review of the main sentences with their terminologies adequately supported; (b) in dealing with security aspects, the imposition of multiple levels of access produces undesirable latencies, with the exponential growth of indexed bases of legal documents and other collections with Apache Solr; (c) the use of free visual components that have a high impact on the execution of screen templates, creates situations of anticipated and synchronous treatment in the construction of screens, which results in an increase of time in the total loads of the results data; (d) the analytical view of the data in clusters and in correlations in general is still being implemented, one of the fundamental issues of support to the internal activities of analysis and interpretation of legal pieces referring to some theme or subject; and (e) regarding the ontological treatment of data, such as, for example, the one that allows the observation of the interconnection model among the several named entities and their physical and legal relations within their activity sectors and shareholding, object of CADE's observation, has not been implemented in the available solution yet. As a mitigation to these factors, we are developing new functionalities that will meet the needs of the Jurisprudence Search system.

### 6. Conclusions

This paper presents a solution for CADE's Jurisprudence Search system to perform textual database processing. First, we performed the collection, retrieval, and indexing of the terms to build the database. Afterward, we used Artificial Intelligence techniques, statistical methods, summarization techniques, indexing, named entity recognition, natural language processing, and ontology. These techniques provided the Jurisprudence Search system a differential to other Jurisprudence Search systems used by other Brazilian agencies. The main contributions of this system are more accurate search results, treatment of structured and unstructured data from different sources and formats, aiming to build a knowledge base that supports legal decision-making processes and analysis.

We also surveyed to identify which Brazilian public administration agencies have textual database processing systems, which use technological resources and artificial intelligence and morphological construction techniques. Our findings revealed that Apache Solr is the main indexing engine used by the systems, and Apache Lucene and the Java language were the most used in developing these systems. However, no agency participating in the survey stated that it uses ontology to organize and structure its information. In addition, more than 85% of the Jurisprudence Search systems used by the agencies are not LGPD compliant.

As future work, we will implement further improvements to the Jurisprudence Search system to build a knowledge base, using templates to facilitate the analysis of Jurisprudence opinions. In addition, the use of the system will monitor the users' perceptions regarding the decision support.

**Author Contributions:** Conceptualization, V.A.M., V.C.R. and V.E.d.R.; methodology, E.D.C. and R.T.d.S.J.; software, V.A.M., L.A.C.C. and F.L.L.d.M.; validation, R.M.G. and F.L.L.d.M.; formal analysis, R.T.d.S.J.; investigation, V.E.d.R., V.A.M. and F.A.M.D.; resources, L.A.C.C.; data curation, R.B.; writing—original draft preparation, A.L.S.O. and E.D.C.; writing—review and editing, A.L.S.O. and E.D.C. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported in part by CNPq-Brazilian National Research Council (Grant 312180/2019-5 PQ-2 and Grant 465741/2014-2 INCT on Cybersecurity), in part by the Administrative Council for Economic Defense (Grant CADE 08700.000047/2019-14), in part by the Brazilian Ministry of the Economy (Grant DIPLA 005/2016 and Grant ENAP 083/2016), in part by the General Attorney of the Union (Grant AGU 697.935/2019), in part by the National Auditing Department of the Brazilian Health System SUS (Grant DENASUS 23106.118410/2020-85), and in part by the General Attorney's Office for the National Treasure (Grant PGFN 23106.148934/2019-67).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** R.T.d.S.J. would like to thank the support of the Brazilian research, development and innovation agency CNPq (grant 312180/2019-5 PQ-2).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Constâncio, A.S. *Ontologia Para um Motor de Busca Semântica para Recuperação Jurisprudencial no Brasil*; Universidade Federal do Paraná: Curitiba, Brazil, 2017; pp. 1–221.
2. Lee, R.W. *Pesquisa Jurisprudencial Inteligente*; Universidade Federal de Santa Catarina: Florianópolis, Brazil, 1998; pp. 1–163.
3. Bourguet, J.; Costa, M.Z. About the Exposition of Brazilian Jurisprudences. In Proceedings of the IX ONTOBRAS Brazilian Ontology Research Seminar, Curitiba, Brazil, 3 October 2016; Volume 1862, pp. 138–143.
4. da Costa Calheiros, T.; Monteiro, S.D. Mecanismos de busca de jurisprudência: Um instrumento para a organização do conhecimento e recuperação da informação no ambiente jurídico virtual. *Em Questão* **2017**, *23*, 146–166. [[CrossRef](#)]
5. Canedo, E.D.; do Vale, A.P.M.; Patrão, R.L.; de Souza, L.C.; Gravina, R.M.; dos Reis, V.E.; Dias, F.A.M.; Mendonça, F.L.L.; de Sousa, R.T. Usability Assessment of a Jurisprudence System. In *International Conference on Human-Computer Interaction*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 482–499.
6. Mahdi, M.N.; Ahmad, A.R.; Ismail, R.; Natiq, H.; Mohammed, M.A. Solution for Information Overload Using Faceted Search—A Review. *IEEE Access* **2020**, *8*, 119554–119585. [[CrossRef](#)]
7. D'Amore, R.J.; Mah, C.P. One-Time Complete Indexing of Text: Theory and Practice. In Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Montréal, QC, Canada, 5–7 June 1985; pp. 155–164. [[CrossRef](#)]
8. Sarwar, M.A.; Ahmed, M.; Habib, A.; Khalid, M.; Ali, M.A.; Raza, M.; Hussain, S.; Ahmed, G. Exploiting Ontology Recommendation Using Text Categorization Approach. *IEEE Access* **2021**, *9*, 27304–27322. [[CrossRef](#)]
9. Kaushik, A.; Naithani, S. A comprehensive study of text mining approach. *Int. J. Comput. Sci. Netw. Secur.* **2016**, *16*, 69.
10. Loutsaris, M.A.; Charalabidis, Y. Legal informatics from the aspect of interoperability: A review of systems, tools and ontologies. In Proceedings of the 13th International Conference on Theory and Practice of Electronic Governance, Athens, Greece, 23–25 September 2020; pp. 731–737. [[CrossRef](#)]

11. van Engers, T.M.; Boer, A.; Breuker, J.; Valente, A.; Winkels, R. Ontologies in the Legal Domain. In *Digital Government: E-Government Research, Case Studies, and Implementation*; Integrated Series in Information Systems; Chen, H., Brandt, L., Gregg, V., Traummüller, R., Dawes, S.S., Hovy, E.H., Macintosh, A., Larson, C.A., Eds.; Springer: Berlin/Heidelberg, Germany, 2008; Volume 17, pp. 233–261. [\[CrossRef\]](#)
12. Avgerinos Loutsaris, M.; Lachana, Z.; Alexopoulos, C.; Charalabidis, Y. Legal Text Processing: Combing Two Legal Ontological Approaches through Text Mining. In Proceedings of the 22nd Annual International Conference on Digital Government Research, Omaha, NE, USA, 9–11 June 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 522–532. [\[CrossRef\]](#)
13. Martins, V.A. *Arquitetura de um Ambiente Colaborativo de Business Intelligence Baseado em um Repositório de Ontologias e Serviços de Dados*; Universidade de Brasilia (UnB): Brasilia, Brazil, 2012.
14. Broughton, V. The need for a faceted classification as the basis of all methods of information retrieval. *Aslib Proc.* **2006**, *58*, 49–72. [\[CrossRef\]](#)
15. Suominen, O.; Viljanen, K.; Hyvänen, E. User-centric faceted search for semantic portals. In Proceedings of the European Semantic Web Conference, Innsbruck, Austria, 11–15 November 2007; pp. 356–370.
16. Tunkelang, D. Faceted Search. *Synth. Lect. Inf. Concepts Retr. Serv.* **2009**, *1*, 1–80. [\[CrossRef\]](#)
17. Lachana, Z.; Loutsaris, M.A.; Alexopoulos, C.; Charalabidis, Y. Automated Analysis and Interrelation of Legal Elements Based on Text Mining. *Int. J. E Serv. Mob. Appl.* **2020**, *12*, 79–96. [\[CrossRef\]](#)
18. Barros, R.; Peres, A.; Lorenzi, F.; Wives, L.K.; da Silva Jaccottet, E.H. Case law analysis with machine learning in Brazilian court. In Proceedings of the International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Montreal, QC, Canada, 25–28 June 2018; pp. 857–868. [\[CrossRef\]](#)
19. Gomes, T.; Ladeira, M. A new conceptual framework for enhancing legal information retrieval at the Brazilian Superior Court of Justice. In Proceedings of the 12th International Conference on Management of Digital EcoSystems, Virtual Event, United Arab Emirates, 2–4 November 2020; pp. 26–29. [\[CrossRef\]](#)
20. Bueno, T.C.D.; von Wangenheim, C.G.; da Silva Mattos, E.; Hoeschl, H.C.; Barcia, R.M. JurisConsulta: Retrieval in jurisprudential text bases using juridical terminology. In Proceedings of the Seventh International Conference on Artificial Intelligence and Law, Oslo, Norway, 14–17 June 1999; pp. 147–155.
21. Ordoñez, H.A.; Ordoñez, C.C.; Ordoñez, J.A.; Urbano, F.A. Jurisprudence search in Colombia based on natural language processing (NLP) and Lynked Data. *INGE CUC* **2020**, *16*. [\[CrossRef\]](#)
22. Aletras, N.; Tsarapatsanis, D.; Preotiuc-Pietro, D.; Lampsos, V. Predicting judicial decisions of the European Court of Human Rights: A Natural Language Processing perspective. *PeerJ Comput. Sci.* **2016**, *2*, e93. [\[CrossRef\]](#)
23. Canedo, E.D.; Mendes, B.C. Software Requirements Classification Using Machine Learning Algorithms. *Entropy* **2020**, *22*, 1057. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Silva, N.; Braz, F.; de Campos, T. Document type classification for Brazil’s supreme court using a Convolutional Neural Network. In Proceedings of the Tenth International Conference on Forensic Computer Science and Cyber Law (ICoFCS), Sao Paulo, Brazil, 29–30 October 2018; pp. 7–11. [\[CrossRef\]](#)
25. da Costa, R.P.; Canedo, E.D.; de Sousa Júnior, R.T.; de Oliveira Albuquerque, R.; García-Villalba, L.J. Set of Usability Heuristics for Quality Assessment of Mobile Applications on Smartphones. *IEEE Access* **2019**, *7*, 116145–116161. [\[CrossRef\]](#)
26. Alshammari, N.; Alanazi, S. An Arabic Dataset for Disease Named Entity Recognition with Multi-Annotation Schemes. *Data* **2020**, *5*, 60. [\[CrossRef\]](#)
27. Weber, R. Intelligent jurisprudence research: A new concept. In Proceedings of the Seventh International Conference on Artificial Intelligence and Law, Oslo, Norway, 14–17 June 1999; pp. 164–172.
28. Giacalone, M.; Cusatelli, C.; Romano, A.; Buondonno, A.; Santarcangelo, V. Big Data and forensics: An innovative approach for a predictable jurisprudence. *Inf. Sci.* **2018**, *426*, 160–170. [\[CrossRef\]](#)
29. Houy, C.; Niesen, T.; Fettke, P.; Loos, P. Towards automated identification and analysis of argumentation structures in the decision corpus of the German Federal Constitutional Court. In Proceedings of the 7th IEEE International Conference on Digital Ecosystems and Technologies, DEST 2013, Menlo Park, CA, USA, 24–26 July 2013; pp. 72–77. [\[CrossRef\]](#)
30. Pasquale, F.; Cashwell, G. Prediction, persuasion, and the jurisprudence of behaviourism. *Univ. Tor. Law J.* **2018**, *68*, 63–81. [\[CrossRef\]](#)
31. Nagumothu, D.; Eklund, P.W.; Ofoghi, B.; Bouadjeneq, M.R. Linked Data Triples Enhance Document Relevance Classification. *Appl. Sci.* **2021**, *11*, 6636. [\[CrossRef\]](#)
32. Sun, X.; Applebaum, T.H. Intonational phrase break prediction using decision tree and n-gram model. In Proceedings of the 7th European Conference on Speech Communication and Technology, Aalborg, Denmark, 3–7 September 2001; pp. 537–540.
33. Wawrzyński, A.; Szymański, J. Study of Statistical Text Representation Methods for Performance Improvement of a Hierarchical Attention Network. *Appl. Sci.* **2021**, *11*, 6113. [\[CrossRef\]](#)
34. Stephan, H. Application of Methods for Syntax Analysis of Context-Free Languages to Query Evaluation of Logic Programs. *arXiv* **2014**, arXiv:1405.3826.
35. Kirk, D.; MacDonell, S.G. An Ontological Analysis of a Proposed Theory for Software Development. *arXiv* **2021**, arXiv:2103.10623.
36. Shi, C.; Wei, B.; Wei, S.; Wang, W.; Liu, H.; Liu, J. A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *EURASIP J. Wirel. Commun. Netw.* **2021**, *2021*, 31. [\[CrossRef\]](#)

37. Sáiz-Manzanares, M.C.; Pérez, I.R.; Rodríguez, A.A.; Arribas, S.R.; Almeida, L.; Martín, C.F. Analysis of the Learning Process through Eye Tracking Technology and Feature Selection Techniques. *Appl. Sci.* **2021**, *11*, 6157. [[CrossRef](#)]
38. Fuad, A.; Al-Yahya, M. Analysis and Classification of Mobile Apps Using Topic Modeling: A Case Study on Google Play Arabic Apps. *Complexity* **2021**, *2021*, 6677413:1–6677413:12. [[CrossRef](#)]
39. Arnaiz-González, Á.; Díez-Pastor, J.; Díez, J.J.R.; García-Osorio, C.I. Instance selection for regression by discretization. *Expert Syst. Appl.* **2016**, *54*, 340–350. [[CrossRef](#)]
40. Apache Solr Reference Guide. Available online: [https://lucene.apache.org/solr/guide/8\\_4/](https://lucene.apache.org/solr/guide/8_4/) (accessed on 20 April 2021).
41. Guntupally, K.; Dumas, K.; Darnell, W.; Crow, M.C.; Devarakonda, R.; Prakash, G. Automated Indexing of Structured Scientific Metadata Using Apache Solr. In Proceedings of the IEEE International Conference on Big Data, Big Data 2020, Atlanta, GA, USA, 10–13 December 2020; pp. 5685–5687. [[CrossRef](#)]
42. Duan, J.; Kashima, H. Learning to Rank for Multi-Step Ahead Time-Series Forecasting. *IEEE Access* **2021**, *9*, 49372–49386. [[CrossRef](#)]
43. Lin, L. Learning information recommendation based on text vector model and support vector machine. *J. Intell. Fuzzy Syst.* **2021**, *40*, 2445–2455. [[CrossRef](#)]
44. Zhang, G.; Zhao, J. Visualizing Interactions in AngularJS-based Single Page Web Applications. In Proceedings of the 30th International Conference on Software Engineering and Knowledge Engineering, Hotel Pullman, Redwood City, CA, USA, 1–3 July 2018; pp. 402–403. [[CrossRef](#)]
45. Han, Q. Inventory System Based on ThinkPHP and Bootstrap Framework. *Am. J. Theor. Appl. Res.* **2019**, *1*, 1–16.
46. Aamulehto, R.; Kuhna, M.; Tarvainen, J.; Oittinen, P. Stage framework: An HTML5 and CSS3 framework for digital publishing. In Proceedings of the ACM Multimedia Conference, MM '13, Barcelona, Spain, 21–25 October 2013; pp. 851–854. [[CrossRef](#)]
47. Theisen, K.J. Programming languages in chemistry: A review of HTML5/JavaScript. *J. Cheminform.* **2019**, *11*, 11:1–11:19. [[CrossRef](#)]
48. Costa, B.; Pires, P.F.; Delicato, F.C.; Merson, P. Evaluating a Representational State Transfer (REST) Architecture: What is the Impact of REST in My Architecture? In Proceedings of the 2014 IEEE/IFIP Conference on Software Architecture, WICSA 2014, Sydney, Australia, 7–11 April 2014; pp. 105–114. [[CrossRef](#)]
49. Sun, H.; Bonetta, D.; Humer, C.; Binder, W. Efficient dynamic analysis for Node.js. In Proceedings of the 27th International Conference on Compiler Construction, CC 2018, Vienna, Austria, 24–25 February 2018; pp. 196–206. [[CrossRef](#)]
50. Dobrea, D.; Diosan, L. A Hybrid Approach to MVC Architectural Layers Analysis. In Proceedings of the 16th International Conference on Evaluation of Novel Approaches to Software Engineering, ENASE 2021, Online Streaming, 26–27 April 2021; pp. 36–46. [[CrossRef](#)]
51. Belshe, M.; Peon, R.; Thomson, M. Hypertext Transfer Protocol Version 2 (HTTP/2). *RFC* **2015**, *7540*, 1–96. [[CrossRef](#)]
52. Saundariya, K.; Abirami, M.; Senthil, K.R.; Prabakaran, D.; Srimathi, B.; Nagarajan, G. Webapp Service for Booking Handyman Using MongoDB. In Proceedings of the 2021 3rd International Conference on Signal Processing and Communication, ICPSC, Coimbatore, India, 16 June 2021; pp. 180–183. [[CrossRef](#)]
53. Kiran, P.R.; Krishna, Y.S. MIDP based J2ME driver for accessing MySQL from mobile devices. *Int. J. Innov. Sci. Eng. Technol.* **2014**, *1*, 164–168.
54. Dang, T.K.; Ta, M.H.; Dang, L.H.; Hoang, N.L. An Elastic Data Conversion Framework: A Case Study for MySQL and MongoDB. *SN Comput. Sci.* **2021**, *2*, 325. [[CrossRef](#)]
55. Nash, T.; Olmsted, A. Performance vs. security: Implementing an immutable database in MySQL. In Proceedings of the 12th International Conference for Internet Technology and Secured Transactions, ICITST 2017, Cambridge, UK, 11–14 December 2017; pp. 290–291. [[CrossRef](#)]