ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 09, Issue, 07, pp. 28642-28646, July, 2019

RESEARCH ARTICLE OPEN ACCESS

MINERJUS: SOLUTION TO THE PROCESSUAL CLASSIFICATION WITH USE OF ARTIFICIAL INTELLIGENCE

[1,*]Rogério Nogueira de Sousa, [1]Leandro O. Ferreira, [2]Jefferson David Asevedo Ramos and [1]David N. Prata

[1]Programa de Pós-Graduaçãoem Modelagem Computacional de Sistemas, UFT, TO, Brazil
[2]Programa de Pós-Graduaçãoem Prestação Jurisdicional e Direitos Humanos, UFT, TO, Brazil

ARTICLE INFO

Article History:
Received 22nd April, 2019
Received in revised form
19th May, 2019
Accepted 29th June, 2019
Published online 28th July, 2019

Key Words:
Artificial Intelligence,
Machine Learning,
Jurisdictional Provision.

ABSTRACT

The electronic judicial process is a reality in Brazil, where 70% of the new cases in all judicial power are virtual, making proper use of this reality and improving it is paramount to give a flow to the demand of approximately 25 million new cases per year. This project proposes an improvement in the electronic processes, through the use of Artificial Intelligence, to assist legal operators responsible for registering the initial petition document (creation of the process), as well as those responsible for the analysis, through automatic and assertive suggestion to the subject of the process, printing greater agility of procedure and quality in the information contained in the Brazilian judicial records.

Citation: Rogério Nogueira de Sousa, Leandro O. Ferreira, Jefferson David Asevedo Ramos and David N.Prata. 2019. "Minerjus: solution to the processual classification with use of artificial intelligence", *International Journal of Development Research*, 09, (07), 28642-28646.

INTRODUCTION

In 2016, Judiciary Power spent R $ 2.248.734.431 with Information Technology (IT) and a workforce of 442.345 employees, divided among magistrates, servers and auxiliaries, to face the 79.7 million pending cases in the year in the Brazilian justice system (CONSELHO NACIONAL DE JUSTIÇA, 2017). In the year 2017, IT spending decreased by R $ 2.207.995675 and the number of processes in process passed the 80 million, with the same amount of components of 2016 (CONSELHO NACIONAL DE JUSTIÇA, 2017). Faced with this scenario with significant numbers, there is a worrying situation of increased judicial demand and scarce resources. The search for increasingly efficient solutions, which can maximize the employees' work capacity, as well as reduce process costs, become imperative for jurisdictional provision in Brazil. Information Technology is then accessed as one of the ways to speed up judicial activities, with less time spent by the professionals involved and, consequently, with greater resource savings (FELIPE; PERROTA, 2018). The duty of efficiency implies the requirement that the Public Administration incorporate the technological advances in its activity (JUSTEN FILHO, 2016). The Brazilian judiciary is fully aware of the importance of IT for jurisdictional provision, so much that it allocates around 25% of its budget (excluding personnel expenses) to IT (CONSELHO NACIONAL DE JUSTIÇA, 2017). Among the system-oriented technological solutions of justice, we highlight the use of electronic judicial processes, since 70% of the new lawsuits are electronic. Some Brazilian courts stand out for having 100% of electronic processes in the two degrees of jurisdiction, among them (CONSELHO NACIONAL DE JUSTIÇA, 2017), the Court of Justice of the Tocantins (TJTO), which at the forefront of the electronic judicial process, implemented the E-Proc/TJTO in 2011. Still in 2011, 100% of the new cases became virtual. After 4 years, all the processes in process were digitized, in 2015, the first court to have all the judicial processes in digital format(TJTO, 2015). The digitization of legal data is a mega-trend, transforming workflows and business models. The volume of data used in legal counseling has increased exponentially (VEITH *et al*., 2016), generating greater demand for selection, analysis and interpretation of an unprecedented

Corresponding author: **Rogério Nogueira de Sousa,**
Programa de Pós-Graduaçãoem Modelagem Computacional de Sistemas, UFT, TO, Brazil.

amount of data. In contrast, such virtualization facilitates the automation process, enabling productivity growth while still reducing costs; increasing the quality and minimizing the downtime of the operators of the right. We are experiencing a new age of automation, in which robots and computers can not only perform a series of routine physical work activities more efficiently and cheaply than humans. But, they are also increasingly capable of performing activities that include cognitive abilities (Mckinsey global institute, 2017). With recent developments in robotics, artificial intelligence, and machine learning, technologies not only do things we thought only humans could do, but can also do things more and more at superhuman levels of performance (Mckinsey global institute, 2017). The initial petition, as the name says, is the first act for the formation of the judicial process (TJDFT, 2014).The process happens to exist electronically when the registration of the same in electronic process systems occurs. In the face of this fact, a group of collaborators formed by magistrates of the Court of Justice of the Tocantins pointed out that, not rarely, the judicial offices carry out the reclassification of the process, generating rework, or simply the process classified Wrongly goes into the system.

The National Council of Justice (CNJ), with the objective of improving the administration of justice and judicial performance, defined interoperability standards to be used in the Judiciary. Among them, the standardization of the basic tables of procedural classification, movement, procedural phases, subjects and parts(TJRR, 2008). Therefore, ensuring greater reliability of the classification of the process at the time of registration of the initial petition is vital. Not only for the promotion of reliable statistical data, but also for future integrations between information systems. The implementation of an automated system that assists in the process classification process based on the information contained in the initial petition has the potential to directly impact on the efficiency of the judicial employees responsible for the preliminary analysis of the initial. And also of the lawyers who register the petition. Producing the benefits of automation. For this, this system uses records contained in the tables of procedural classifications managed by the CNJ. In the face of explicit demand, this project aims to present an automation tool for process classification using the Learning Machine, which is an extremely important follow-up in Artificial Intelligence(JR, 2016). Machine learning techniques will be used to significantly reduce the number of documents that today require manual overhaul(JR, 2016). For example, the Learning Machines (ML) are able to predict the classification of documents from documents previously classified correctly (ZAKI; MEIRA, JR, 2014) in their training base. Initially, the tool will focus on the prediction of the subject of the judicial process from the extraction of data from the initial petition. Through the preprocessing of the petition, which initially consists of removing the textual content of the digital documents, will usually be in PDF format, which composes petition, later are applied techniques of Natural Language Processing (NLP) that will convert the texts into vectors of relevant terms for classification intended and understood by the computer.

For the formation of the predictive model (Learning Machine) will be used initial petitions of processes that process in the region of Augustinópolis-TO. Where the researcher and Chief Judge of the Special Civil Court, Dr. Jefferson David Asevedo Ramos with his team, carried out the work of screening and validation of the subjects of a group of processes that process in that specialized branch, acting as supervisor of the contents to be submitted to the machine. The selected processes will be divided into two groups, being called training corpus and test corpus. The first group will be used to teach the machine learning patterns of data related to the content of the documents of a given subject, generating an analytical model. The second will be used to validate the learning process by comparing the assertiveness and performance of the technological solution with the process currently used, which consist of assigning subjects manually after individual process analysis. As a result, we expect to have a considerable reduction in the time for filing initial petitions, increasing assertiveness in relation to the subject of the petition and mitigating rework performed by employees of judicial registries, thus promoting greater procedural speed and confidence in the data attributed to judicial processes.

METHODOLOGY

A solution was developed, capable of extracting information from documents in PDF format, processing the text from these documents, and subjecting them to a previously trained machine learning algorithm capable of suggesting the subject of the information contained in the document. Concomitant to the creation of the software, a training corpus was created, consisting of documents (initial petitions) carefully classified in relation to the subject. This mass of data was responsible for presenting to the learning machine the contexts and subjects, which will be used to classify the initial petitions submitted to it. The processes from which the initial petitions were extracted based on the subject, are in the electronic system of judicial processes of the Court of Justice of Tocantins (E-Proc), these processes are processed in the special civil court of the region of Augustinópolis, which are in phase of knowledge. Subsequently, the prediction capacity of the solution was evaluated, quantitatively calculating accuracy and precision indexes. The artificial intelligence techniques employed in the development of the solution, as well as the composition corpus training approaches can be adapted, aiming at the improvement of the tool, improving the prediction capacity and computational performance.

Materials

Part of this work will be the selection of the judicial process in processing and their respective initial petitions (figure 1), which have undergone rigorous analysis, one by one, and when necessary, corrected the subject that was assigned to it in the act of the initial registration in the electronic process system. This action was carried out by a team of Magistrates and judicial analysts of the Court of Tocantins with experience in jurisdictional rendering, thus seeking to minimize possible misinterpretations regarding the subject, which would later be mistakenly passed on to the learning machine, negatively reflecting on the tool prediction capability. The internal legislation of the TJTO states that only files in PDF (Portable Document Format) format should be used for texts(TJTO, 2011), so all the documents that we will use will be in PDF. So it is imperative for this project that the system be able to extract texts from this document format. All documents must pass through a text extractor, allowing the manipulation of the content by the computer.

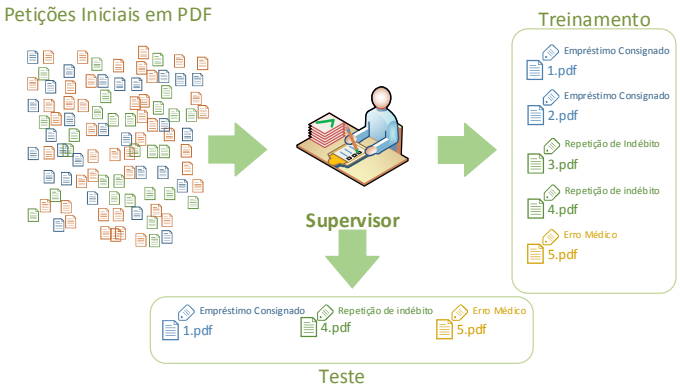

Figure 1. Formation of the corpus

Tools: To extract the texts contained in the initial petitions that are in PDF format, the system of extracting metadata and texts, called apache Tika will be used. Written in Java and maintained by the Apache Software Foundation, it is free software capable of analyzing different types of files and returning extracted information in plain text, which will be subjected to a set of natural language processing (NLP) techniques. The language used in the development of the NLP techniques that make up the solution is Python, in version 3.7. Because it is a high-level, object-oriented language, Python can be used across platforms because it is interpreted(PYTHON.ORG, 2019). This language has been shown to be a good choice for the speed of development and maintenance and has been established as one of the most popular languages of scientific computing(PEDREGOSA *et al*., 2011). Python still has a community that prides itself on its development culture, tradition of well-defined APIs, and capillarity of machine learning methods (Szymański; kajdanowicz, 2019).

Some files written in Python can contain definitions, functions, and classes, which are initialized when invoked within an application, are usually object abstractions, these are called modules. It will be mentioned some possible module that will compose the proposed solution:

Pandas (Data Structures for Statistical Computing in Python): high-performance module, used for data manipulation through matrix-based structures, which implements various statistical models(MCKINNEY, 2010).

Scikit-learn: set of tools used to implement machine learning, supervised and unsupervised, with an easy-to-use interface and strictly integrated with the Python language(Pedregosa *et al*., 2011).

NLTK(Natural Language Toolkit): platform built to work with human language, using techniques of Natural Language Processing(BIRD; KLEIN; LOPER, 2009).

Methods

The text to be classified will be extracted from the initial requests that are in PDF format, through the service of extracting texts of files in PDF format, known as Apache Tika. The documents that make up the training corpus will also have their contents extracted and processed (Figure 2). To reduce the complexity of the texts, all tokens will be converted to lowercase and the special characters of the texts as well as the accents, numbers and punctuations will be removed. Another way to reduce the vocabulary, consequently the computational complexity is the withdrawal of stop words, which are words that do not have relevance for the classification of the text (Lane; howard; hapke, 2017). In order to give more generality to the terms, a processing is carried out in each term, where morphologically complex chains are identified, decomposed into radicals and affixes, the affixes are discarded and the term becomes only the radical, known as stemming (Lane; Howard; Hapke, 2017). When adopting the stemming technique for token formation, removing the suffix and prefix, we have a more generic term, for example the words 'book', 'booklet', 'books' and 'booklets', all have similar or close meanings and in common the string of characters 'book', being the base element for the meaning. One can then substitute the four words for the radical 'book' that there is no considerable loss of meaning. Even though 'free' is not an existing word, it does not matter why the goal is to marry words into queries and documents and not show them to the user (Coppin, 2017). The term occurrence matrix undergoes a correction and normalization with the application of TF-IDF (Term Frequency-Inverse Document Frequency), which calculates how relevant is the term for classification in relation to the total set of documents, by calculating the quantity of times a term appears in the document, balanced by the amount of documents it appears(JONES, 1972).

The prediction model is generated by a machine learning algorithm, in this work we will use the algorithm Support Vector Machine (SVM), since it offers important advantages, which makes it attractive for classifying texts, where in its implementation is used techniques of linear and discriminatory classification, thus being able to deal with problems of high dimensionality and sparse data(JOACHIMS, 1998). These characteristics are important to work with the classification of large collections of documents(D'ORAZIO *et al*., 2014).

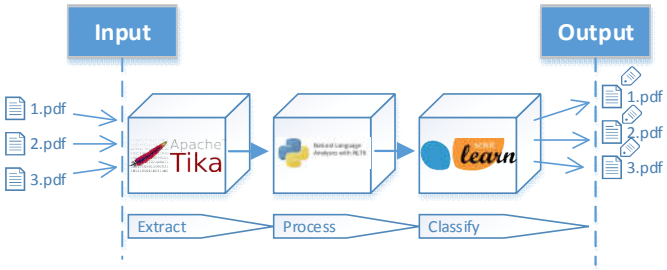

Figure 2. Text classify process

The SVM algorithm is of the supervised type, that is to say that the algorithm learns through a data set of properly labeled examples (HAN; PEI; KAMBER, 2011), in this case it is the training corpus. The new documents that must be classified, will be submitted to the processes of extraction of the text, processing and classification.

RESULTS

To evaluate the solution, we will use a test corpus composed of properly labeled data, which were not previously submitted to the predictor model, these data helped to verify the accuracy and accuracy of the model. By analyzing in this way one can make sure there is no over fit. Accuracy, which is the simplest metric that can be used to evaluate a classifier, measures the percentage of correctly classified data (Bird *et al.,* 2009),

calculated by summing the correct classifications divided by the number of documents in the test corpus(ZAKI; MEIRA, JR, 2014). According to the type of problems other metrics can be used, for example in problems that present a very unbalanced test base, such as the task of classifying whether a document is relevant or not at the time of a search, logically the amount of irrelevant documents (BIRD; KLEIN; LOPER, 2009), so a model that asserts that no document has relevance will present an accuracy of close to 100%. The confusion matrix (Figure 1) is a technique used to formulate other evaluation metrics, it is an *N*x*N* matrix, where *N* is the number of classes contained in the test set, each cell [*i, j*], indicates how many classes *i* were predicted while the right would be *j*. Thus the diagonal of this matrix informs how many hits have occurred and what is not diagonal are the errors(BIRD; KLEIN; LOPER, 2009). In other words the lines represent the classes informs by the classifier, and the columns are reference classes, coming from the test set.

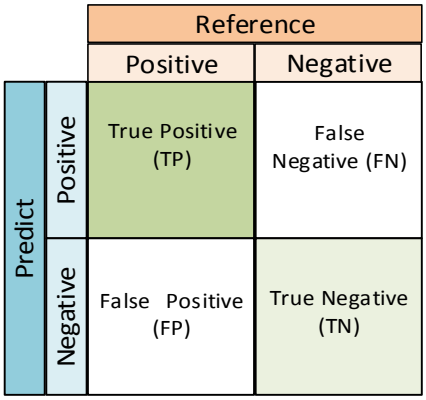

Figure 3. Confusion matrix

True positive (TP) is number of correct positive ratings.
True negatives (VN) are total negative ratings correct.

False positive (FP) is given by the portion of incorrect positive rating. False negatives (FN) are formed by the number of incorrect negative classifications. When one wants to know which proportion of positive classification is correct, we use the metric precision, denoted by *P* and given (1):

$$P = \frac{TP}{(TP+FP)} \quad (1)$$

Another metric is the recall which measures the ratio between the positive potential classification and how many were classified correctly calculated and denoted by *R* (2):

$$R = \frac{TP}{(TP+FN)} \quad (2)$$

One way to combine precision and recall to provide a single index is by using the harmonic mean between precision and recall(INDURKHYA; DAMERAU, 2010), this index is known as *F1 score* (3).

$$F1 = \frac{2 \cdot P \cdot R}{(P+R)} \quad (3)$$

In order to test the solution, we randomly selected 78 initial petitions (pdf) of cases that are processed in the civil court of Augustinópolis -TO of the Court of Justice of Tocantins and submitted to the system in order to predict the issues. The table 1 presents precision, recall and *F1-score* by subject:

Table 1. Detail precision, recall and F1-scores by subjects

Subject	Precision	Recall	F1-scores
Acidente de Trânsito	1.00	1.00	1.00
AssinaturaBásica Mensal	1.00	1.00	1.00
Cancelamento de Voo	1.00	1.00	1.00
Cartão de Crédito	1.00	1.00	1.00
Cheque	1.00	1.00	1.00
Compromisso	0.00	0.00	0.00
Dever de informação	1.00	1.00	1.00
Direito de imagem	0.00	0.00	0.00
Direito de vizinhança	1.00	1.00	1.00
Empreitada	0.00	0.00	0.00
EmpréstimoConseguindo	1.00	1.00	1.00
Fornecimento de energiaelétrica	1.00	1.00	1.00
Fornecimento de água	1.00	1.00	1.00
Inclusãoindevidacadastro de inadimplentes	1.00	0.95	0.97
Locação de imóveis	0.00	0.00	0.00
Nota Promissória	0.00	0.00	0.00
Obrigação de fazer/nãofazer	1.00	1.00	1.00
Oferta e publicidade	1.00	1.00	1.00
Perdas e Danos	1.00	1.00	1.00
Rescisão de contrato e devolução do dinheiro	1.00	1.00	1.00
Substituição e produto	1.00	1.00	1.00
Tarifas	1.00	1.00	1.00

The solution presented an accuracy of 93.58%, when tested with an unknown set of documents by the learning machine and an accuracy of 72.72% in the same set.

Conclusion

This work presents the development of a solution that aims to guarantee greater reliability to the classification of legal proceedings in the act of registering the initial petition with respect to the subject, through the use of natural language processing techniques for extraction of information contained in documents in format pdf, a learning machine was built capable of automatically predicting the subject of the process.
Currently there is no tool to assist legal operators in the task of classifying the process, which leads to several inconsistencies in the process classification data, such as the Bahia Court of Justice (TJBA), which analyzed 404.3 thousand cases and identified that 56% presented errors in the registry of the initial petition, being that of the analyzed ones, 176,598 present deficiencies in the classification with respect to the subject, representing 78% of the errors found (TJBA, 2017). When it obtained an accuracy of 93.58%, the solution proves its feasibility and ability to support the legal operators, since in the case of the TJBA the classification performed by humans without technological support reaches a success rate of 56% with respect to the subject. An intrinsic feature of learning machines is the ability to learn, soon to improve current levels of accuracy and precision of the software should "teach" it continuously. In the specific case of precision, it is essential to improve the examples and / or expand the training base prioritizing those subjects who had low precision. At the end of the work, we can conclude that when we opt for open source technologies we guarantee a flexible and easy maintenance solution. Duly effective, capable of promoting agility in the procedural process mitigating rework and improving the quality of information contained in court proceedings. Providing a significant contribution to the jurisdictional provision.

REFERENCES

Bird, S.; Klein, E.; Loper, E. Natural Language Processing with Python. 1. ed. Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472: Julie Steele, 2009.

Conselho Nacional De Justiça. RelatórioJustiçaemNúmeros 2017, ano-base 2016. Disponívelem: <http://www.cnj.jus.br/files/conteudo/arquivo/2017/12/b60a659e5d5cb79337945c1dd137496c.pdf>.

Coppin, B. Inteligência Artificial. Rio de Janeiro: LTC, 2017.

D'orazio, V. *et al*. Separating the Wheat from the Chaff: Applications of Automated Document Classification Using Support Vector Machines. Political Analysis, v. 22, n. 2, p. 224–242, 1 jan. 2014.

Felipe, B. F. da C.; Perrota, R. P. C. Inteligência Artificial no Direito – umarealidade a serdesbravada. Revista de Direito, Governança e Novas Tecnologias, v. 4, n. 1, p. 1–16, 21 ago. 2018.

Indurkhya, N.; Damerau, F. J. Handbook of Natural Language Processing. 2nd. ed. [s.l.] Chapman & Hall/CRC, 2010.

Joachims, T. Text categorization with Support Vector Machines: Learning with many relevant features. In: NÉDELLEC, C.; ROUVEIROL, C. (Ed.). Machine Learning: ECML-98. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998. 1398p. 137–142.

Jones, K. S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, v. 28, p. 11–21, 1972.

Jr, D. W. L. What We Know and Need to Know About Legal Startups. SOUTH CAROLINA LAW REVIEW, v. 67, p. 31, 2016.

Justen Filho, M. Curso de Direito Administrativo. n. 12ª, p. 1861, 2016.

Lane, H.; Howard, C.; Hapke, H. M. Natural Language Processing in Action. 3. ed. [s.l.] Manning Publications Co., 2017.

Mckinney, W. Data Structures for Statistical Computing in Python. p. 6, 2010.

Mckinsey Global Institute. A Future That Works: Automation, Employment and Productivity. Disponívelem: <https://www.mckinsey.com/~/media/mckinsey/featured%20insights/Digital%20Disruption/Harnessing%20automation%20for%20a%20future%20that%20works/MGI-A-future-that-works-Executive-summary.ashx>. Acessoem: 30 out. 2018.

Pedregosa, F. *et al*. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, n. Oct, p. 2825–2830, 2011.

Python.ORG. The Python Tutorial - Documentação do Python 3.7.1. Disponívelem: <https://docs.python.org/3/tutorial/index.html>. Acessoem: 18 nov. 2018.

Szymański, P.; Kajdanowicz, T. scikit-multilearn: A Python library for Multi-Label Classification. Journal of Machine Learning Research, v. 20, n. 6, p. 1–22, 2019.

TJBA. Projeto Cadastrar Melhoravança para sanear 1 milhão de processosaté o final do ano. Disponívelem: <http://www5.tjba.jus.br/portal/projeto-cadastrar-melhor-avanca-para-sanear-1-milhao-de-processos-ate-o-final-do-ano/>. Acessoem: 10 jul. 2019.

Tjdft. Petição Inicial - ondetudocomeça — TJDFT - Tribunal de Justiça do Distrito Federal e dos Territórios. Disponívelem: <http://www.tjdft.jus.br/institucional/imprensa/direito-facil-1/peticao-inicial-onde-tudo-comeca>. Acessoem: 30 out. 2018.

Tjrr. Conhecer a Tabela Processual Unificada do CNJ. Disponívelem: <http://www.tjrr.jus.br/sistemas/php/metas_cnj/index.php?option=com_content&view=article&id=55&Itemid=66>. Acessoem: 29 out. 2018.

Tjto. Instrução Normativa Nº 5, DE 24 DE Outubro DE 2011.Disponívelem: <http://wwa.tjto.jus.br/ elegis/Home/Imprimir/423>. Acessoem: 9 nov. 2018.

Tjto. Justiça 100%digital: Processoeletrônicotrazuma nova realidade para o Judiciário do Tocantins. Disponívelem: <http://www.tjto.jus.br/index.php/noticias/3698-justica-100-digital-processo-eletronico-traz-uma-nova-realidade-para-o-judiciario-do-tocantins>. Acessoem: 17 maio. 2019.

Veith, C. *et al*. How Legal Technology Will Change the Business of Law. n. Report, 2016.

Zaki, M. J.; Meira, JR, W. Data Mining and Analysis: Fundamental Concepts and Algorithms. 1. ed. [s.l.] Cambridge University Press, 2014.

****