Document type classification for Brazil's supreme court using a Convolutional Neural Network

N. Correia da Silva, F. A. Braz, D.B. Gusmão,
F.B. Chaves, D.B. Mendes, D.A. Bezerra, G.G. Ziegler,
L.H. Horinouchi, M.H.P. Ferreira,
P.H.G. Inazawam, V.H.D. Coelho, A.B.S. Guedes
Faculdade de Engenharias do Gama, FGA
University of Brasília, UnB
Gama, Brazil

T.E. de Campos
Departamento de Ciência da Computação
University of Brasília, Brasília, Brazil
{niltoncs,fabraz,teodecampos}@unb.br

G.H.T.A. Carvalho
Departamento de Engenharia Elétrica, ELE
University of Brasília, UnB
Brasilia, Brasil

R. V. C. Fernandes, F. H. Peixoto
M. S. Maia Filho, B. P. Sukiennik, L. S. Rosa
R. Z. M. Silva, T. A. Junquilho
Faculdade de Direito
University of Brasilia, UnB
Brasilia, Brazil

***Abstract*—The Brazilian Court System is currently the biggest judiciary system in the world, and receives an extremely high number of lawsuit cases every day. These cases need to be analyzed in order to be associated to relevant tags and allocated to the right team. Most of the cases reach the court as single PDF files containing multiple documents. One of the first steps for the analysis is to classify these documents. In this paper we present results on identifying these pieces of document using a simple convolutional neural network.**

***Index Terms*—Natural Language Processing, Convolutional Neural Networks, Document Classification.**

I. Introduction

The research and development project, entitled VICTOR, aims to solve problems of pattern recognition in texts of legal processes that reach the Brazilian Supreme Court (*Supremo tribunal Federal - STF*) [7], [8].

According to the STF, it would take 22,000 man-hours by its employees and trainees to analyze the approximately 42,000 processes received per semester. The court also points out that the time its employees spend on classifying these processes could be better applied at more complex stages of the judicial work flow.

The aim of VICTOR is to speed up the analysis of lawsuit cases that reach the supreme court by using document analysis and natural language processing tools. Most of the cases reach the court in the form of unstructured PDF volume which encloses several documents that have not been indexed. Therefore, in the first phase of this project, our goal is to classify these documents within PDF volumes.

This paper reports results of a preliminary evaluation on a dataset containing 6,814 documents from the STF. We propose a simple convolutional neural network architecture for this task and show that it obtains 90.35% accuracy on this dataset. In the next section we briefly describe our dataset and in section III we describe our method and summarize its results. This paper concludes in section IV.

II. Dataset for classification of Brazilian judicial documents

Our work focuses on classifying five main types of legal documents that make up the cases that are dealt by the STF. These are listed below, keeping their original label in Portuguese:

1) *Acórdão*
2) *Recurso Extraordinário (RE)*
3) *Agravo de Recurso Extraordinário (ARE)*
4) *Despacho*
5) *Sentença*
6) Others.

Note that the legal cases incluse several other types of documents, which we grouped in the class *Others*.

We developed an annotation tool which was used by a team of four lawyers who manually classified 6,814 documents. Figure 1 presents a pie diagram showing the proportion of documents in each of these classes.

The standard practice to train and evaluate machine learning methods requires that datasets are split into three parts: train, validation and test subsets [5]. We use stratified splits for each document class, maintaining the proportions of class samples in each subset. We used the following proportions, as detailed in Figure 2:

- 70% for the training set,
- 20% for validation and
- 10% for the test set.

We acknowledge the support of “Projeto de Pesquisa & Desenvolvimento de aprendizado de máquina (machine learning) sobre dados judiciais das repercussões gerais do Supremo Tribunal Federal - STF”.

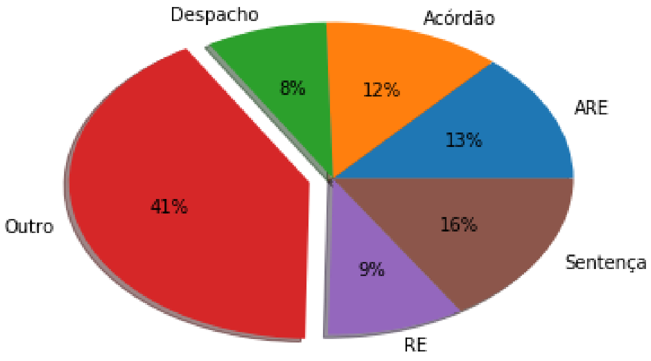


Fig. 1. Distribution of document classes in the dataset.

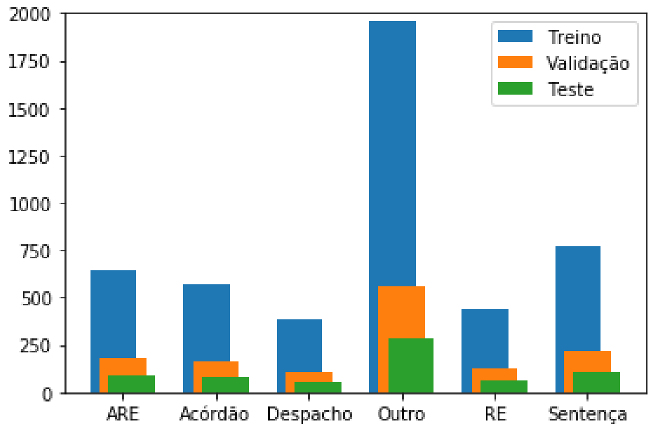


Fig. 2. Training, validation and test set distribution for each of the document classes.

A. Dataset issues

This dataset was provided by Supreme Court of Brazil to be used in the VICTOR project. Below we list some features that made the original dataset quite challenging.

- The STF receives processes from all the Brazilian courts of second instance and there is no pattern in the way they are written. The only requirement for admission is that the process is classified as a “Repercursão Geral” case, i.e., one of the predefined law process categories[1]. The flow of processes is detailed in [6].
- A significant part of the documents available in the court are in the form of raster images obtained by scanning printed documents, which often contain handwritten annotations, stamps, stains, etc.
- Furthermore, many of the processes are stored in the form of a series of PDF volumes, rather than a single PDF file that contains all its document. This was done to avoid file handling problems in legacy systems. The problem is that each a PDF volume often finishes in the middle of a document and the next PDF volume starts in the next page of that document.

III. Proposed method

A. Text extraction

The first step is to extract text from the PDF files, given all the challenges discussed in Section II-A. Figure 3 is a workflow showing how each page of the lawsuit processes is analyzed. First it is checked if its content is a raster (scanned) image or text. In case it is an image we apply an OCR (Optical Character Recognition) system [9] and then the resulting text is stored. In case that page embeds its text, its quality is verified by means of regular expressions. If the quality level is acceptable the text is stored, otherwise, the OCR is applied as if the page was in raster image format and the result is stored. The final result is that all pages of all suitcases processes are stored as text in a database to be used for further classification phase.

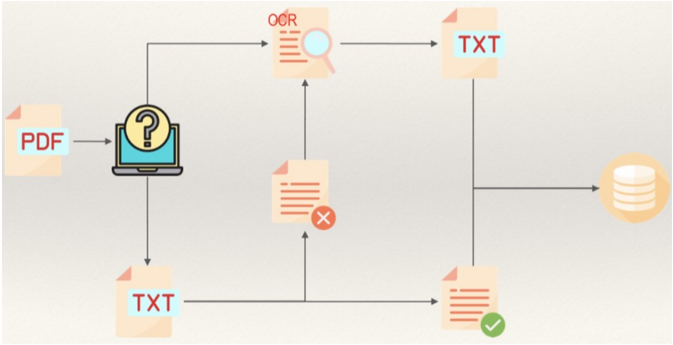


Fig. 3. Workflow for the extraction of text from the legal processes PDF files.

B. Preprocessing

To reduce the complexity of the dataset and improve the model’s accuracy, we applied regular expressions in order to filter special characters and recurring words as well as to emphasize important terms in the original texts [2]. The following steps were applied:

- Removal of special characters, such as # , @, $.
- Removal of alphanumeric terms with numbers and letters in the same “words”.
- Lowercasing of all characters.
- Transformation of terms that are emails or links into tokens "EMAIL" and "LINK".
- Transformation of terms referring to numbers of laws and articles in the tokens “*LEI_X*” (*LEI* means LAW) and “*ARTIGO_X*” (*ARTIGO* means ARTICLE) where X represents the respective law or article quoted. Figure 4 illustrates this process with an example.
- Stemming to reduce words into their stem, reducing the number of words with similar semantics. For that, we use the Natural Language Toolkit in Python, which implements the method of [4].

[1] The task of classifying legal processes as a whole (rather than their document parts) constitutes the main goal of the second phase of VICTOR project.

Text example before preprocessing:

Juiz Federal Relator, na\nforma do artigo 1o , inciso III, da Lei 11.419, de 19 de dezembro de 2006 e Resolução TRF 4a\nRegião no 17, de 26 de março de 2010. A conferência da autenticidade do documento está\ndisponível no endereço eletrônico http://www.jfpr.jus.br/gedpro/verifica/verifica.php

Text example after preprocessing:

juiz federal relator forma inciso iii da LEI_11419 de de dezembro resolução trf região março a conferência autenticidade document está disponível endereço eletrônico SITE

Fig. 4. Example of text before and after the application of all preprocessing steps except for stemming.

C. Convolutional neural network

The literature shows that one of the state-of-the-art approaches for document classification consists in applying a Convolutional Neural Network (CNN) on embedded text. We designed a system that was inspired by the framework proposed by Conneu et al. [1], though our model is much simpler and therefore training requires less computational power and has a lighter GPU memory footprint.

Our architecture is illustrated in Figure 5. The first step is to apply an embedding method that transforms the data into a 2D tensor with the dimensions of (2000, 100). Next, a convolutional layer is added with kernel size 4 and 256 filters resulting in a output of dimensions (2000, 256). Then, a max pooling layer chooses the part of the data with greater relevance for classification of documents. The resulting tensor is flattened, leading to a one-dimensional array of 256000 dimensions. The last layer is a fully connected layer with a softmax activation function. This network was trained using the categorical cross-entropy as its loss function and the Adam optimization method.

D. Results

Our CNN-based method achieves an accuracy of 90.35% and F1 score of 0.91. Figure 6 shows the confusion matrix obtained in our test set.

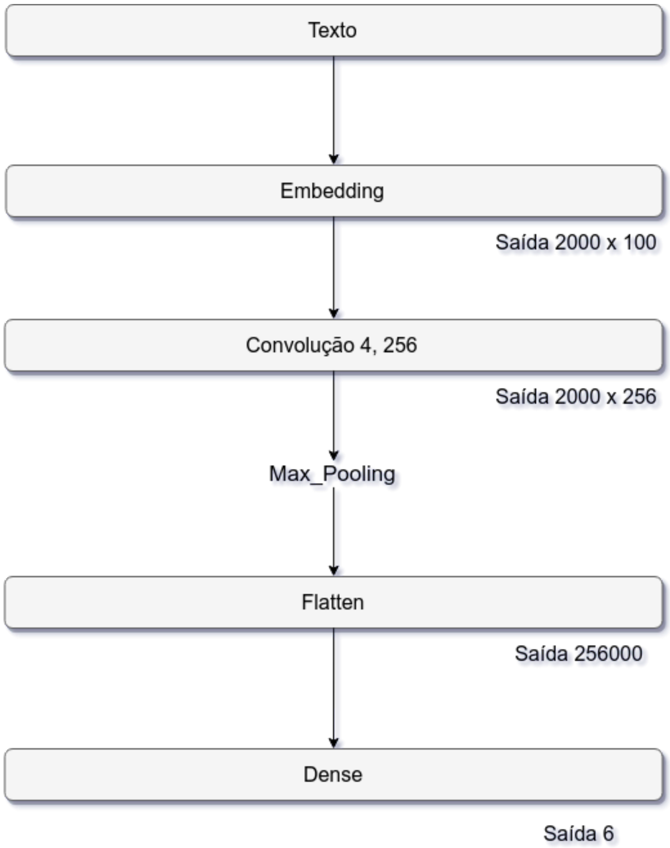
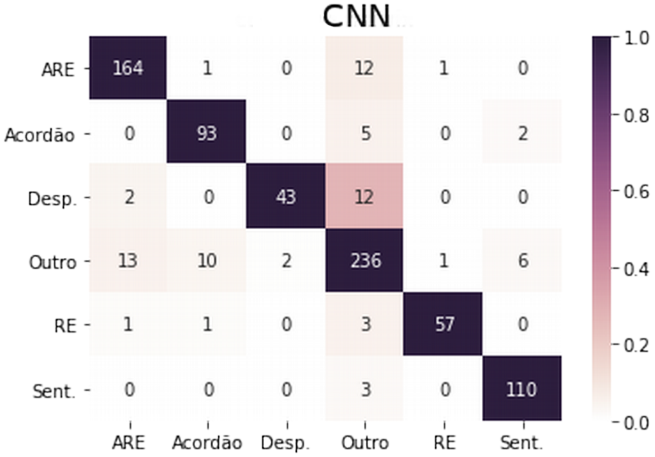


Fig. 5. CNN Architecture diagram



	ARE	Acordão	Desp.	Outro	RE	Sent.
ARE	164	1	0	12	1	0
Acordão	0	93	0	5	0	2
Desp.	2	0	43	12	0	0
Outro	13	10	2	236	1	6
RE	1	1	0	3	57	0
Sent.	0	0	0	3	0	110

Fig. 6. Confusion Matrix.

IV. Conclusion

The development of a pattern recognition applied to the lawsuit cases of the Brazilian Supreme Court is a very challenging task. It is necessary to face a huge amount of non structured data (about 350 new lawsuit cases per day which contents can be either raster images or texts. The results obtained show that our CNN model was a good choice for document piece classification.

The introduction of a simple convolutional network can efficiently reduce the burden of the laborious and time consuming task of process management in the supreme court. The document classification bottleneck can be eliminated, which means that lawyers who work at STF can dedicate their time to less mechanical tasks.

A more efficient workflow benefits not only the court, but the whole society in Brazil, inspiring other institutions to invest in machine learning solutions to improve their activities.

One of the possible directions for future work is the use a method for automatic named entity recognition, such as that of [3], to replace or complement the regular expressions used in the preprocessing steps. Another direction of work is the use these document classification results as part of a process to automatically classify lawsuit cases as a whole (rather than only their parts).

References

[1] Alexis Conneau, Holger Schwenk, Loïc Barrault, and Yann Lecun. Very deep convolutional networks for text classification. Technical report, Cornell University Library, CoRR/cs.CL, 2016. arXiv:1606.01781.

[2] Yang Liu Li Deng. *Deep Learning in Natural Language Processing*. Springer, 2018.

[3] Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. Lener-br: a dataset for named entity recognition in brazilian legal text. In *International Conference on the Computational Processing of Portuguese (PROPOR)*, Canela, RS, Brazil, September 24-26 2018.

[4] Robert A. N. de Oliveira and Methanias C. Junior. Experimental analysis of stemming on jurisprudential documents retrieval. *Information*, 9(2), 2018.

[5] Joaquin Qui nonero Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. *Dataset Shift in Machine Learning*. Neural Information Processing. The MIT Press, 2009.

[6] Portal do Supremo Tribunal Federal. Estatísticas do STF. Online: http://www.stf.jus.br/portal/cms/verTexto.asp?servico=estatistica&pagina=comrecvisaogeral, Website accessed in September 2018.

[7] Portal do Supremo Tribunal Federal. Inteligência artificial vai agilizar a tramitação de processos no STF. Online: http://www.stf.jus.br/portal/cms/verNoticiaDetalhe.asp?idConteudo=380038, May 30 2018.

[8] Portal do Supremo Tribunal Federal. Ministra Cármen Lúcia anuncia início de funcionamento do projeto Victor, de inteligência artificial. Online: http://www.stf.jus.br/portal/cms/verNoticiaDetalhe.asp?idConteudo=388443, August 30 2018.

[9] Ray Smith. An overview of the Tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR)*, volume 2, pages 629–633. IEEE, 2007.