Application of artificial intelligence in the automatic identification and classification repetitive demand resolution incident in the Brazilian Court of Justice

Antonio P. Castro Jr[1]

Dr. Gabriel A. Wainer

Dr Wesley P. Calixto[2]

Abstract: One of the areas of knowledge with several possibilities for applying artificial intelligence is Law. Recent changes in Brazilian legislation have facilitated the use of information technology resources to streamline the progress and judgment of cases, such as repetitive demand resolution incident (IRDRs). The aim of this paper is to develop and apply an AI method that can identify and relate new lawsuits with consolidated repetitive judgments (IRDRs). The datasets used in this research are judges' repetitive judgment documents, and consolidated in IRDRs. Court documents are transformed into weighted vectors. The construction of the weights in the vector is based on the co-occurrence of the terms, calculated from the combination of the term frequency-inverse document frequency and their similarity in the corpus of the same IRDR. Artificial neural networks are trained with these vectors

[1] Possui graduação em Ciência da Computação pela PUC-Goiás (1999) e mestrado em Ciência da Computação pela UNICAMP (2001). Está cursando doutorado na Escola de Engenharia de Computação, EMC/UFG, em Inteligência Artificial. Tem experiência na área de Ciência da Computação, com ênfase em Sistemas de Informação, atualmente atuando nos seguintes temas: ciência de dados, inteligência artificial, aprendizagem de máquinas e algoritmos de aprendizagem e predição, ênfase em processamento de linguagem natural. É Analista Judiciário, área de Tecnologia da Informação e Ciência de Dados, do Tribunal de Justiça do Estado de Goiás desde 1997.

[2] Possui graduação em Física pela Pontifícia Universidade Católica de Goiás (2002), mestrado em Engenharia Elétrica e de Computação pela Universidade Federal de Goiás (2008) e doutorado em Engenharia Elétrica pela Universidade Federal de Uberlândia (2011) com período na Universidade de Coimbra (UC), Portugal. Realizou pós-doutorado em modelagem de sistemas eletromagnéticos aplicado a geoprospecção na Carleton University (CU), Ottawa/Canadá no Visualization and Simulation Centre (VSIM). Atualmente é docente permanente no programa de pós-graduação da Universidade Federal de Goiás e professor full do Instituto Federal de Educação, Ciência e Tecnologia de Goiás. Atua na área de modelagem de sistemas com ênfase em sistemas inteligentes, processo de otimização, modelos computacionais e inteligência artificial.

to recognize whether new lawsuits are related to an IRDR. As the methodology obtained 93% accuracy, 97% precision, and 93% in recall in the simulations, the method can streamline the work of the Court of Justice, seeking to solve society's conflicts as quickly as possible. Although the method can be used in several scenarios, the simulations were carried out in judicial documents.

Keywords: Artificial intelligence. Similarity-learning. Repetitive demand resolution incident. Machine learning. Artificial neural networks.

Data de submissão do artigo: Agosto de 2021

Data de aceite do artigo: Novembro de 2021

Introduction

One of the main innovations in the Law area in the last years is the use of artificial intelligence (AI) in the classification and grouping of texts (CASTRO *et al.,* 2020; *RANI et al.,* 2017). The volume of lawsuits filed in the judiciary (Kim *et al.: 2019), as opposed to the task* force of the judges, as well as the possibilities of flows imposed in the legislation of the countries, has hindered the celerity in the attendance to the rights of the society (CASTRO *et al.,* 2017, 2017a). Thus, since human resources to judge the number of lawsuits that are brought to justice is scarce, it is believed that optimization techniques with information retrieval technologies, using AI, can be applied to improve this scenario (CASTRO *et al., 2017, 2017a,* 2020).

Some works that use AI in law, such as Zhang *et al. (2015), Zhang et al. (2017)* and Rani *et al. (2017), bring difficulties in the use of computational resources for automation* and prediction of texts in law, in situations that really bring benefits to society, works discussed in section 2. The change in the Brazilian Civil Procedure Code in 2015 brought the possibility of standardizing the understanding of the Brazilian Justice on legal facts and theses, and implementing, in the life of Brazilian society, the constitutional principle of isonomy. The Repetitive Demand Resolution Incident (IRDR) is the materialization of this change. The standardization of legal understanding based on lawsuits is a raw material for information technology, especially for training in AI solutions, to frame new actions and relate to the consolidated IRDRs.

In this sense, this work aims to train an artificial neural network to learn to recognize whether an action that reaches the Judiciary is related to a consolidated IRDR. Upon recognition, the AI tool will notify the magistrate of the case's relationship with the IRDR before judgment of the case.

This work **innovates** in three aspects: (a) using the IRDRs of the Court of Justice of Goiás for training in AI solutions, (b) after training the AI solution, the solution will be able to relate the consolidated IRDRs to the lawsuits that come to court, and (c) develop an integration solution with the electronic process system to inform magistrates before the judgment of lawsuits.

The **originality** of the studies and researches carried out in this article is in the application in the corpus of real lawsuits documents, in the governmental institution of justice in Goiás, Brazil. The dataset used in this research are documents of judgment of judges, known as decisions and consolidated in so-called IRDRs. These documents describe the event that occurred, covering aspects of the law inherent in the case and solicitations by lawyers. They are documents rich in information, integrating generated facts and applied laws. These decisions shape the way the law is interpreted and applied by the many attorneys and access professionals across the government (Castro *et al.:* 2020, 2017, 2017a; Jasanoff: 2018).

The **relevance of** this research is to provide celerity in the judgments of the judicial proceedings, since it groups, identifies, and classifies the received lawsuits, allowing connection with cases already judged. In addition, the **relevance** of the AI method applied in this paper, makes it possible to comply, mainly, with the provisions of article 332 of Law nº. 13.105/2015. To **apply** the method proposed in this study, an API was created, allowing to insert the text classification functionality in the software of the Court of Justice of Brazil, known as Projudi.[3] In addition to the **contribution to society** applied to the Court of Justice, this article brings relevant contributions to the academic environment, by building and making the source code available on GitHub[4] to perform preprocessing on justice text documents, in Portuguese, which allows its use and comparison with other research works. The results found have a low time of processing in the learning of the proposed method and the prediction of new documents, its application is not onerous in a production environment.

[3] Software currently used by the Court of Justice in Goiás to create and maintain electronic lawsuits.

[4] GitHub is a hosting tool where 65 million developers and researchers shape the future of software, together contributing to the open-source community (https://GitHub.com/apcastrojr).

The AI solution applied in this paper uses the term frequency-inverse document frequency (*tf-idf) model together with the Jaccard similarity metric to transform the set of* repetitive judgments, the IRDRs, into vectors with the co-occurrence weights of the terms of these text documents. Vectors, with assigned values, are used to train the artificial neural network (ANN). After the ANN training, the AI solution will be able to predict whether the lawsuit that arrived at the Court of Justice is related to any IRDR in the training. To evaluate the proposed method, accuracy, precision, and recall metrics are used. The results obtained are:

(a) 93% in accuracy, (b) 97% in precision, and (c) 93% in recall.

This paper contains the following structure: Section 2 describes the theoretical basis of this work with related works, description of the judiciary, information retrieval, and artificial neural networks. Section 3 details the proposed methodology, Section 4 and Section 5 present the results obtained and the discussion of the work, respectively, while the conclusions are provided in Section 6.

1 Theoretical background

In the world there is no standard of the functioning of the judiciary, each country has established its way of judging and structuring its judiciary. Some countries prioritize laws in codes, while other countries prioritize recurring judgments over laws in codes. In the first case, countries are known as Civil Law, while in the latter they are known as Common Law (DAVID, 2014). However, despite differences in the structures and the functioning of the judiciary in countries, when there are conflicts and there is no dissolution, this is triggered in an attempt to resolve the divergence (Moura and Sousa, 2013).

Courts of justice are important structures for the society of each country (DAVID, 2014). The volume of information generated, from the various judgments made, allows to know both the history of a nation and the understanding in the dissolution of social problems (CASTRO *et al., 2014; MOURA and SOUSA, 2013*). *With the high volume of information generated* and the scarcity of human resources, it is difficult to compose standardized judgments, especially linking judgments with similar cases already judged. In this situation, the application of computational resources, mainly AI, can help in the consistent resolution of society's problems. (CASTRO *et al.,* 2020, 2014).

In addition to law (ZHANG *et al.,* 2017),*, other areas of knowledge have also been* working on document classification through machine learning, such as i) medicine (Arsene *et al.,* 2011), ii) biology (LAMY, 2017) iii) engineering (RANI *et al.,* 2017), iv) education (Grubisic *et al.: 2013), among others.* It is noticed that the use of AI in the classification of texts is relevant and applied in several areas of knowledge.

1.1 Related Works

Zhang *et al. (2015) discusses the difficulties encountered in the construction of* Chinese lawsuit ontology. The authors argue that the challenges lie not only in the ontology, but also in the organization of data in the Chinese judicial system. There are three obstacles to the Chinese judicial system: i) ambiguity of lawsuit language, ii) deficiency in the inference of judicial decisions, and iii) limited role of lawsuit in China. The authors state that, based on these difficulties, the judicial ontology proposed is the mixture between normative documents and lawsuits. Because it is a conceptual work, the results focus on presenting the challenges for the construction of judicial ontology.

Zhang *et al. (2017) continue the work in Zhang et al.* (2015), creating the information retrieval system based on judicial precedents and lawsuits in the Chinese judicial system. The system allows recovering judgments and norms by relevance. The methodology used supports logical reasoning and incorporates a hierarchical structure. In addition to the ontological contribution generated, the authors say that they can improve the accuracy metrics in information retrieval using genetic algorithms integrated with the K-Nearest Neighbor (KNN).

Rani *et al. (2017) claim* that the greatest challenge in the application of the methodology is in its automatic construction, says that there are still no mechanisms to generate the ontology of textual bases in a fully automated way. Rani *et al. (2017) article* addresses two algorithms that can model ontology by topic, LSI & SVD and Mr.LDA. The purpose of these algorithms is to determine the statistical relationship between the documents and the terms to construct the topic ontology with minimal human intervention. The method proposed by Rani *et al.* (2017) demonstrates efficacy in semantic retrieval relating to searches performed.

Ceci and Gangemi (2016) develops work using the semantic web, a library of lawsuit knowledge that is based on the metadata contained in judicial documents, where are constructed semantic relationships by extracting fragments in lawsuit texts. As a result, is generated the library called JudO, where it presents the interpretations of the judges in the conduct of their lawsuit reasoning. The authors, who are from the area of computer science, contribute to the modeling of judicial knowledge, resulting from studies based on cases and design patterns using the ontology.

Fawei *et al.* (2015) states that laws are an explicit system of rules to govern the behavior and legal practitioners must learn to apply legal knowledge to the facts at hand. Thus, Fawei *et al.* (2015) describes an initial attempt to model and implement the automatic application of legal knowledge using a rule-based approach, presents the AI as a promising method in the judiciary, applied in knowledge management where ontological elements are associated with legal rules.

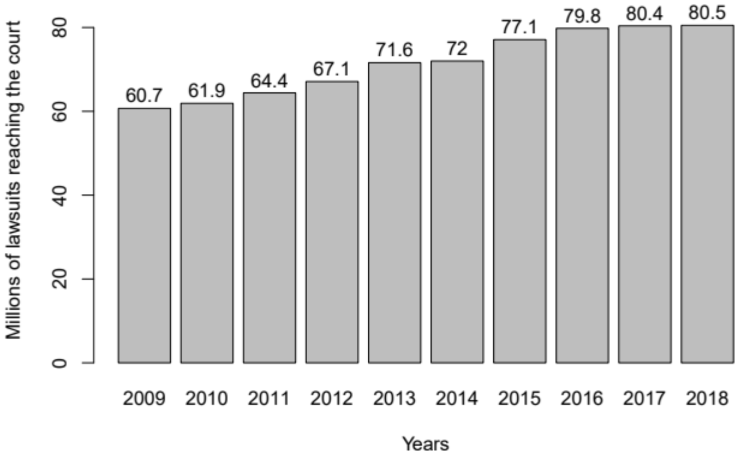
Calambas *et al. (2015) uses the AI in the judicial area,* being constructed a semantic relation of the base of lawsuit decisions uttered. Authors present progress in system development using natural language processing technology, as well as clustering to optimize the recovery process and analysis of the bases of judicial decisions.

1.2 Environment for case study: data from the Judiciary

In Brazil, the judiciary is one of the three pillars that controls and organizes its democratic society, being responsible for enforcing the laws in the country, becoming the guardian of the rules, which establish the relationship between people. The structure of the judiciary in Brazil is regionalized, established in the states of the federation, and having a centralized controlling unit in the capital of the country. The judiciary is divided into federal and state structures. Federal structures are subdivided into central and regional courts, while state structures are organized in state justice. Some branches of justice are created with the purpose of promoting the specialties, such as i) labor justice, ii) electoral justice, iii) military justice, iv) superior court of law, and v) supreme court federal (DONATO, 2006).

According to the National Council of Justice in Kim and Toffoli (2019), the judiciary in Brazil ends the year 2018 with more than 80 million lawsuits in progress. Figure 1, adapted from Kim and Toffoli (2019), shows the lawsuit stock increased to the previous year, and it is increasing year by year (Rocha: 2018; Kim and Toffoli: 2020). Figure 1 shows that the judiciary is unable to reduce pending court cases.

Figure 1: Department of Justice branch's historical lawsuit series.



The judiciary in state of Goiás, Brazil, has almost two million cases in process, and arrives nearly four hundred thousand a year (Kim and Toffoli, 2020). The volume is high to be analyzed and judged, thus increasing the congestion rate year after year. The judiciary in Goiás, Brazil, has an annual average of 6,000 lawsuits per magistrate, where each magistrate has only three advisors (KIM and TOFFOLI, 2020). The judiciary needs to increase the number of judges and their advisors, or constructs software tools capable of speeding up the procedures for judging and filing lawsuits. The human resource is scarce, and there is also a budget limitation for new hires (CASTRO *et al.*, 2017a; KIM and TOFFOLI, 2020). *At the moment, in the* Goiás, Brazil Judiciary, there is big data with information on the court decisions of all the lawsuits already judged, with a rich and powerful knowledge base, which until now has not been explored with AI methods to speed up the judgment (CASTRO *et al.*, 2020). *It is noticed* that it is necessary to think differently and to apply new tools to try to contain the increase in the passive lawsuits. This work uses artificial intelligence software to assist in this scenario. Judging the lawsuits, the Departments of Justice manage to reduce this large volume, shown in Figure 1.

Brazil is known as Civil Law, but changes to the civil procedure code in 2015 introduced Common Law features, for example, the IRDRs. They are repetitive demands in which judges seek to consolidate judgments so that there are no differences. The IRDRs applied in this work are admitted by the Court of Justice of Goiás. The complete list of IRDRs can be consulted on the website of the Court of Justice of Goiás, at the link: https://www.tjgo.jus.br/index.php/nugepnac-irdr.

1.3 Information Retrieval

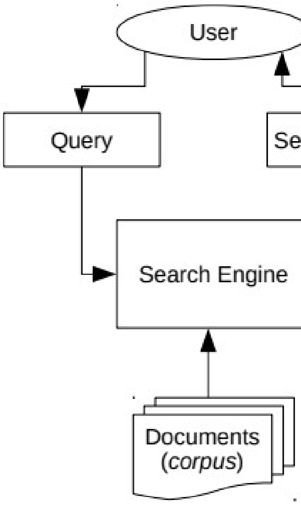
Information retrieval is responsible for handling and retrieving data objects such as text, images, sounds, and so on. Mooers (1951) makes information retrieval (IR) refined and advanced, and after fifty years of evolution it becomes highly sophisticated, bringing interactivity and being accompanied by problems in human-machine interaction. Almeida (2013) describes that IR is the goal of the information and knowledge management process. This line of thinking is also reinforced by Delicato *et al.* (2001), *which describes that the main* intention when performing the database search is to find documents that are useful to satisfy the information needs and avoid retrieval of items useless.

The relevant document is any document that contains information related to the query. The purpose of the information retrieval system is to compare the query with the collection and return the set of related documents to the user, often classified according to their presumed relevance (MANNING *et al.,* 2010).

The effectiveness of the IR system is generally assessed by two measures: i) recall and ii) precision. Recall is the proportion of relevant material retrieved from the archive, while precision is the proportion of recovered material that is found to meet the user's need. Recall and precision tend to vary in inverse way, and this is the difficulty in recovering everything that is desired while rejecting everything that is undesirable.

The objective of IR is to find and present the correct information, from the contents of the document to the user, satisfying their need in the search expression. Figure 2, adapted from Buttcher *et* *al.* (2016), *represents in a simplified way the process of retrieving the* information.

Figure 2: Model of the information retrieval process.



In Figure 2, the search engine is the most relevant points, it compares the query of the users/systems with corpus, retrieving items that are probably the correct ones. Most of the search engines are quantitative in nature, based on disciplines such as i) logic, ii) statistics, and

iii) set theory. The efficiency of the IR system is directly linked to the applied model.

These IR models are also known as search engines. Some search engines were created in the 1960s and 1970s and perfected in the 1980s and are still used today. The ideas of these search engines are present in most of today's retrieval systems, as well as in search engines on the internet. Boughanem *et al. (2009) state that quantitative models have boosted* the development of information retrieval systems, including i) Boolean, ii) Vector, iii) Probabilistic, and iv) clustering.

1.3.1 Quantitative model applied in information retrieval

The quantitative models help in the search of documents using the terms inserted in the search process, in this way, it can be considered that each document in the corpus is represented by the set of terms. The term is the word that represents the concept or meaning present in the document, and to identify the relevance of the term in the description of the content of the document is an onerous task (PONTE and CROFT,1998, 2017).

The quantitative models used in information retrieval can associate weights both in terms of indexing and in terms of the search expression. These weights are used to calculate the degree of similarity between search expressions established by the user for each document. Thus, it is possible to obtain documents ordered by degree of similarity based on search expression (PONTE and CROFT, 1998, 2017).

Buttcher *et tal. (2016) informs that to increase the accuracy of the search it is* necessary to perform actions in the procedure of identifying the terms in the documents, such as i) identify and isolate each word in the document, ii) eliminate words with little semantic value, such as terms common to business, articles, prepositions, and numerals (stopwords), (iii) reduce words to their root, remove suffixes and prefixes, (iv) incorporate the terms into the vectors of the documents (index terms), and (v) assign the weight values to the terms. These listed actions are known as preprocessing.

The calculation of the weight is an important aspect, and it can be applied in several ways: i) manually by specialized personnel, which makes the process costly, ii) measure the number of times the term appears in the document or in the corpus, (iii) calculate the number of documents by the number of documents that have a given term, (iv) calculate the product of the frequency of the term in the document by the number of documents that have a term, described in the paper by Salton and McGill (1986), and (v) other methods (PONTE and CROFT, 1998, 2017; SALTON, 1989). The corpus is represented in a matrix with various documents and indexing terms, where each line represents the document and each column represents the term in the document (PONTE and CROFT, 1998, 2017; SALTON, 1989).

In the literature, several works on the classification of text documents and vectorization of terms apply the concept of bag--of-words (MURPHY, 2013; GARG *et al.,* 2011; BOSCH *et al.,* 2008; LAZEBNIK *et al.,* 2006; FERGUS *et al.,* 2003), *others bag-of-concepts* (LI *et al.,* 2020; MIKOLOV *et al.,* 2013; MILNE *et al.,* 2007; MIHALCEA *et al.,* 2006; GABRILOVICH and MARKOVITCH, 2005). and others apply both solutions (bag-of-words and bag-of-concepts) (LI *et al.,* 2020; KIM, 2014; HUANG *et al.,* 2012).

In the bag-of-words model, the document is transformed into a vector of size n, in which *n is the number of terms/words used to represent the document, each vector field is* associated with the term/word weight, calculated by traditional methods such as term- frequence (*tf),* term frequency-inverse document frequency (*tf-idf),* Okapi BM25, and others. Bag-of-concepts model, also known as word embeddings, the document is transformed into a vector space of size *n* × *d, in which n is the number of terms/words and d the number of* dimensions of the word embeddings, considering that the dimension *d of a term/word is* generally defined by the own term/word and by the terms/words that accompany the term/word in the text, creating a co-occurrence among the terms/words. Methods known and applied in the construction of words embeddings are word2vec (MIKOLOV *et al.,* 2013), GloVe (PENNINGTON *et al.,* 2014), ELMo (Peters *et al.,* 2018), BERT (DEVLIN *et al.,* 2018), *and* others.

After obtaining the document attributes, through the one-dimensional model, bag- of-words, or the multi-dimensional model, bag-of-concepts, it is possible to train different machine learning algorithms to classify documents.

1.4 Machine learning algorithms

Machine learning algorithms receive the vectors representing the documents, as well as informing which category of the document. The category of each vector or document is called tag. This phase of loading and processing the vectors and their respective tags is called training. Machine learning algorithms after being trained can take an unknown vector, unknown tag/category of document, and predict its tab/category, this is called prediction. These algorithms are also called text classifiers. Machine learning is part of artificial intelligence. Machine learning are used in a wide variety of applications, such as in email filtering, computer vision, medicine, text classification, firewall, antivirus, multiple pattern recognition.

Examples of machine learning algorithms are: artificial neural networks, support vector machines, Bayesian networks, genetic algorithms and others. In this article, artificial neural networks with multilayer perceptron with backpropagation (MLPNN) are used.

1.5 Similarity between text documents

Thada *et al.* (2013) report some traditional models to calculate similarity, such as i) Jaccard; ii) Cosine and iii) Dice coefficient. All being statistical models, generating their results between [0,1]. In the studies of Thada *et al. (2013) it is noticed that Jaccard's model* presents the lowest statistical percentage compared to the Cosine and Dice models. Since it is the objective of this work to find the relationship between the terms by calculating their similarities, we have that if Jaccard's expression finds values close to 1, it is believed that the degree of similarity is stronger than in other models.

Jaccard's expression measures the similarity between finite sample sets and is defined as the size of the intersection divided by the size of the union of the sample sets. Given the expression in (1), Jaccard measures the similarity between documents D1 and D2.

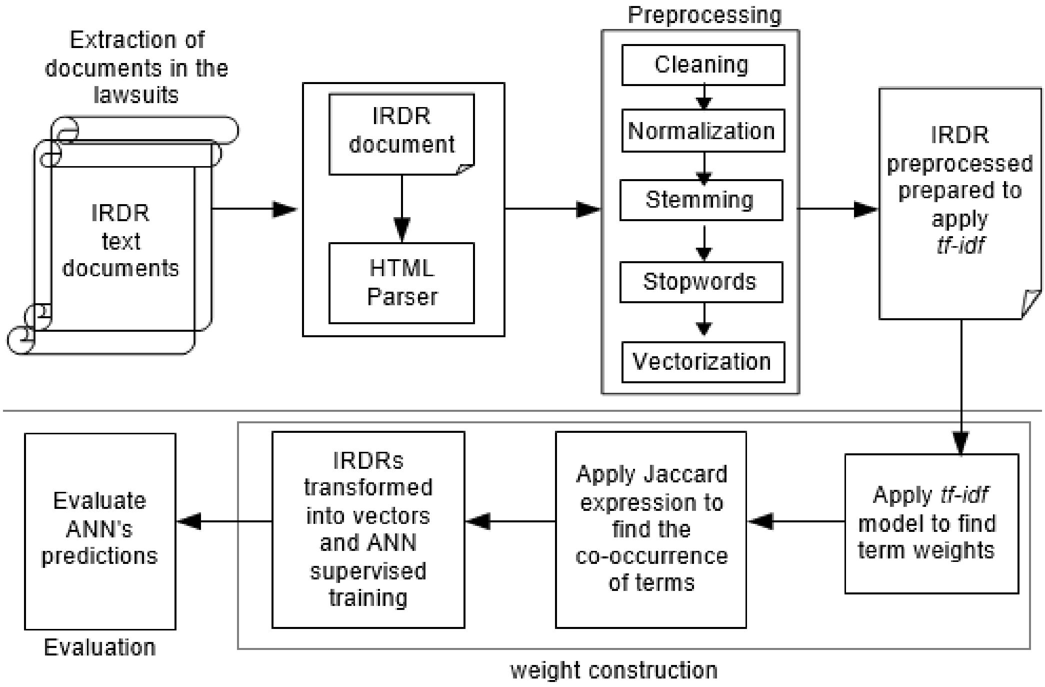
$$Jaccard(D_1, D_2) = \frac{|V_{D1} \cap V_{D2}|}{|V_{D1} \circ V_{D2}|} = \frac{|V_{D1} \cap V_{D2}|}{|V_{D1}| + |V_{D2}| - |V_{D1} \cap V_{D2}|} \quad (1)$$

Where V_{D1} and V_{D2} are vectors containing the tokens of your D1 and D2 documents. The result of the expression (1) is presented in percentage of similarity of document D1 with document D2.

2 Methodology

The methodology developed in this paper has the main object of providing celerity in the judgments of the judicial proceedings because it identifies and classifies the received lawsuits, allowing connection with already judged and consolidated cases (IRDRs). The method proposed collaborates with researches in the field of machine learning in text classification, in three aspects: (a) using the IRDRs of the Court of Justice of Goiás for training in AI solutions; (b) after training the AI solution, the solution will be able to relate the consolidated IRDRs to the lawsuits that come to court, and (c) develop an integration solution with the electronic process system to inform magistrates before the judgment of lawsuits. The stages to be implemented are: i) select documents from some processes related to IRDRs, such as initial petitions and judgments, ii) preprocessing documents to remove tags, unnecessary HTML resources, articles, prepositions, pronouns, commonly used legal terms, and common special characters, iii) separation of terms by term frequency-inverse document frequency model (*tf-idf); iv) application of the similarity metric in (2) to find co-occurrence of* terms and represent the documents in vectors with the weights of the co-occurrence of terms, and vi) supervised training of classification model. The flow constructed to apply the proposed methodology is illustrated in Figure 3.

Figure 3. Overview of the proposed methodology.


Extraction of documents in the lawsuits
IRDR text documents
IRDR document
HTML Parser
Preprocessing
Cleaning
Normalization
Stemming
Stopwords
Vectorization
IRDR preprocessed prepared to apply *tf-idf*
Apply *tf-idf* model to find term weights
Apply Jaccard expression to find the co-occurrence of terms
IRDRs transformed into vectors and ANN supervised training
Evaluate ANN's predictions
Evaluation
weight construction

3.1 IRDR text documents

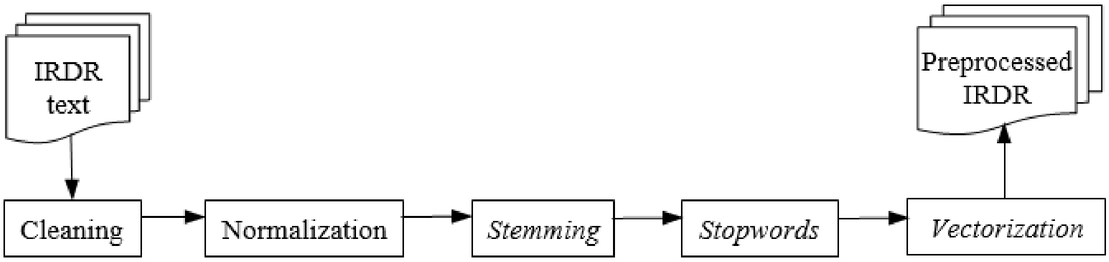
Currently, the Court of Justice of Goiás has 24 IRDRs, 6 IRDR themes were separated for analysis and validation of the proposed AI methodology, with themes numbered 16, 19, 20, 21, 22, and 23. The Court of Justice informed that they are the topics that occur most nowadays, despite being empirical information for research, it serves to validate the machine learning methodology and solution. With the definition of the themes that will be used, it is necessary to separate the documents from the lawsuits to apply the computational model that can represent the document in a vector of weights, to later train the machine learning solution to predict whether a new document is related to one of the five IRDRs used in this research. In this step, several documents from each of the six IRDRs were extracted from certain lawsuits.

Documents extracted directly from the Projudi database are in HTML format. Before moving on to the preprocessing phase, it is necessary to extract this HTML language from the IRDR documents. For this, an HTML extractor of the Ruby-on-Rails programming language was used.[5]

2.2 Preprocessing

In the context in which the universe of unstructured documents *D of the court of* justice is inserted, it is necessary to apply preprocessing, since there are countless words/terms that are not necessary and may interfere with the proposed method. Knowing in advance that the electronic judicial texts for analysis and processing are voluminous and in great quantity for the project at hand, and that they are not structured in standardized models, it is necessary to apply computational techniques of Natural Language Processing (NLP). NLP is a field of artificial intelligence that enables computers to analyze, manipulate and interpret human language. After the step of separating the IRDR´s documents, the preprocessing methods generally used in NLP are applied to each of the received judicial texts, according to the sequence in Figure 4.

Figure 4. NLP preprocessing for incoming IRDR text documents.



a) cleaning, remove special characters, punctuation marks, parentheses, hyphens and other characters of no value in context. After converting the document from HTML format to text raw format, it is necessary to clean and delete several language-specific. It's common in all judgments documents

[5] The HTML parser function used was *ActionView::Base.full sanitizer.sanitize()*, from the Ruby-on-Rails programming language.

to include special characters, like ampersand, double quotes, left and right single quotation, ordinal indicator, section sign, vertical bar, semicolon, brace, bracket, hyphen, and others, they also need to be removed;

b) normalization, remove accents, cedilla and put all words in lower case. Words with the same meaning are also treated in this phase of normalization, such as pharmacy and drugstore. This solution was implemented by the authors and is available on GitHub[4;]

c) stemming, to reduce inflected or derived words to their base. Several Latin words are found in the judgment texts. At first, a module was developed to convert Latin into Portuguese, but it became clear, after the first simulations, that it was not necessary. In the similarity learning technique, the objective is to find situations of combined words that allow classifying texts, so words in Latin can be useful in this process. So, the conversion Latin to Portuguese module was no longer used;

d) stopwords, remove articles, prepositions, pronouns, conjunctions without meaning, words that have no semantic meaning to the text, including in this preprocessing common words in the area of law without importance to the context. Examples of words in the field of law, in Portuguese, that are not relevant to the ranking method are: *autor, réu, requerente, requerido, demandante, demandado, decisão, sentença, fls, PRI, juiz, direito, intimem-se, código, processo, civil, penal, petição, autos* and others;

e) vectorization, transforming the text into a word co-occurrence vector, with their respective weights, removing their repetitions. Since this step occurred after several others during preprocessing, the vector is now ready to be used by the text classification algorithm.

The authors did not find open solutions on the Internet that preprocess common legal terms in Portuguese, so the authors developed a solution that removes irrelevant legal terms from the classification of texts and make the solution available for free on the Internet, in the GitHub[6] repository. This is one of the contributions of this article to the AI research community in the area of Law in Brazil. Preprocessing is done in the entire source text. After this preprocessing step, treated text is generated and saved, facilitating the analysis of the next step.

2.3 Defining weight of terms in IRDR documents

Transforming the court document into a computer model for the machines to understand is an important phase in the AI process. The *tf-idf model* transforms documents of a given category into structures that allow machine learning techniques to learn and recognize other documents in the same category. The *tf-idf model transforms the court document into a* vector of document terms, applying weights to each term. Salton and McGill (1986), and Ali Qaiser and Ali (2018) describes how the *tf-idf metric* works in the construction of the weighting terms. The term is the same thing as the word.

However, the *tf-idf model will not be applied alone to generate the vector that* represents the document. This article uses the vector with the *tf-idf model together with the* Jaccard similarity to computationally represent court documents.

2.4 Transforming IRDR documents into vectors with their terms co-occurring

The *tf-idf model* finds weight for each term, but cannot find weight in the relationship between terms in court documents, that is, it cannot represent the co-occurrence of terms in the vector (Agarwal et al.: 2020; Seo et al.: 2020; Li et al.: 2020).

[6] https://GitHub.com/apcastrojr/court_of_law_pre_processing

The model in (1) is changed so that it can identify weights in the co-occurrence of terms in the document. Thus, although the model in (1) calculates the similarity between documents D1 and D2, this work changes this expression so that it is possible to calculate the similarity between the terms, expression in (2). The *tf-idf model finds the weight of each term* and the model in (2) analyzes the similarity coefficient between the terms, analyzing 2 to 2. Thus, in each field of the computational vector, we do not only have the weight of each term for representing the document, but the weight of the co-occurrence of the terms. The *tf-idf model finds the weight for each term, while Jaccard's altered expression in (2) analyzes the* relevance of the co-occurrence of the terms to be used in the vector that computationally represents the text document.

$$Jaccard(t_i, t_j) = \frac{|t_i \cap t_j|}{|t_i \cup t_j|} = \frac{|t_i \cap t_j|}{|t_i| + |t_j| - |t_i \cap t_j|} \qquad (2)$$

Where t_i e t_j are the values of terms found by the *tf-idf* model in the set of IRDRs documents of the same category, same theme. The terms t_i e t_j represent all *n*terms, *i*= *j*={1,2,3 , … , *n*}, found with a coefficient ≥ 50%. In this sense, the vector that represents the document D_1 is formed by the weight of the co-occurrence of the term t_i e t_j in document 1. The co-occurrence weight is the greatest weight between t_i e t_j, therefore, the vector of D_1, known as V_{D1} is formed by the co-occurrence terms and not just by the occurrence of a term.

This method created to represent documents improves the identification of documents of a technical nature, such as the documents used in this research, as they are text documents prepared by professionals in the field of law.

2.5 Artificial Neural Networks

In this work, Artificial Neural Networks (ANNs) are trained to recognize text documents and classify which IRDR that document belongs to. Samples from each text document need to be used to carry out the ANN training. Examples of IRDR documents are petitions and case judgments. These samples are called datasets. Table 1 presents the dataset used in this work, which are the IRDRs categories and the number of text documents.

The ANN used is a multilayer perceptron with backpropagation neural network (MLPNN) and the training applied is the supervised, having the activation function the logistic. After training, the MLPNN is prepared to predict new lawsuits, by the Petition.[7] To train this neural network, it is necessary to feed it with several vectors, V_{D1} ,V_{D2},V_{D3},… , V_{Dm}

, where m is the total number of documents used in the supervised training. In addition to the weight vectors, it is necessary to inform the neural network which is the category of each vector or each document. Table 1 presents the number of text documents for each IRDR category, used in the supervised training to MLPNN.

Table 1. Category and number of IRDR documents used in MLPNN supervised training.

Category of IRDR document	Number of IRDR documents
16	77
19	30
20	35
21	101
22	115
23	50
Total	408

[7] Petition is the document that initiates the lawsuits and describes the history, related legal codes, and request of the claimant. A Petition is a document prepared by lawyers and starts lawsuits

3 Results

This section presents the results obtained by the proposed methodology, using artificial intelligence applied to predict IRDRs in lawsuits. The analysis of the proposed methodology uses real documents from the Court of Justice of Goiás and goes through two different types of tests, namely: (a) one to measure the accuracy[8], a precision[9], and recall[10] in the application of the proposed AI method and the other (b) use the proposed AI method to identify, in the database of initial petitions that reach the Judiciary, lawsuits that are related to the categories of IRDRs used in the work.

3.1 IRDR Documents

The documents reported in Table 1 were taken from lawsuits currently suspended in the Goiás Judiciary. The documents extracted from the lawsuits are the initial petitions and the judgments that suspended the proceedings. The number of lawsuits in each IRDR is listed in Table 2.

Table 2. Number of lawsuits currently suspended in each IRDR.

Category of IRDR	Number of lawsuits suspended
16	22
19	18
20	24
21	20
22	49
23	17

Table 3 shows the number of words found in each IRDR category, without applying preprocessing, with the total number

[8] Accuracy indicates the percentage, among all classifications, how many were performed correctly by the AI model.

[9] Precision is the number of relevant documents retrieved by a search divided by the total number of documents retrieved by that search.

[10] Recall is the number of relevant documents retrieved by a search divided by the total number of existing relevant documents.

of words and the number excluding term repetitions. Table 3 presents the size of the data volume that the proposed AI model will process.

Table 3. Number of terms found in the datasets shown in Table 1.

Category of IRDR document	Number of terms in IRDR documents with repetitions	Number of terms in IRDR documents without repetitions
16	1.954.184	272.976
19	176.359	49.975
20	70.204	19.975
21	1.834.208	334.850
22	711.078	244.784
23	801.626	198.536

Comparing the columns of quantities with repetitions and without repetitions of terms in Table 3, the high number of repetitions of terms in court documents is observed. This volume makes human analysis difficult in an eventual IRDR classification process. However, the machines do not present difficulties and can process in seconds.

3.2 Preprocessing and vectorization results

The preprocessing step is important in any AI methodology applied to text classification, as there are several terms that do not contribute to document identification. As an example of the relevance of using the preprocessing phase, the *tf-idf model was applied to* the dataset in Table 1, and the result is shown in Table 4.

Table 4. Shows the terms in Portuguese with the highest weights, calculated by *tf- idf*, before and after applying preprocessing.

Category of IRDR document	Five terms with higher weights without applying preprocessing	Five terms with higher weights applying preprocessing

16	da=1507.3780175395457
	o=1779.9643935013253
	e=1936.8493658185846
	a=2069.1843960166007
	de=2398.6624764572125
		salarial=677.1682077966802
		piso=702.9594783027799
		magistério=707.0299683014484
		profissional=734.4971848040876
		educação=904.2878448679812

19	da=232.19596561244217
	o=236.8430892811736
	do=271.883321179333
	a=278.7122776860285
	de=382.42644169034935
		fiscal=91.3242976647417
		tributaria=93.74484314675333
		imposto=101.95340432691351
		multa=118.77924994146105
		valor=145.25902359736176

20	da=105.01671846066877
	o=131.6498002069678
	a=133.173472021566
	do=152.58672311309863
	de=187.43842062965777
		valor=53.67830215840388
		fiscal=58.15739948745119
		advocatícios=59.03978576215159
		execução=68.5797812080749
		honorários=77.16947504824773

21	do=1686.067133079404
	e=1709.6878640445361
	o=1739.82235477815
	a=2054.9202125460774
	de=2458.328760124574
		pagamento=703.0455386490694
		credito=742.500213879804
		contrato=742.7324755639847
		emprestimo=754.1780441180449
		valor=851.3693213781881

22	do=1341.1722111301603
	e=1412.5676513486078
	o=1427.0204702889182
	a=1598.2171280381306
	de=1919.3965114624932
		atraso=475.04028879609905
		valor=479.65251498434026
		entrega=483.2755595334913
		dano=630.0928299991141
		contrato=743.3316019925719

23	o=888.58459382515
	e=902.0948701588957
	a=1021.9512556160885
	do=1036.2921605675076
	de=1496.1431221996831
		fomentar=305.3113737265507
		imposto=319.8586354530146
		fiscal=357.4889234785092
		repasse=396.61779546578833
		icms=460.61496815307123

Table 4 lists five terms that had greater weights generated by the *tf-idf model* without and with the application of preprocessing in IRDR documents. It is noticed that without the application of preprocessing, the five terms with the highest weights are common terms, found in any text and that do not help the AI method in differentiating the categories. It is also noticed that the terms found without applying preprocessing are identical between the different categories of IRDRs. Now, applying the preprocessing to later calculate the weights by the *tf-idf model, it is noticed that there are different words between the different categories* of IRDRs and related to the themes, which helps the AI method in the classification of texts. Thus, the importance of the preprocessing step in the proposed method is clear.

Preprocessing is performed using routines built in the Ruby language. These libraries were made available on GitHub,[11] under GNU General Public License v 3.0, aiming to share them in new research. The text documents used are petitions and judgments made by judges of the Court of Justice.

Following the flow of the method proposed in Figure 3, after finding the weights of terms by the *tf-idf model,* Jaccard's expression (2) is applied to find the co-occurrence of terms with a similarity coefficient greater than 50%. Table 5 shows five examples of co-occurrence of terms found by the proposed model, with their similarity coefficients, for each category of IRDR documents. A coefficient close to 1 means that the co-occurrence of terms generally occurs in the same IRDR document category.

Tabela 5. Five examples of co-occurrence of terms, in Portuguese, found by the expression of Jaccard in (2).

IRDR documents	Co-occurrence of terms in Portuguese and similarity coefficients found
16	atividades-auxiliar=0.96, profissional-atividades=0.96, profissional-básica=0.97,magistério-básica=0.99, piso-salarial=0.99, ...
19	imposto-cálculo=0.88, imposto-credito=0.91, imposto-icms=0.91, icms--credito=0.92,credito-cálculo=0.95, ...
20	aplicação-advocatícios=0.95, fixados-pagamento=0.95, fixados-salários-mínimos=0.95,fiscal-aplicação=0.95, fiscal-advocatícios=0.99, ...
21	consignado-consumidor=0.94, cartão-consumidor=0.95, valor-pagamento=0.95,consignado-cartão=0.96, emprestimo-bancario=0.98, ...

[11] https://GitHub.com/apcastrojr/court_of_law_pre_processing

22	entrega-multa=0.94, atraso-multa=0.94, empreendimento-prazo=0.94, dream-park=0.98,entrega-atraso=0.98, ...
23	fomentar-programas=0.99, imposto-constitucional=0.99, fomentar-produzir=0.99, valor-fomentar=0.99, produzir-programas=1.0, ...

The two-by-two terms shown in Table 5 are used to structure the vectors. These vectors are used to train artificial neural networks (MLPNN). With MLPNN training, it is necessary to assess the accuracy, precision, and recall of the proposed solution.

3.3 Accuracy, precision, and recall metric results of the proposed AI model

To measure the accuracy, precision, and recall of the method proposed in section 3, the dataset used is shown in Table 1. The sample of IRDR documents in Table 1 is randomly divided into IRDRs for training and IRDRs for testing, applying the percentage of 80% of documents for training and 20% of documents for testing, as shown in Table 6.

Table 6. Number of IRDR documents used for training and testing.

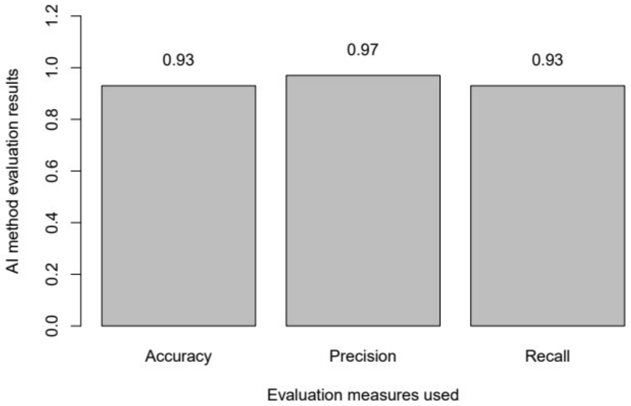
Category of IRDR document	Number of IRDR documents for training	Number of IRDR documents for testing
16	62	15
19	24	6
20	28	7
21	81	20
22	92	23
23	40	10
Total	327	81

Computer programs were developed using Ruby-on-rails and Python languages. The weighting construction using *tf-idf and* similarity-learning expression in (2) were developed in Ruby-on-

Rails. The vectors imports and classification models were developed in Python. The *tf-idf-similarity ruby gem is used to implement term frequency-inverse document* frequency (*tf-idf). The similarity-learning using Jaccard in (2) was built by the authors using* Ruby-on-Rails. The Scikit-learn package is used to implement multilayer perceptron with backpropagation neural network (MLPNN).

After applying the proposed methodology, as shown in Figure 3, in its last step, evaluation, the values for the accuracy, precision, and recall metrics were obtained, as shown in the graph in Figure 5. It is observed that the values achieved of 93% in accuracy, 97% in precision, and 93% in recall, in the tests performed, demonstrate the success of the method proposed in this article and give the necessary confidence to apply the studies in a production environment in the Court of Justice.

Figure 5. Results of assessment metrics obtained from simulations performed on real IRDR court documents.



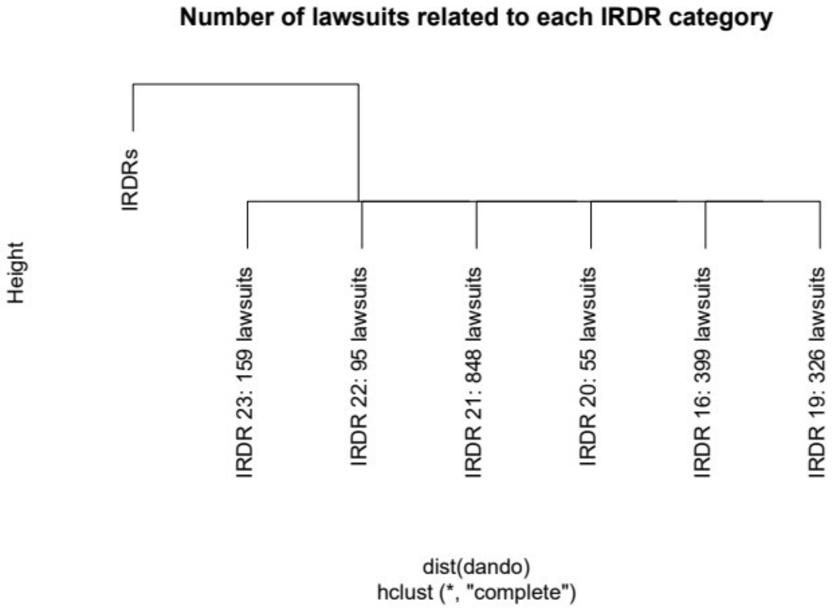
Regarding processing time, the proposed model was able to analyze and test all 408 IRDR documents, reported in Table 1, in less than 50 seconds, generating the metric values stamped in Figure 5.

3.4 Integration of the AI solution in the day-to-day of justice

It is relevant to integrate the proposed AI method to identify daily lawsuits in the electronic system of the judiciary in Goiás, Brazil. Castro et al. (2020) developed a solution that searches the judicial petitions in the Projudi system, since 05/17/2020, and files these petitions in the IA solution database of the work of Castro et al. (2020), known by the name Berna. Berna's database currently has 514,347 initial petitions filed, all in HTML format. This judicial initial petition basis is used by the methodology proposed in this paper to identify whether any of them are related to IRDRs of categories 16, 19, 20, 21, 22 and 23.

With MLPNN classification model training, the AI solution will analyze all 514,347 text documents and report the amount found by IRDR category. The Dendrogram in Figure 6 shows the number of lawsuits found in Projudi's database related to each IRDR category.

Figure 6. Number of lawsuits found by the AI method related to each IRDR category.



After identifying the lawsuits related to the IRDRs applied in this study, the AI solution informs the magistrate or secretary, to which the case was distributed, the lawsuit numbers and the respective IRDRs. The information reaches the magistrate or secretary through the pending functionality in the Projudi system. The magistrate or secretary of the court will analysis the pending issue in Projudi, being able to dismiss it or admit it as true. Thus, the AI solution will help the judiciary in the day-to-day study and analysis of lawsuits, acting as a court clerk, with agility and speed in the processing, reading, and reasoning of documents. This is how the AI solution fits into the everyday life of court.

4 Discussion

The work proposes a method that transforms unstructured technical legal documents into known computer-manipulated data structures, called vectors. The method fills the vector with the weights, values, of the co-occurrence of terms in the documents. The vector represents the document and is used in the training of the artificial neural network that starts to learn to identify other new documents.

In the simulations, 408 legal technical documents, initial petitions and judgments are used, separated into 6 categories of IRDRs. These legal documents come from 150 different lawsuits, as shown in Table 2. In the tests, various information was extracted from these legal documents, such as (a) the total terms in all documents add up to more than five million, (b) if we count the terms without the repetitions, the value is low to one million and one hundred thousand, that is, 80% of the terms are repeated, as shown in Table 3. It is noticed that in legal documents there is a considerable volume of words or terms commonly used by professionals in the field of law.

In the studies carried out, it is clear that preprocessing is an important step for the classification of legal documents. Table 4 shows that when the preprocessing step is not applied, the terms with greater weight are articles, conjunctions, and prepositions that do not collaborate in the classification process of new legal documents. However, applying to the preprocessing, the five terms with higher weights are relevant words in identifying the IRDR categories. The results in Table 4 show the relevance of preprocessing. This work contributes to new studies in the area of law, as it develops and makes available on the Internet the source code of the preprocessing software built and applied in this paper in the GitHub repository,[12] it is a preprocessing of legal texts in Portuguese.

Table 5 presents five examples of co-occurrence of terms and their weights found by the proposed AI model. The relationship of the two by two terms (co-occurrence), identified for each IRDR theme, is used to train the MLPNN network. After training the neural network, you can use it to recognize new documents. Table 6 shows the number of documents used for

(a) training and (b) classification tests. For evaluation, accuracy, precision, and recall metrics are used, and the results are shown in Figure 5.

The results of the simulations, carried out on real documents, show that the AI model proposed in the article can predict which IRDR category a new legal document belongs to, meeting the main objective of the paper. The percentage of 97% in precision, 93% in accuracy, and 93% in recall brings confidence and credibility in the application of the methods and their integration with the electronic software that controls court lawsuits in Goiás, Brazil. When applying the proposed method to ongoing cases in Projudi, the number of lawsuits listed in each IRDR category is identified. Figure 6 shows the values by IRDR category.

[12] https://GitHub.com/apcastrojr/court_of_law_pre_processing

5 Conclusion

The studies developed in this article are relevant because they are applied research in a Court of Justice in Goiás, Brazil. Kim and Toffoli (2019) report that the Court of Justice of Goiás has a volume of approximately two million cases under study, with four hundred thousand new cases being received each year, with approximately three hundred and eighty judges to judge them. Bolsonaro et al. (2020) report that the estimated population in 2020 in Goiás is 7,113,540 people. The works are applied in this scenario.

Recent changes in Brazilian legislation have facilitated the use of information technology resources to streamline the progress and judgment of cases, such as IRDRs. In this context, the application of artificial intelligence to relate the new lawsuits with IRDRs already handed down would become an important tool for judges, as they can speed up the judgment, linking related lawsuits already judged and avoiding divergent judgments for related lawsuits. For the integration, an API was installed and is currently working in the Court of Justice.

This work proposes an AI method to computationally represent legal documents and a model for text classification. The simulation results with values of 93% accuracy, 97% precision, and 93% recall indicate that the proposed AI method was successful and brings confidence in its application in a production environment in Goiás Justice. With these numbers, it is possible to affirm that the objective of this work was achieved.

The integration software developed in the work by Castro et al. (2020) is used to integrate the AI model of this work to the electronic process software, called Projudi, from the court in Goiás, Brazil. With this integration, the AI solution analyzes the initial petitions that reach the Court of Justice, being able to classify them in some IRDR. In cases where there is identification, the AI solution informs the judge or secretary that the judicial process is related to the IRDR of a certain category.

It is believed that the AI model applied will expedite the progress of legal proceedings, if the judge understands that he can: (a) suspend the lawsuit based on the informed IRDR, (b) analyze the precedent informed by the category, if it meets the provisions of the article 332 of the CPC, (c) analyze the feasibility of applying Pronouncement FONAJE 73, in cases related to Special Courts, and (d) establish new routines and management of lawsuits in judicial units.

Regarding the difficulties found, the quality of the documents used in the processing of the proposed method is highlighted. The proposed model finds problems in identifying the entire content of some pieces, not allowing the extraction of their terms. It is noticed that some documents are inserted in the court system as images and without the care of correctly registering their typology. As these documents are the raw material of the AI model, it is necessary to establish stricter standards in the act of filing the lawsuit documents.

Work continues to be developed to include new IRDR themes as well as STJ themes and standardization of jurisprudence. Still, other text classification models are also under study and development. Word embedding solutions are being evaluated and compared with the results of this work.

References

A. Bosch, A. Zisserman, X. Munoz, Scene classification using a hybrid generative/ discriminative approach, IEEE Trans. Pattern Analysis and Machine Intelligence, 30 (2008).

A. Devitt, B. Danev, K. Matusikova, Constructing Bayesian networks automatically using ontologies, Applied Ontology Journal, ISSN 1570–5838 (2006).

A. Grubisic, S. Stankov, I. Peraic, Ontology based approach to Bayesian student model design, Expert systems with applications 40 (2013) 5363–5371.

A. P. Castro, W. P. Calixto, C. H. Araujo, Application of artificial intelligence in the identification of connections by fact and thesis in the judicial complaint and integration with the electronic system of lawsuits (in Portuguese), CNJ Magazine 4 (2020) 10.

A. P. Castro, W. P. Calixto, V. M. Gomes, E. F. Veiga, L. F. Silva, L. L. O. P. Castro, J. L. F. Barbosa, P. H. Campos, Ontology applied in the judicial sentences, in: 2017 CHILEAN Conference on Electrical, Electronics Engineering, Information and Communication Technologies (CHILECON), IEEE, pp. 1–6.

A. P. Castro, W. P. Calixto, V. M. Gomes, E. F. Veiga, L. F. Silva, P. H. Campos, Ontology to mining judicial sentence´s big data, in: Alive Engineering Education, UFG, 2017a, pp. 187–196.

B. Smith, Ontology: philosophical and computational, The Blackwell Guide to the Philosophy of Computing and Information. Blackwell, Oxford (2003).

Bolsonaro, J. M.; Guedes, P.R.N.; Junior, W.R.; Guerra, S.C.; de Araújo Abrantes,

C. L. A. Rocha, Justice in Numbers: document produced by the Brazilian judiciary (in Portuguese), Digital Magazine of the National Council of Justice (in Portuguese) 1, Conselho Nacional de Justiça – CNJ, 2018.

C. Manning, P. Raghavan, H. Schutze, Introduction to information retrieval, Natural Language Engineering 16 (2010) 100–103.

C. N. Mooers, Zatocoding applied to mechanical organization of knowledge, American documentation 2 (1951) 20–32.

Corcho, M. Fernandez-Lopez, A. Gomez-Perez, Methodologies, tools and languages for building ontologies. where is their meeting point?, Data & knowledge engineering 46 (2003) 41–64.

D. N. Milne, I. H. Witten, D. M. Nichols, A knowledge-based search engine powered by wikipedia, In Proceedings of the 16th Association for Computing Machinery (ACM) Conference on Information and Knowledge Management (2007).

E. Gabrilovich, S. Markovitch, Feature generation for text categorization using world knowledge, In Proceedings of the 19th International Joint Conference on Artificial Intelligence (2005).

F. Coimbra Delicato, L. Pirmez, L. Fernando Rust da Costa Carmo, Fenix– personalized information filtering system for WWW pages, Internet Research 11 (2001) 42–48.

F. Glitz, Incoterms and Brazilian legislation on contracts, Education and Science without borders Journal 2 (2011) 40–44.

F.J. Estimates of the resident population (In Portuguese). Technical report, IBGE, 2020.

Fawei, A. Wyner, J. Z. Pan, M. Kollingbaum, Using legal ontologies with rules for legal textual entailment, in: AI Approaches to the Complexity of Legal Systems, Springer, 2015, pp. 317–324.

G. Salton, Automatic text processing: The transformation, analysis, and retrieval of information, Reading: Addison-Wesley (1989).

G. Salton, M. J. McGill, Introduction to modern information retrieval, McGraw- Hill, Inc., 1986.

Giaretta, N. Guarino, Ontologies and knowledge bases towards a terminological clarification, Towards very large knowledge bases: knowledge building & knowledge sharing 25 (1995) 307–317.

Guarino, D. Oberle, S. Staab, What is an ontology?, in: Handbook on ontologies, Springer, 2009, pp. 1–17.

J. C. C. Moura, M. T. C. Sousa, Towards judiciary: brief psychoanalyst and historical considerations about voluntary subjection to the law and judiciary (in Portuguese), Cad. Pesq., Sao Luis, v. 20, n. 3 (2013).

J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, Computing Research Repository (CoRR) (2018).

J. M. Ponte, W. B. Croft, A language modeling approach to information retrieval, Ph.D. thesis, University of Massachusetts at Amherst, 1998.

J. M. Ponte, W. B. Croft, A language modeling approach to information retrieval, in: ACM SIGIR Forum, volume 51, ACM, pp. 202–208.

J. Pennington, R. Socher, C. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2014, pp. 1532– 1543.

J.-B. Lamy, Owlready: Ontology-oriented programming in python with automatic classification and high level constructs for biomedical ontologies, Artificial intelligence in medicine 80 (2017) 11–28.

K. P. Murphy, Machine Learning: A Probabilistic Perspective, The MIT Press (Cambridge Massachusetts), 2013.

L. Huang, D. Milne, E. Frank, I. H. Witten, Learning a concept-based document similarity measure, Journal of the American Society for Information Science and Technology, ISSN 1532–2882 (2012).

L. Uusitalo, Advantages and challenges of Bayesian networks in environmental modelling, Ecological modelling 203 (2007) 312–318.

M. A. Calambas, A. Ordnez, A. Chacon, H. Ordonez, Judicial precedents search supported by natural language processing and clustering, in: 2015 10th Computing Colombian Conference (10CCC), IEEE, pp. 372–377.

M. B. Almeida, Revisiting ontologies: A necessary clarification, Journal of the American Society for Information Science and Technology 64 (2013) 1682–1693.

M. B. Almeida, Revisiting ontologies: A necessary clarification, Journal of the American Society for Information Science and Technology 64 (2013) 1682–1693.

M. Boughanem, A. Brini, D. Dubois, Possibilistic networks for information retrieval, International Journal of Approximate Reasoning 50 (2009) 957–968.

M. Ceci, A. Gangemi, An owl ontology library representing judicial interpretations, Semantic Web 7 (2016) 229–253.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: North American Chapter of the Association for Computational Linguistics - NAACL, ACM Digital Library, 2018.

M. Horridge, S. Jupp, G. Moulton, A. Rector, R. Stevens, A practical guide to building owl ontologies using protege 4 and co-ode tools edition v. 1.3, The University of Manchester 178 (2011).

M. Rani, A. K. Dhar, O. Vyas, Semi-automatic terminology ontology learning based on topic modeling, Engineering Applications of Artificial Intelligence 63 (2017) 108– 125.

N. Agarwal, G. Sikka, L. K. Awasthi, Enhancing web service clustering using length feature weight method for service description document vector space representation, Expert Systems with Applications Journal (2020).

N. Guarino, Formal ontology in information systems: Proceedings of the first international conference (FOIS’98), June 6-8, Trento, Italy, volume 46, IOS press, 1998.

N. Zhang, P. Wang, Y. Pu, Challenges and related issues for building Chinese legal ontology, in: 2015 International Conference on Mechanics, Electronic, Industrial and Control Engineering (MEIC-15), Atlantis Press, pp. 1260–1265.

N. Zhang, Y.-F. Pu, S.-Q. Yang, J.-L. Zhou, J.-K. Gao, An ontological chinese legal consultation system, IEEE Access 5 (2017) 18250–18261.

O. Arsene, I. Dumitrache, I. Mihu, Medicine expert system dynamic Bayesian network and ontology based, Expert Systems with Applications 38 (2011) 15253–15261.

P. Castro, W. P. Calixto, B. F. Franco, Information Management in Big Data Volumes in the Judiciary (in Portuguese), volume V, V Luso-Brazilian Collection - Information Management, Cooperation in networks and Competitiveness, 2014.

P. Li, K. Maoa, Y. Xu, Q. Li, J. Zhang, Bag-of-concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base, Knowledge-Based Systems (2020).

P. Li, K. Maoa, Y. Xu, Q. Li, J. Zhang, Bag-of-concepts representation for document classification based on automatic knowledge acquisition from probabilistic knowledge base, Knowledge-Based Systems (2020).

R. David, Major Legal Systems in the World Today (in Portuguese), Martins Fontes, 2014.

R. Fergus, P. Perona, A. Zisserma, Object class recognition by unsupervised scale- invariant learning (2003).

R. Mihalcea, C. Corley, C. Strapparava, Corpus-based and knowledge-based measures of text semantic similarity, In Proceedings of the 21st National conference on Artificial Intelligence (2006).

R. P. Kim, J. A. D. Toffoli, Justice in Numbers: document produced by the Brazilian judiciary (in Portuguese), Digital Magazine of the

National Council of Justice (in Portuguese) 1, Conselho Nacional de Justiça – CNJ, 2019.

R. P. Kim, J. A. D. Toffoli, Justice in Numbers: document produced by the Brazilian judiciary (in Portuguese), Digital Magazine of the National Council of Justice (in Portuguese) 1, Conselho Nacional de Justiça – CNJ, 2020.

S. Blackburn, The Oxford dictionary of philosophy, OUP Oxford, 2005.

S. Buttcher, C. L. Clarke, G. V. Cormack, Information retrieval: Implementing and evaluating search engines, Mit Press, 2016.

S. H. Bailey, M. J. Gunn, The Modern English Legal System, Sweet & Maxwell, 5th edition edition, 2007.

S. Jasanoff, Science, common sense judicial power in u.s. courts, Daedalus – Journal of the American Academy of Arts & Sciences (2018).

S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories (2006).

S. Qaiser, R. Ali, Text mining: Use of *tf-idf to examine the relevance of words to* documents, International Journal of Computer Applications 181 (2018).

S. Seo, D. Seo, M. Jang, J. Jeong, P. Kang, Unusual customer response identication and visualization based on text mining and anomaly detection, Expert Systems with Applications Journal (2020).

S. Staab, R. Studer, Handbook on ontologies, Springer Science & Business Media, 2010.

T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, Computing Research Repository (CoRR) (2013).

V. C. C. Donato, The Judiciary in Brazil: structure, criticism and control (in Portuguese), Ph.D. thesis, University of Fortaleza (UNIFOR) – Brazil, 2006.

V. Garg, S. Vempati, C. V. Jawahar, Bag of visual words: A soft clustering based exposition, Third National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (2011).

V. Thada, V. Jaglan, Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm, International Journal of Innovations in Engineering and Technology 2 (2013) 202–205.

Y. Kim, Convolutional neural networks for sentence classification, Computing Research Repository (CoRR) (2014).

Y. Wand, R. Weber, Mario bunge's ontology as a formal foundation for information systems concepts, Studies on Mario Bunge's treatise 18 (1990).

Y. Wand, V. C. Storey, R. Weber, An ontological analysis of the relationship construct in conceptual modeling, ACM Transactions on Database Systems (TODS) 24 (1999) 494–528.

Acknowledgements

The authors would like to thank National Council for Scientific and Technological Development (CNPq), Foundation for Research Support of the State of Goiás (FAPEG),

Brazilian Federal Agency for Support and Evaluation of Graduate Education (CAPES) and Court of Justice of the State of Goiás, Brazil.

Additional information Competing interests

The author(s) declare no competing interests