Revista de Casos e Consultoria, v. 15, n. 1, e36701, 2024
ISSN 2237-7417 | CC BY 4.0

Processamento de Linguagem Natural como ferramenta de suporte em documentos jurídicos: uma revisão sistemática

Natural Language Processing as a support tool in legal documents: a systematic review

Procesamiento de Lenguaje Natural como herramienta de apoyo en documentos legales: una revisión sistemática

Recebido: 21/06/2024 | Revisado: 27/06/2024 | Aceito: 20/08/2024 | Publicado: 21/08/2024

Filipe Machado da Costa Barros
ORCID: https://orcid.org/0000-0003-1879-4200
Universidade Federal do Pará - UFPA, Brasil
E-mail: filipe.barros@tucurui.ufpa.br
Cleison Daniel Silva
ORCID: https://orcid.org/0000-0001-8280-2928
Universidade Federal do Pará - UFPA, Brasil
E-mail: cleison@ufpa.br
Igor Rosberg de Medeiros Silva
ORCID: https://orcid.org/0009-0003-1138-1806
Universidade Federal do Rio Grande do Norte - UFRN, Brasil
E-mail: igor.silva@ufrn.br
Victor Simões Martins
ORCID: https://orcid.org/0000-0002-4789-2646
Universidade Federal do Pará - UFPA, Brasil
E-mail: victorsm@ufpa.br
Antônio Jhoseph Silva de Araújo
ORCID: https://orcid.org/0009-0006-0651-8400
Universidade Federal do Pará - UFPA, Brasil
E-mail: j.garibald@gmail.com

Resumo

A sobrecarga de processos judiciais tem aumentado cada vez mais, interferindo diretamente na execução das atividades nos tribunais. Começamos a buscar auxílio na inteligência artificial, através do uso de ferramentas e técnicas de processamento de documentos, promovendo uma mudança significativa na forma como as atividades jurídicas são realizadas. Nesse sentido, foi realizada uma revisão sistemática, onde foram consultados o Google Scholar, Portal de periódicos Capes, Science Direct - Elsevier e IEEE Xplore. As publicações foram obtidas com o intuito de responder a 4 perguntas: (1) Quais são as publicações científicas mais relevantes relacionadas à aplicação de NLP em documentação jurídica no período de 2017 a 2022; (2) quais técnicas e ferramentas de NLP foram aplicadas no tratamento de documentos no domínio jurídico; (3) o desempenho obtido ao aplicar NLP em novos documentos do âmbito jurídico brasileiro; (4) quais bases de dados jurídicas existentes no contexto brasileiro possuem algum pré-processamento que auxilie a NLP. A literatura recomenda o uso de algoritmos de deep learning para resolver problemas envolvendo NLP, onde sua aplicação, combinada com técnicas de embedding de texto em domínios específicos, melhora significativamente os modelos gerados

Palavras-chave: Processamento de linguagem natural; Documentos judiciais; Aplicação jurídica.

Abstract

The burden of court cases has been increasing more and more, directly interfering with the performance of activities in the courts. We started looking for aid in artificial intelligence, through the use of document processing tools and techniques, promoting a significant change in the way legal activities are carried out. In this sense, a sistematic review was conduced, where Google Scholar, Portal de periódicos Capes, Science Direct - Elsevier and IEEE Xplore were consulted. The publications were obtained in order to answer 4 questions: (1) What are the most relevant scientific publications related to the application of NLP in legal documentation for the period from 2017 to 2022; (2) which NLP's techniques and tools were applied in dealing with documents in the legal domain; (3) the performance obtained when applying NLP in new documents of the Brazilian legal scope; (4) which legal databases exist in the Brazilian context have some pre-processing that helps the NLP. The literature recommends the use of deep learning algorithms to solve problems involving NLP, where their application, combined with embedding text on specific domain techniques, greatly improves the generated models.

Keywords: Natural language processing; Court documents; Legal application.

Resumen

La carga de casos judiciales ha ido aumentando cada vez más, interfiriendo directamente en la ejecución de las actividades en los tribunales. Comenzamos a buscar ayuda en la inteligencia artificial, a través del uso de herramientas y técnicas de procesamiento de documentos, promoviendo un cambio significativo en la forma en que se llevan a cabo las actividades legales. En este sentido, se realizó una revisión sistemática, donde se consultaron Google Scholar, Portal de periódicos Capes, Science Direct - Elsevier e IEEE Xplore. Las publicaciones se obtuvieron con el fin de responder a 4 preguntas: (1) ¿Cuáles son las publicaciones científicas más relevantes relacionadas con la aplicación de NLP en la documentación legal en el período de 2017 a 2022?; (2) ¿qué técnicas y herramientas de NLP se aplicaron en el tratamiento de documentos en el ámbito legal?; (3) el rendimiento obtenido al aplicar NLP en nuevos documentos del ámbito legal brasileño; (4) ¿qué bases de datos legales existentes en el contexto brasileño tienen algún preprocesamiento que ayude al NLP? La literatura recomienda el uso de algoritmos de deep learning para resolver problemas que involucran NLP, donde su aplicación, combinada con técnicas de embedding de texto en dominios específicos, mejora considerablemente los modelos generados.

Palabras clave: Procesamiento de lenguaje natural; Documentos judiciales; Aplicación legal.

Introduction

The applications of Natural Language Processing (NLP) have been gaining momentum in various sectors, including the legal and educational environments. In education, NLP enables the creation of tools like the *Assistente Virtual Educacional (AVE)* (PINTO; ARAÚJO; SANTANA JÚNIOR, 2024), which facilitates communication between students and teachers. These assistants, based on NLP, understand and respond to questions in natural language, providing personalized support in the teaching-learning process. By assisting with tasks, they make education more accessible and effective, allowing students to receive immediate guidance while teachers focus on more complex aspects.

Brazilian courts ended the year 2016 with 79.7 million ongoing cases. In 2017, the National Council of Justice (CNJ), through the data disclosed in the report Justice in Numbers, makes it clear that every year the congestion rate caused by insoluble cases in the Brazilian Judiciary remains at unsustainable levels (FILHO e JUNQUILHO, 2018).

In the search for technological support, the Judiciary has been increasingly using NLP in its work routine, which is an important tool that contributes quickly, cheaply and efficiently to organize data, extract information, and even help with the classification of cases (LUZ DE ARAUJO et al., 2020).

Works developed involving NLP applied to legal documentation reveal important contributions to the sector, such as: increased efficiency and accuracy in the analysis of legal documents, which can lead to a reduction in costs and time. Possibility of extracting relevant information and insights from large volumes of documents, which can help in strategic decision-making. Ease of access and retrieval of information, allowing lawyers and legal professionals to find the necessary information more quickly and efficiently.

Despite the presented benefits and gains, there are still gaps to be overcome in future works, such as: the lack of resources and data for training NLP models specific to the legal domain. The difficulty of dealing with the diversity of terminology and expressions used in legal documents, often specific to a region or country. The challenge of dealing with the complexity and ambiguity of legal language, which often involves long sentences and technical terms. The need to ensure the privacy and security of sensitive data contained in legal documents, which may limit access to this data for training and validating NLP models.

Justice 4.0 is an initiative of the Conselho Nacional de Justiça (CNJ) that seeks to modernize and make legal services more efficient through the use of digital technologies and artificial intelligence. One of the tools that can be used in this context is NLP, which is an artificial intelligence technique that allows machines to understand and interpret human language (VASCONCELOS et al., 2020).

The use of natural language processing as a support tool in legal documents can bring many benefits to Justice 4.0. For example, this technology can be used to analyze large volumes of cases and identify patterns and trends, helping judges and lawyers to make more informed and fair decisions. In addition, natural language processing can be used to improve the search for jurisprudence and laws, which can save time and resources.

However, it is important to highlight that the use of natural language processing cannot replace human analysis in legal documents, since understanding the context and nuances of language is still a difficult task for machines. Therefore, technology should be seen as a support tool for legal decision-making rather than a replacement for legal professionals.

A systematic review proposed by MARTINS e SILVA (2021) addresses the use of NLP for text classification in the legal field. Here, it is noticed that the migration of processes from physical to electronic has been the driving force behind research development in these topics. This study led to the conclusions that in Brazil there is the use of automatic text classification, but it is less frequent when compared to studies conducted in English. It is essential to train language models specifically for the legal context to achieve better results. Additionally, there are two studies that provide Brazilian Portuguese language models and one that introduces a dataset in the Brazilian legal field.

The systematic review conducted here serves as the foundational basis for the research presented by BARROS et al. (2023). This research aims to apply NLP techniques and tools to a dataset maintained by the *Tribunal Regional do Trabalho 8ª Região (Pará/Amapá)* (TRT 8) to develop classifier models capable of identifying which cases have the potential for success in conciliation hearings. The ultimate goal is an evaluative and comparative study of Machine Learning models used to classify labor cases and identify those related to conciliation. The research presents preliminary results to provide comparable models for the development of a study focused on NLP applied to legal documentation in the public sector.

Given the previous considerations, this study aims to answer four questions: (1) What are the most relevant scientific publications related to the application of NLP in legal documentation for the period from 2017 to 2022. (2) Identify the NLP techniques and tools used for processing legal documents. (3)

Evaluate the performance achieved by applying NLP to new documents within the Brazilian legal context, and, finally, (4) identify legal databases in the Brazilian context that offer pre-processing features to assist NLP tasks.

This paper aims to propose a systematic review of natural language processing tools and techniques in legal documentation used in the public sector. The review is organized as follows: section 2 presents and describes the research protocol used in the review, along with the questions that make up the review. In sections 3 to 5, analyses are made on the selected works in order to answer the questions of section 2. Finally, in section 6, the main conclusions and gaps on the subject are discussed, and the bibliographic references that make up this systematic review are listed.

Methodology

In an age of prolific legal data generation, exploring pertinent scientific publications regarding the NLP in legal documents within the public sector over recent years yields valuable insights into evolving trends. Moreover, comprehending the specific NLP techniques and tools employed in processing legal documents facilitates the discovery of inventive strategies to enhance legal processes.

Evaluating the performance of NLP methods within the framework of Brazilian law is pivotal in assessing their viability and potential influence on the local legal landscape. Lastly, delving into the existence of pre-processed legal databases optimized for NLP integration in the Brazilian context not only economizes time but also establishes the groundwork for crafting efficient and contextually pertinent NLP solutions for legal objectives.

Thus, the review protocol was produced with the aim of answering the following review questions:

- Q.1 What are the most relevant scientific publications related to the application of NLP in legal documents in the public sector between the years 2017 and 2022?
- Q.2 How NLP techniques and tools have been applied in dealing with documents in the legal area?
- Q.2 How are NLP techniques and tools (such as TF-IDF, Word2Vec, Long Short-Term Memory, XGBoost, word embedding, BERT) applied in the legal area context?
- Q.3 What is the performance obtained when applying NLP techniques and tools in documents within the Brazilian legal framework?
- Q.4 What legal databases exist in the Brazilian context that already have some pre-processing that helps to apply the NLP?

The searches were carried out in the following databases: Google Scholar, Portal de Periódicos Capes, Science Direct - Elsevier, and IEEEXplore Digital Library.

To collect the articles of interest in this systematic review, search strings were defined with the aim of answering the raised questions. Table 1 presents the associations between each review question and its respective search string.

Questions	Search String
Q.1	“natural language processing” AND (“court documents” OR “legal application”)
Q.2	(“court documents” OR “legal application” OR “textual mining”) AND (“word embedding” OR TF-IDF OR Word2Vec OR "Long Short-Term Memory" OR XGBoost OR BERT)
Q.3	(legal OR "legal domain") AND "natural language processing" AND ("Brazilian legal documents" OR Brazil)
Q.4	"legal documents" AND ("natural language processing" OR "language resources") AND (Brazil OR "Portuguese processing")

Table 1. Search strings associated with each of the Review Questions.

The year 2023 was not included in the systematic review due to the review work being conducted in 2022 and the decision to select works within a timeline of 5 years backward. This methodological choice was made to ensure that the review covered a recent and relevant period, allowing for the analysis of more current trends and developments in the application of natural language processing in legal documents.

The systematic review was conducted from January 2017 to April 2022, totaling 5 years and 4 months. Table 2 demonstrates the filters applied by the search base considering the defined period.

Google Scholar	***Portal de Periódicos Capes***	**Science Direct – Elsevier**	**IEEE Xplore**

Sort by relevance;
10 first pages;
Do not include citations.
	Expand my results;
	10 first pages;
	Online Resource;
	Peer-reviewed journals;
	Open Access;
	Articles, newspaper and newsletter articles;
	English and Portuguese languages.
		Sort by relevance;
		10 first pages;
		Research articles.
			Sort by relevance;
			10 first pages;

Table 2. Filters associated with each search databases.

Following a similar systematic review planning as used by (SPOLAOR et al., 2020), exclusion criteria (EC) were established to carry out the screening process of articles that will be included in this systematic review:

1. Duplicate works;
2. The work does not primarily focus on the subject of NLP applications in legal documentation, and it presents topics directed towards private services, not preserving the idea that the review should have themes related to the public sector.
3. The work consists of only one page (abstract paper), poster, presentation, conference proceedings, or tutorial slides;
4. The work was published outside the period from January 2017 to April 2022;
5. The work does not fit the research questions of the review;
6. The work was written in a language other than English or Portuguese;
7. The work does not have open content or is hosted on web pages that cannot be accessed using the login credentials of the educational institution;
8. The work does not present results from experimental evaluation (quantitative study);

The initial filtering is based on the first two exclusion criteria. Therefore, the works were initially gathered in a total list containing 1028 articles, of which 618 were excluded, as they were not related to the topic of the review or were duplicates, resulting in a list with 410 works.

After that, the remaining titles were read, considering the others exclusion criteria, resulting in a filtered list containing 61 publications. For this last group of works, a quality criterion was applied, identifying the most cited articles from 2017 to 2022, based on the H5 index.

The H5 index is a metric used to evaluate the productivity and impact of a researcher or author of scientific articles. It is based on the number of citations received by the articles published by that author. The H5 index is considered a more robust measure than the impact factor of a single article because it takes into account both the quantity of articles published and the quantity of citations received by those articles. It is widely used to compare the productivity and impact of researchers in a specific field of knowledge. Figure 1 represents this last group of papers distributed by the H5 index classification.

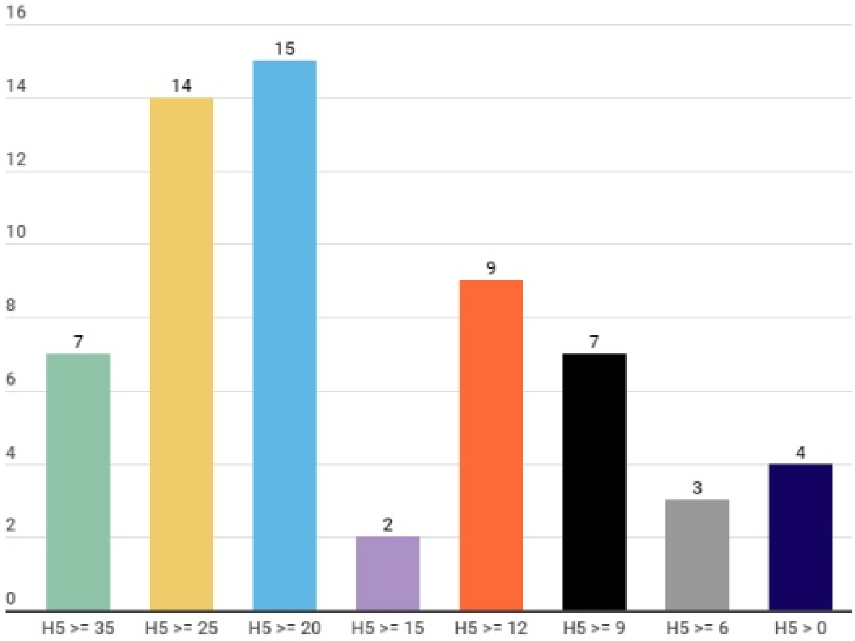


Figure 1. Quantities of articles grouped by H5 index.

From this final filtered list of articles, 11 titles were selected from the group where the H5 indexes are greater than or equal to 9, which answer the review questions raised at the beginning of this section, becoming the research base of this systematic review. These articles are indicated with an asterisk (*) in the References section. Figure 2 contains a funnel-shaped diagram to illustrate the filtering steps performed until reaching the final selection of review articles.

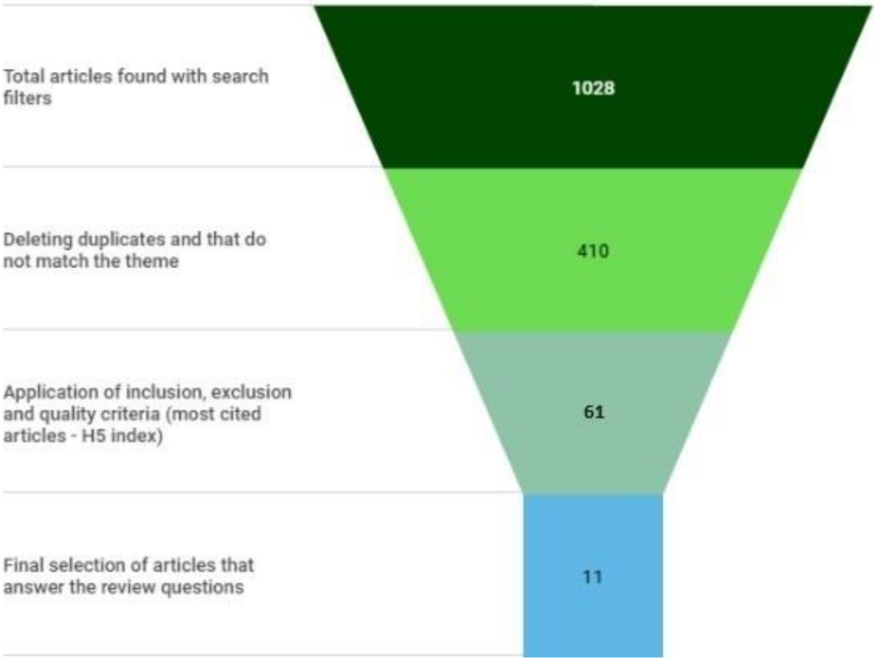

Total articles found with search filters
1028
Deleting duplicates and that do not match the theme
410
Application of inclusion, exclusion and quality criteria (most cited articles - H5 index)
61
Final selection of articles that answer the review questions
11

Figure 2. Selection flow of review articles

The criteria applied ensure that the review answers the questions raised, restricting the research to the application of NLP in documents of the legal domain with a greater focus on existing solutions in the Brazilian Judiciary, where the datasets used have textual data in Portuguese. For works applied outside Brazil, the state of the art of NLP is sought, in addition to the tools and techniques used in applications that involve documentation in the judicial context.

Results

The following section will demonstrate the results obtained in the research. This section is structured to address the questions raised in the methodology, in order to consolidate the applied systematic review process.

Highlights of high-impact publications on applied NLP in legal documentation in the public sector (Q.1)

This section is dedicated to presenting the most relevant scientific publications from the period of 2017 to 2022 related to the application of NLP in legal documents in the public sector. The selected works for this systematic review are indicated with an asterisk (*) in the References section.

The focus of this article is on works that use NLP in the general legal domain, containing citations from international works. Despite this, specific sections are reserved to address research focused on the use of NLP in the Brazilian legal context, working with language in Portuguese.

For text classification problems in legal texts NOGUTI et al. (2020) using data from the Public Prosecutor's Office of the State of Paraná (*Ministério Público do Paraná*), demonstrate that semantic approaches, such as the use of word embeddings, are more efficient than simple approaches like TF-IDF, requiring less preprocessing.

The application of deep learning models in textual classification tasks involving legal documents is introduced by NOGUTI et al. (2020). This approach utilizes large volumes of data and texts containing domain-specific words for model training.

A Long Short-Term Memory (LSTM) model trained on a corpus of legal texts from International Arbitration Awards, EU legislation documents, European Court of Human Rights decisions, and European Court of Justice decisions is presented by CLAVIÉ et al. (2021), achieving high performance in multiple legal text classification tasks.

For problems involving decision prediction, LUZ DE ARAUJO et al. (2018) generate a model to predict decisions of the Supreme Court of the United States, determining whether a decision would be affirmed or reversed. Another example of a predictive model can be found in Philippine Supreme Court decisions, which achieved an accuracy of 59% using a Random Forest classifier, while using an SVM algorithm showed a 4% reduction in accuracy, reaching 55% in the best case (VIRTUCIO et al., 2018).

In association tasks, CHAU et al. (2020) in their research generate a model to select legal responses for questions in Vietnamese legal documents, based on a corpus containing question-answer pairs.

With the aim of detecting similarity between judicial documents in the Brazilian judicial system, DE OLIVEIRA e NASCIMENTO (2022) utilize transformer-based techniques such as BERT, GPT-2, and RoBERTa, creating models specialized for Brazilian Portuguese.

Lastly, HSIEH et al. (2022) develop a model called LSTMensembler, based on LSTM, to predict the success of mediations in conciliation cases in the Labor Court. The model was trained on mediation cases from Tainan, Taiwan, and employed different classifiers, considering previous processes, resulting in more accurate predictions. XGBoost and LightGBM classifiers were used to handle process metadata, while the textual descriptions of the cases were processed with BERT and the TextCNN algorithm.

Figure 3 shows the quantity of articles distributed by the year of their publication.

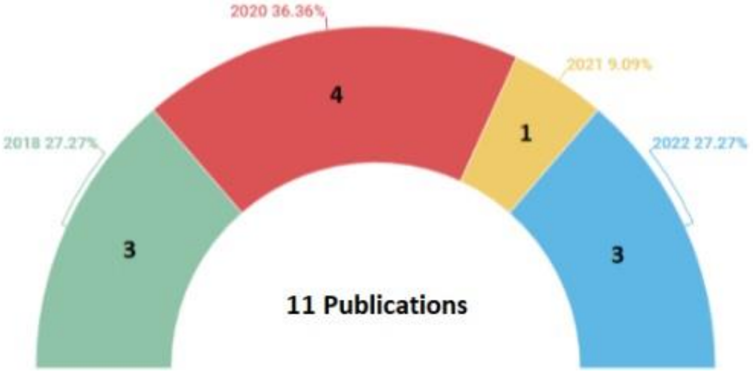


Figure 3. Number of publications of the review distributed per year.

NLP techniques and tools for dealing with legal documents (Q.2)

The methodology for building models for NLP, with a focus on the processing of legal texts, is generally composed of four steps: (1) the creation of a corpus of texts relevant to the domain in question; (2) the pre-processing of this corpus, in order to generate a dataset as output to be delivered to the NLP algorithms; (3) tagging, automatically or manually, appropriately marking the relevant terms to the approach domain; and (4) training the model that will make the intended predictions (MARANHÃO et al., 2018).

NLP techniques to group documents in the Tribunal Regional do Trabalho 5ª Região (TRT5)

Machine learning algorithms have demonstrated, through recent research, that they are powerful tools capable of solving complex problems using natural language processing tools.

The generation of word embeddings, a way of representing terms through a vector that takes into account the context, has been playing a fundamental role in carrying out analyzes on unstructured data sets, given these are present on a large scale in court documents.

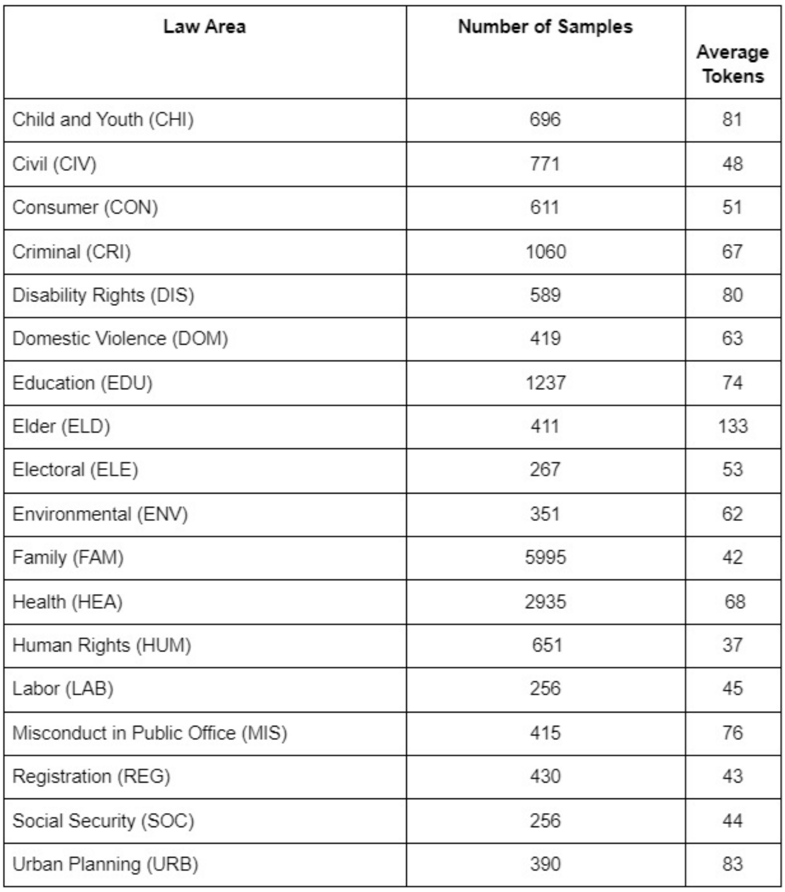
The Ordinary Appeal was selected in the research of DE OLIVEIRA e NASCIMENTO (2022) as a reference file due to its crucial role in forwarding judicial proceedings to the higher court, alongside a significant presence of pending cases awaiting judgment. This type of process represents a fundamental step in the judiciary system, directly influencing the flow and outcome of judicial cases.

In the evaluation conducted, six NLP techniques based on the transformers architecture were applied, using BERT, GPT-2, and RoBERTa, with two variations: one for Brazilian Portuguese in a general-purpose context and another for specific demands of the legal sector using an extensive set of 210,000 judicial proceedings. This strategy enabled the models to be adapted to handle the complexity and particularities of legal language, resulting in superior performance when applied in a case study involving judicial processes in Brazil (DE OLIVEIRA e NASCIMENTO, 2022).

In the research it was demonstrated that the RoBERTa pt-BR technique achieves the best results when performing the task of grouping judicial documents of the Ordinary Appeal type, surpassing BERT and GPT-2. In terms of processing power, the RoBERTa pt-BR technique stands out, proving to be very efficient when applied to models trained for longer periods and subjected to greater data loads.

4.2 LSTM with domain-specific word embeddings for document classification in the *Ministério Público do Paraná*

The PRO-MP system of the Ministério Público of the State of Paraná, in Brazil, contains 17,740 documents categorized into 18 different classes of the branch of law (NOGUTI et al., 2020). The distribution of classes in the dataset is shown in Figure 4.



Law Area	Number of Samples	Average Tokens
Child and Youth (CHI)	696	81
Civil (CIV)	771	48
Consumer (CON)	611	51
Criminal (CRI)	1060	67
Disability Rights (DIS)	589	80
Domestic Violence (DOM)	419	63
Education (EDU)	1237	74
Elder (ELD)	411	133
Electoral (ELE)	267	53
Environmental (ENV)	351	62
Family (FAM)	5995	42
Health (HEA)	2935	68
Human Rights (HUM)	651	37
Labor (LAB)	256	45
Misconduct in Public Office (MIS)	415	76
Registration (REG)	430	43
Social Security (SOC)	256	44
Urban Planning (URB)	390	83

Figure 4. Distribution of the dataset. Adapted from NOGUTI et al. (2020)

Through a base assembled with petitions from this system, containing documents from 2016 to 2019, NOGUTI et al. (2020) showed that simpler approaches, such as TF-IDF, require more extensive pre-processing to achieve good results, while semantic-oriented approaches, such as the use of word embeddings, require only minimal pre-processing to achieve good results. be applied efficiently.

According to NOGUTI et al. (2020), the model that presented the best results in the classification of short texts in the legal area was generated through a combination of Word2Vec, trained with a domain-specific corpus, applied in an Recurrent Neural Network (RNN) architecture, more specifically Long Short-Term Memory (LSTM).

In this way, this research concludes that models built using domain-specific embeddings, in this case through legal documentation in Brazilian Portuguese, are superior to models trained with embeddings based on generic documents.

LSTM to predict successful cases mediations of court lawsuits in Tainan, Taiwan

An LSTM model developed by HSIEH et al. (2022), which they named LSTMensembler, applying it in legal proceedings in order to predict whether a mediation will be successful or not. The dataset used has 5,776 cases from the mediation committee in Tainan, Taiwan, from March 2009 to January 2017. The model was generated by combining the advantages of different classifiers, in addition to considering lawsuit dependencies with previous cases, increasing the mediation prediction performance.

To perform the task of predicting mediations, HSIEH et al. (2022) used the XGBoost and LightGBM classifiers to lawsuit metadata related to the case (case sub-category, number of participants, mediator involved, among others), due to the numerical and categorical nature of their attributes. As for the textual description of the cases, the content extraction tool was used to generate embedding vectors: BERT.

XGBoost is a decision tree based algorithm that uses a Gradient Boosting structure. LightGBM, a decision tree algorithm developed by Microsoft in 2017, when compared to XGBoost, proves to be able to process data at a higher speed, using a smaller amount of memory.

In the text mining stage, in addition to BERT, the NLP tool was also used: TextCNN. TextCNN is a deep learning algorithm suitable for performing sentence classification tasks, using a focused Convolutional Neural Network (CNN) architecture for text processing.

Figure 5 is demonstrating the general steps of the LSTMensembler model generation process, with the phases for processing the lawsuit metadata on the left and the steps applied in the textual treatment of the case involved on the right.

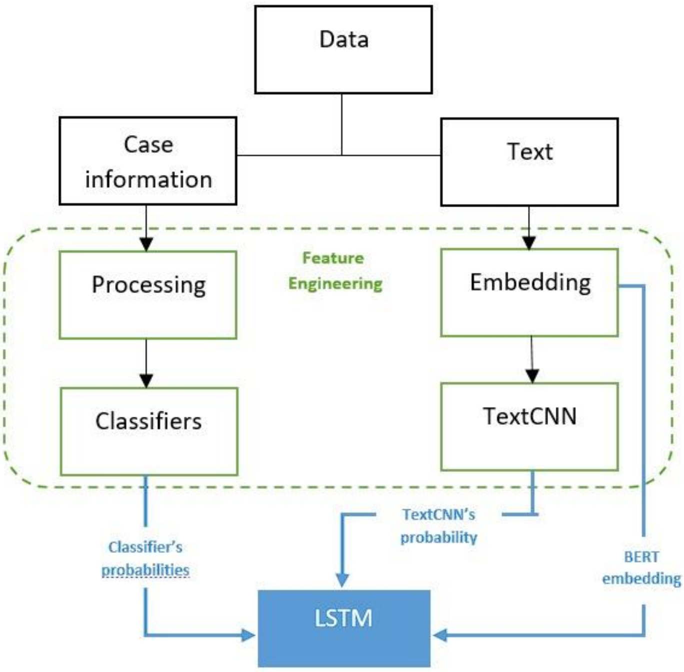


Figure 5. LSTMensembler model overview. Adapted from HSIEH et al. (2022).

Another example of documentary classification was demonstrated using the LegaLMFiT, a Long Short-Term Memory (LSTM) model, which is a type of neural network used to perform deep learning. This model was trained through a corpus formed by legal texts of the International Arbitration of Awards, documents of the EU legislation, decisions of the European Court of Human Rights and decisions of the European Court of Justice, being able to reach a high performance when performing multiple tasks of classification of legal texts, completing two of the three tasks that were given with better performance (78.8% and 83.6% of F1-Score, respectively) than the current models considered state of the art (CLAVIÉ et al., 2021).

In this way, it is concluded that the application of deep learning models in textual classification tasks involving legal documents emerges as the main solution to solve this type of problem. Neural network algorithms used in deep learning achieve optimal results when performing document classification tasks, being trained through large volumes of data and texts containing specific words of the addressed domain (WEI et al., 2018).

Performance of NLP techniques and tools in legal documents in the Brazilian context (Q.3)

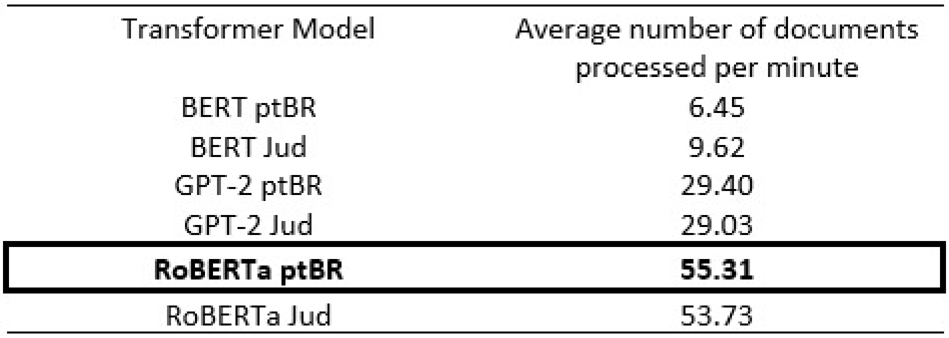
The validation metrics are used to analyze the quality of generated learning models. They provide an overview of how a model performs, as well as comparative data that allows for the selection of the best model for each case.

NLP techniques for document clustering at the *Tribunal Regional do Trabalho da 5ª Região* (TRT5)

In the research conducted by DE OLIVEIRA e NASCIMENTO (2022), three Transformer techniques were analyzed: BERT, GPT-2, and RoBERTa. Each technique generated two models, which were trained on two respective corpora: one for general-purpose Brazilian Portuguese (pt-BR) and another specialized in Brazilian labor law (Jud).

For each Transformer technique, the time to perform the processing was considered in the generation of numerical representation of approximately 210,000 judicial documents of the Ordinary Appeal type (DE OLIVEIRA e NASCIMENTO, 2022).

Figure 6 demonstrates that RoBERta pt-BR reached the objective faster, processing more documents per minute, surpassing all other models, including those specialized in labor justice corpus.



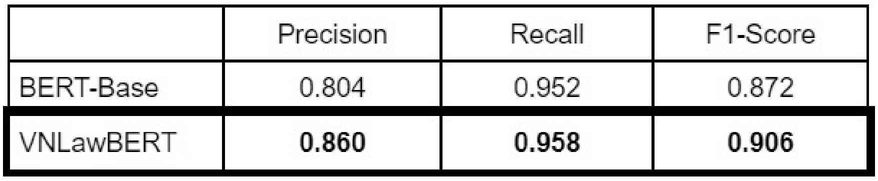
Transformer Model	Average number of documents processed per minute
BERT ptBR	6.45
BERT Jud	9.62
GPT-2 ptBR	29.40
GPT-2 Jud	29.03
RoBERTa ptBR	**55.31**
RoBERTa Jud	53.73

Figure 6. Average documents processed per minute for each model. Adapted from DE OLIVEIRA e NASCIMENTO (2022).

The use of NLP to perform textual association tasks is based mainly on models that take into account the syntactic context of words, which are capable of understanding the meaning and relationships between words and sentences in the text, such as the case of Bidirectional Encoder Representations from Transformers (BERT) language models.

A case that illustrates the association task is the model that was used to select legal answers to questions in Vietnamese legal documents. Through a corpus containing the pairs of legal questions and answers, the BERT language model reached an F1-Score of 87% for the execution of the associative task (CHAU et al., 2020). Also, an improvement in the model was registered when working with a specific corpus of the Vietnamese legal domain, reaching an F1-Score of 90.6% when performing the same task.

Figure 7 summarizes the results of the selected metrics for evaluating the BERT-base models, which include general terms of the language in question, and VNLawBERT, a BERT model trained with a specific corpus from the Vietnamese legal domain. The results demonstrate the improvement in performance when training models with a corpus from a specific domain.



	Precision	Recall	F1-Score
BERT-Base	0.804	0.952	0.872
VNLawBERT	**0.860**	**0.958**	**0.906**

Figure 7. Metrics of BERT-Base and VNLawBERT. Adapted from CHAU et al. (2020).

LSTM with domain-specific word embeddings for document classification in *Ministério Público do Paraná*

The PRO-MP system of the *Ministério Público do Paraná*, Brazil, contains 17,740 documents categorized into 18 different classes of the branch of law. Through a base assembled with petitions from this system, containing documents from 2016 to 2019, NOGUTI et al. (2020) showed that simpler approaches, such as TF-IDF, require more extensive pre-processing to achieve good results, while semantic-oriented approaches, such as the use of word embedding, require only minimal pre-processing to achieve good results.

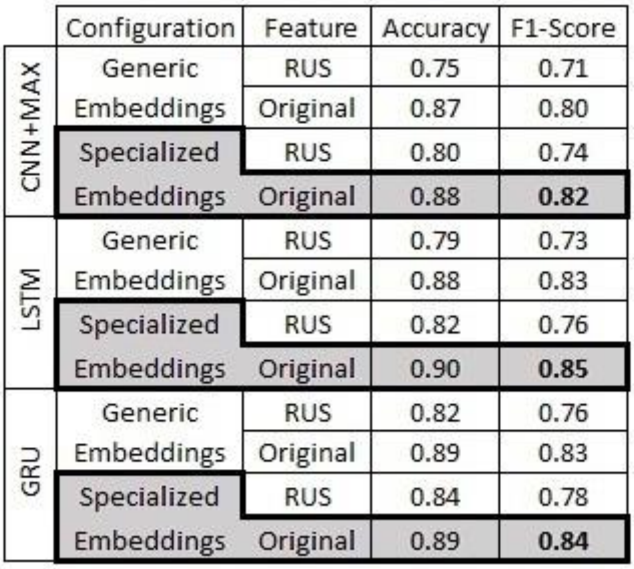
The model that presented the best results in the classification of short texts in the legal area was generated through a combination of Word2Vec, trained with a domain-specific corpus, applied in an Recurrent Neural Network (RNN) architecture, more specifically Long Short-Term Memory (LSTM).

For the comparative analysis proposed by NOGUTI et al. (2020), in addition to LSTM, GRU (Gated Recurrent Unit) and CNN+MAX models are also used. CNN+MAX refers to the combination of a convolutional layer in a convolutional neural network (CNN) with a "max pooling" layer. The CNN is designed to learn local patterns in the input data, while max pooling reduces the dimensionality of the features learned by the CNN by selecting the maximum value in each convolutional region. This approach is often used in text processing tasks, where the input is treated as a sequence of words or characters.

GRU is a type of recurrent layer in neural networks, specifically a variation of Long Short-Term Memory (LSTM) recurrent networks. GRUs are designed to handle sequence data, such as text or time series, capturing long-term dependency relationships. They contain memory units that enable the propagation of information over time and feature update and reset gates to control the flow of information.

For the task of document classification, NOGUTI et al. (2020) showed that the use of specialized embeddings and the use of the original features of the dataset, instead of the Random Under Sampler (RUS) mechanism, led to generate an LSTM model with better performance results. RUS is the technique of reducing all classes to the frequency of the smallest observed category, through a balanced random sample extracted from the classes.

The results obtained by comparing the generated neural network models were summarized in Figure 8 .



	Configuration	Feature	Accuracy	F1-Score
CNN+MAX	Generic Embeddings	RUS	0.75	0.71
		Original	0.87	0.80
	Specialized Embeddings	RUS	0.80	0.74
		Original	0.88	**0.82**
LSTM	Generic Embeddings	RUS	0.79	0.73
		Original	0.88	0.83
	Specialized Embeddings	RUS	0.82	0.76
		Original	0.90	**0.85**
GRU	Generic Embeddings	RUS	0.82	0.76
		Original	0.89	0.83
	Specialized Embeddings	RUS	0.84	0.78
		Original	0.89	**0.84**

Figure 8. Results of neural network models applied in PRO-MP. Adapted from NOGUTI et al. (2020).

Development and Evaluation of Predictive Classifiers for Labor Justice Conciliations and Sentence Reversals in *Tribunal Regional do Trabalho da 1ª Região*

In a practical context applied to Labor Justice, MENEZES NETO (2022) developed three classifiers. The first is responsible for predictive analysis in labor conciliations of the *Tribunal Regional do Trabalho da 1ª Região – Rio de Janeiro* (TRT/RJ), a topic also addressed in the research by BARROS et al. (2023). The second classifier aims to calculate the probability of reversal or modification of sentences issued by the Labor Courts, and the last one focuses on the probability of reversal or modification of decisions made by the TRT/RJ panels.

In the research, classical algorithms were employed to generate models in two fundamental steps: vectorization, which involves converting the input text into numerical representations, and processing these numerical representations with Machine Learning algorithms. Logistic Regression, Random Forest, Light Gradient Boosting Machines, and XGBoost were used, along with a final classifier named Voting Classifier, which is a weighted result of the previous four models.

In addition to the application of classical algorithms, a deep learning approach was implemented to develop the first classifier model in his research, which aims to anticipate conciliations based on the content of the initial petition. In the end, he concluded that, for this case, the bidirectional ULMFiT architecture demonstrated superior performance, achieving the best Matthews Correlation Coefficient (MCC) of 0.3006. Although the Random Forest model had a higher accuracy, reaching 75.69%, the MCC was chosen as the superior quality indicator for this classifier in the methodology.

Databases with pre-processing existing in the Brazilian legal context (Q.4)

The Brazilian legal scenario presents a favorable environment for the application of NLP in legal documentation, given the large number of processes in electronic format available for training and generation of learning models. Despite this, the Brazilian Judiciary still has few databases that have some pre-processing.

Dataset Victor of *Supremo Tribunal Federal*

In the higher courts, the Victor projects in the *Supremo Tribunal Federal* (STF) and Sócrates in the *Superior Tribunal de Justiça* (STJ) stand out. Victor, the result of a partnership between the STF and the *Universidade de Brasília* (UnB), is the AI used in the Brazilian Supreme Court to carry out document conversion activities, sorting and classification of processes, in addition to identifying topics of general repercussion of greater incidence (FILHO e JUNQUILHO, 2018).

The processes that make up the database used by the Victor robot were manually labeled by specialists in the legal area of the STF, in order to facilitate the document categorization process. The labels used to perform the document classification task are: Judgment, for first instance decisions under review; Extraordinary Appeal (RE), for appeal petitions; Extraordinary Appeal (ARE) for motions against appeals; Dispatch, for court orders; Sentence for judgments; and Others for other documents not included in the previous classes (LUZ DE ARAUJO et al., 2018).

There are also labels to classify the lawsuits, allowing the categorization of the main themes of general repercussion of each Extraordinary Appeal. In this case, 28 classes were considered to represent the most frequent themes and one class for the other themes, totaling 29 labels.

Named entity recognition in Brazilian legal documents

Named entity recognition (NER) is the process of finding, extracting and classifying entities from natural language texts, fitting them into predefined classes, such as: people, places and organizations.

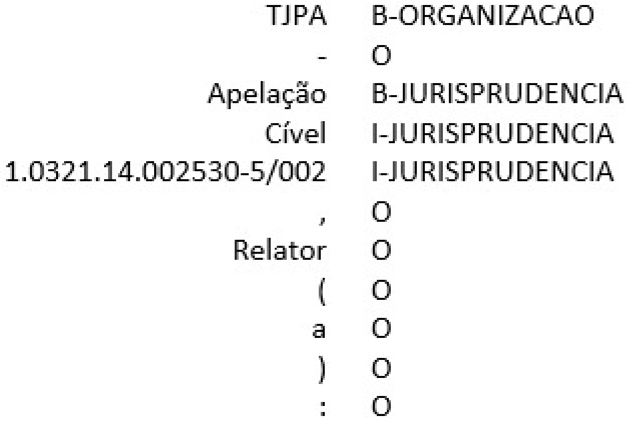
The LeNER-Br, a dataset in Brazilian Portuguese composed entirely of manually annotated court documents for the recognition of named entities, was created by (LUZ DE ARAUJO e CAMPOS, 2020). This set is made up of 66 legal documents originating from various higher and state courts, such as the Federal Supreme Court, Superior Court of Justice, Court of Justice of *Minas Gerais* and the Federal Court of Auditors. In addition to these, four more legislative documents were added, such as the Maria da Penha Law, generating a dataset containing 70 documents.

To construct the dataset, 50 documents were randomly selected for the training set, while 10 documents were chosen for each of the development and test sets. The total number of tokens in the LeNER-Br dataset is comparable to other named entity recognition corpora, such as the Paramopama and CONLL-2003 English datasets, which consist of 318,073, 310,000, and 301,418 tokens, respectively.

Using the Python library NLTK, (LUZ DE ARAUJO e CAMPOS, 2020) assembled the LeNER-BR dataset by dividing the text of the documents into a list of phrases and then performing their tokenization. An output file was generated for each document, containing one word per line and a blank line to demarcate the end of each sentence.

The WebAnno tool was used to manually annotate each document, using the following tags: “ORGANIZACAO” for organizations; “PERSON” for persons; “TEMPO” for time entities; “LOCAL” for locations; “LEGISLATION” for laws and “JURISPRUDENCE” for decisions on legal proceedings.

As a complement, the IOB labeling scheme was also used, where “B-” indicates that a tag is at the beginning of a named entity, “I-” indicates that the tag is inside a named entity, and “O” indicates that a token does not belong to any named entity (LUZ DE ARAUJO e CAMPOS, 2020). Figure 9 illustrates IOB tagging.



TJPA	B-ORGANIZACAO
-	O
Apelação	B-JURISPRUDENCIA
Cível	I-JURISPRUDENCIA
1.0321.14.002530-5/002	I-JURISPRUDENCIA
,	O
Relator	O
(	O
a	O
)	O
:	O

Figure 9. IOB document tagging. Adapted from LUZ DE ARAUJO e CAMPOS (2020).

In the end, (LUZ DE ARAUJO; CAMPOS, 2020) generated an LSTM-CRF model, composed of a Bidirectional LSTM architecture followed by a Conditional Random Field (CRF) layer, which was trained using the LeNER-BR dataset. This model achieves state-of-the-art performance on the English CoNLL-2003 test set (with an F1-score of 90.94%). It offers readily available open-source implementations that have been specifically adapted to cater to the requirements of this project. The utilization of 300-dimensional GloVe word embeddings, pretrained on a diverse corpus encompassing both Brazilian Portuguese and European Portuguese texts, adds further sophistication. Notably, the model achieves remarkable F1 scores of 97.04% and 88.82% for identifying Legislation and Legal Case entities, respectively, providing substantial evidence for the applicability of the proposed dataset within the legal domain.

Another case of NER was carried out for STF legal decisions, where law students acted by generating two levels of annotated nested legal entities: 4 legal entities at a coarser level and 24 entities nested at a more refined level. The final result of the work was a corpus containing 594 annotated decisions of the STF, which is considered the most extensive corpus in Portuguese dedicated to the recognition of named entities in the legal field, serving as a tool to support research and improvements in this context (CORREIA et al., 2022).

The dataset created by (CORREIA et al., 2022) allowed the generation of models capable of accurately extracting, from judicial decisions, central entities for reasoning and legal argumentation.

Conclusion

The analysis of the literature obtained indicates the growing trend of the study around natural language processing applied in dealing with documents in the legal sector, especially in the last five years, showing the importance of using this technology to improve the performance of the activities performed in the courts.

The state of the art of NLP applied to legal documentation demonstrates the use of deep learning models to perform textual classification tasks. For decision prediction, the literature advocates the application of decision tree models. For tasks that involve textual association rules, BERT language models are used.

In the approach of NLP tools and techniques for legal document treatment, an LSTM network working together with domain-specific word embeddings can generate models that present a better performance. In the treatment of numerical and categorical data, the tools XGBoost and LightGBM are highlighted, while for textual mining tasks, BERT and TextCNN appear acting in cases of procedural mediation. Finally, the transformer, called RoBERta pt-BR, presents a linguistic context focused on Brazilian Portuguese, efficiently processing documents in the TRT 5ª Região.

There are few databases in the context of the Brazilian Judiciary that have some prior processing that helps the use of NLP, highlighting as a pre-processed database the Victor dataset of the STF. In addition, the use of word embeddings in Brazilian Portuguese is still under development, highlighting the need for greater investments to generate language models in Portuguese for the Brazilian legal field.

In a country like Brazil, support tools based on NLP, applied in more bureaucratic activities, represent a great differential in the performance of the work performed in the Brazilian courts, mainly taking into account that the country already has a huge number of processes in electronic media, which creates a favorable environment for the application of this type of technology, in addition to providing a real maturation of the country's legal system.

References

BARROS, F. M. C.; SILVA, C. D.; SILVA, I. R. M.; MARTINS, V. S.; ARAÚJO, A. J. S. Machine Learning Algorithms Applied on Classification of Processes for Conciliation on Brazilian Labour Judiciary. In: ENCONTRO NACIONAL DE INTELIGÊNCIA ARTIFICIAL E COMPUTACIONAL (ENIAC), 20., 2023, Belo Horizonte/MG. **Anais [...]**. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 389-402. ISSN 2763-9061. doi: https://doi.org/10.5753/eniac.2023.234189.

CHAU, C.-N.; NGUYEN, T.-S.; NGUYEN, L.-M. VNLawBERT: A Vietnamese Legal Answer Selection Approach Using BERT Language Model. 2020 **7th NAFOSTED** Conference on Information and Computer Science (NICS), Ho Chi Minh City, Vietnam, 2020, p. 298-301, doi: 10.1109/NICS51282.2020.9335906.

CLAVIÉ, B.; GHEEWALA, A.; BRITON, P.; ALPHONSUS, M.; LAABIYAD, R.; PICCOLI, F. LegaLMFiT: Efficient Short Legal Text Classification with LSTM Language Model Pre-Training. **arXiv preprint** arXiv:2109.00993, 2001. doi: https://doi.org/10.48550/arXiv.2109.00993

CORREIA, F. A.; ALMEIDA, A. A. A.; NUNES, J. L.; SANTOS, K. G.; HARTMANN, I. A.; SILVA, F. A.; LOPES, H. Fine-grained legal entity annotation: A case study on the Brazilian Supreme Court. **Information Processing & Management**, v. 59, n. 1, 2022, p. 102794. ISSN 0306-4573, doi: https://doi.org/10.1016/j.ipm.2021.102794.

DE OLIVEIRA, R. S.; NASCIMENTO, E. G. S. Brazilian Court Documents Clustered by Similarity Together Using Natural Language Processing Approaches with Transformers. **arXiv preprint** arXiv:2204.07182, 2022. doi: 10.48550/arXiv.2204.07182

HSIEH, H.-P.; JIANG, J.; YANG, T.-H.; Hu, R.; WU, C.-L. Predicting the success of mediation requests using case properties and textual information for reducing the burden on the court. **ACM Journals**, v. 2, n. 4, 2022, p. 1–18. Nova York, NY, EUA. doi: https://doi.org/10.1145/3469233*

LUZ DE ARAÚJO, P. H.; CAMPOS, T. Topic Modelling Brazilian Supreme Court Lawsuits. **33rd International Conference on Legal Knowledge and Information Systems** (JURIX 2020), v. 334, 2020, p. 113–122. Praga, República Tcheca. doi: http://dx.doi.org/10.3233/FAIA200855.

LUZ DE ARAÚJO, P. H.; CAMPOS, T.; BRAZ, F. A.; SILVA, N. C. VICTOR: a Dataset for Brazilian Legal Documents Classification. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, 2020, p. 1449–1458, Marseille, França. European Language Resources Association.

LUZ DE ARAÚJO, P. H.; CAMPOS, T.; OLIVEIRA, R. R. R.; STAUFFER, M. COUTO, S. BERMEJO, P. LeNER-Br: A Dataset for Named Entity Recognition in Brazilian Legal Text. In: Villavicencio, A., et al. Computational Processing of the Portuguese Language. **PROPOR** 2018. Lecture Notes in Computer Science(), vol 11122. Springer, Cham. doi: https://doi.org/10.1007/978-3-319-99722-3_32.

MAIA FILHO, M. S.; JUNQUILHO, T. A. Projeto Victor: Perspectivas de Aplicação da Inteligência Artificial ao Direito. **Revista de Direitos e Garantias Fundamentais**, v. 19, n. 3, p. 218–237, 2018. doi: 10.18759/rdgf.v19i3.1587. Disponível em: https://sisbib.emnuvens.com.br/direitosegarantias/article/view/1587. Acesso em: 20 ago. 2024.

MARANHÃO, J. S. de A.; FLORÊNCIO, J. A.; ALMADA, M. Inteligência artificial aplicada ao direito e o direito da inteligência artificial. **Suprema - Revista de Estudos Constitucionais**, Distrito Federal, Brasil, v. 1, n. 1, p. 154–180, 2021. doi: 10.53798/suprema.2021.v1.n1.a20. Disponível em: https://suprema.stf.jus.br/index.php/suprema/article/view/20. Acesso em: 20 ago. 2024.

MARTINS, V. S.; SILVA, C. D.. Text Classification in Law Area: a Systematic Review. In: SYMPOSIUM ON KNOWLEDGE DISCOVERY, MINING AND LEARNING (KDMILE), 9. , 2021, Rio de Janeiro. **Anais [...].** Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 33-40. ISSN 2763-8944. doi: https://doi.org/10.5753/kdmile.2021.17458.

MENEZES NETO, E. J. de. **Inteligência Artificial e Eficiência do Judiciário**: Uso de Análise Preditiva em Conciliações, Sentenças e Acórdãos no Tribunal Regional do Trabalho da 1ª Região. Relatório Final. Natal, Rio Grande do Norte: UFRN - Universidade Federal do Rio Grande do Norte, 2022.

NOGUTI, M. Y.; VELLASQUES, E.; OLIVEIRA, L. S. Legal Document Classification: An Application to Law Area Prediction of Petitions to Public Prosecution Service, 2020 **International Joint Conference on Neural Networks** (IJCNN), Glasgow, UK, 2020, p. 1-8, doi: 10.1109/IJCNN48605.2020.9207211.*

PINTO, L. A. S. .; ARAÚJO, I. A. F. .; SANTANA JÚNIOR, O. V. de . Transformando o aprendizado: uma proposta de um bot educacional para auxiliar o professor - RN. **Revista de Casos e Consultoria**, v. 15, n. 1, p. e33870, 2024.

SPOLAOR, N.; LEE, H. D.; TAKAKI, W. S. R.; ENSINA, L. A.; COY, C. S. R.; WU, F. C. A systematic review on content-based video retrieval. **Engineering Applications of Artificial Intelligence**, v. 90, 2020, p. 103557. ISSN 0952-1976, doi:https://doi.org/10.1016/j.engappai.2020.103557.

VASCONCELOS, R. C.; SOUZA, M. A.; PIMENTEL, M. d. G. C. Justiça 4.0: um Panorama das Tecnologias e Soluções Aplicadas ao Poder Judiciário Brasileiro. **SBC**, v. 11, n. 3, 2020, p. 251–265.

VIRTUCIO, M. B. L.; ABONITA, J. K. C.; AVIÑANTE, R.; ABOROT, J. Predicting Decisions of the Philippine Supreme Court Using Natural Language Processing and Machine Learning, 2018 IEEE **42nd Annual Computer Software and Applications Conference** (COMPSAC), Tokyo, Japão, 2018, p. 130-135, doi: 10.1109/COMPSAC.2018.10348.

WEI, F.; QIN, H.; YE, S.; ZHAO, H. Empirical Study of Deep Learning for Text Classification in Legal Document Review, 2018 IEEE **International Conference on Big Data** (Big Data), Seattle, WA, EUA, 2018, p. 3317-3320, doi: 10.1109/BigData.2018.8622157.