Document classification using a Bi-LSTM to unclog Brazil's supreme court

**F.A. Braz, N.C. Silva, T.E. de Campos,[*] F.B.S. Chaves, M.H.P. Ferreira,
P.H.G. Inazawa, V.H.D. Coelho, B.P. Sukiennik, A.P.G.S. Almeida, F.B. Vidal, D. Alves Bezerra,
D.B. Gusmão, G.G. Ziegler, R.V.C. Fernandes, R. Zumblick, F. Hartmann Peixoto**
Universidade de Brasília (UnB), DF, Brazil
{fabraz,niltoncs,teodecampos}@unb.br

Abstract

The Brazilian court system is currently the most clogged up judiciary system in the world. Thousands of lawsuit cases reach the supreme court 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 raster scanned documents with widely variable levels of quality. One of the first steps for the analysis is to classify these documents. In this paper we present a Bidirectional Long Short-Term Memory network (Bi-LSTM) to classify these pieces of legal document.

1 Introduction

Ensuring equal access to justice for all is a real challenge in Brazil due to the sluggishness of judicial process and the high volume of new cases entering in the justice system every year. In the Brazilian higher courts, the average time to reach the sentence is 11 months in the Superior Court of Justice (STJ), 1 year and 2 months in the Superior Court of Labor (TST) and 8 months in the Superior Electoral Court (TSE). This scenario is worsening each year: around 28.8 million new cases reach the Brazilian Judiciary and there are already approximately 79.6 million other cases in stock in Brazilian courts, related to 2016. Adding up these figures, there are approximately 108.4 million legal processes being carried on by Brazilian judiciary [1]. This overload has a cost of USD 22 billion per year, according to National Justice Council (CNJ), and is one of the main causes of legal insecurity and criminal impunity in the country. The congestion rate was 73.0% in 2016. That means that only 27% of all the cases processed were solved in that year.

Our work aims to contribute towards goal 16.3 of the UN agenda 2030: "ensure equal access to justice for all" [15]. To overcome the obstacles that prevent the Brazilian judicial system from democratically guaranteeing the right to due process of law, we propose to apply machine learning methods to classify legal processes that reach the Brazilian Supreme Court (in Portuguese, *Supremo Tribunal Federal - STF*) [9, 10]. 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 [1]. 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.

Most of the cases reach the court are in the form of PDF files with raster scanned documents. About 10% of them are completely unstructured volumes which encloses several documents per file without any digital index. In Brazil's judiciary, documents that compose a legal case are grouped into briefs or parts (called *peças* in Portuguese) which are categorized into a set of classes. In order to speed up the analysis of cases, the first step needed is to automate the classification of these documents.

[*] The authors belong to three diferent insitutions of UnB: Faculdade de Engenharias do Gama, Departamento de Ciência da Computação and Faculdade de Direito. Further information at https://cic.unb.br/˜teodecampos/ViP

This work was presented at NIPS 2018 Workshop on Machine Learning for the Developing World (ML4D).

This paper reports results of a preliminary evaluation on a dataset containing 6,814 documents (*peças*) from the STF. We propose a Bidirectional Long Short-Term Memory network architecture for this task and show that it obtains 84% F_1 score on this dataset with no pre-processing.

2 Proposed method

2.1 Text extraction

The first step is to extract text from the PDF file. If the content of a document page is a raster image, we apply the Tesseract OCR system [13] and store the text. If the page embeds its text as metadata, its quality is verified using 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.

2.2 Bidirectional Long Short-Term Memory Model

The proposed model is based on recurrent models to deal with text as sequential information. Long Short-Term Memory (LSTM) models use the information from the previous status of neurons [5]. More contextual information can be extracted using a bidirectional LSTM (Bi-LSTM) [4]. The data processing flows forward and backward at same time [12] and the output of each LSTM are merged using their sum [4].

Bi-LSTM models have shown to be effective in problems of speech recognition [3, 7, 14, 11, 6], being a state-of-the-art model to classify sequential data into multiple classes. Documents are sequences of words and our problem is also multi-class, therefore Bi-LSTM is a suitable model.

Our architecture (shown in Figure 1) is composed of three main layers with 1000 tokens of input. We followed [11, 6] and used word embedding as an input layer of the Bi-LSTM. This layer transforms each token in a distributed array of 100 dimensions. The recurrent layer has two hidden LSTM, a forward and backward layer model, each with 200 memory blocks and one-cell. The output of this layer uses a ReLu activation. The two hidden LSTMs are combined by adding their outputs. The last layer is dense, with 6 output neurons and a Softmax activation function.

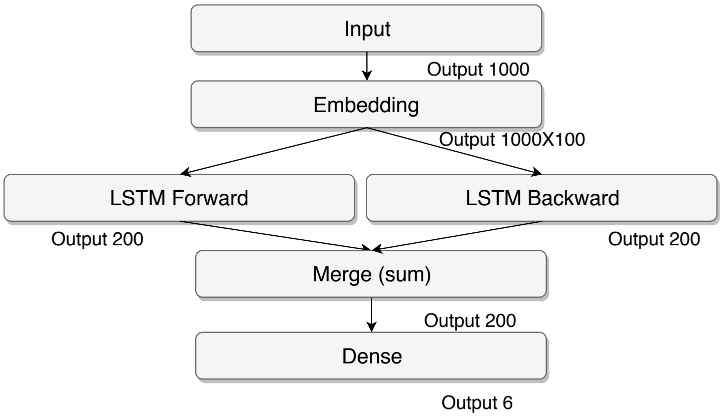

Figure 1: Bi-LSTM Architecture diagram

3 VICTOR project's dataset of brief parts (*peças*)

We used the dataset described in [2]. Our work focuses on classifying five main types of legal documents handled by the Brazilian supreme court (STF). Documents that do not belong to these classes are grouped in a class called *Others*. Table 1 presents the classes used and the number of samples in each of them.

A total of 6,814 text documents were manually labeled by a team of four specialist lawyers. This dataset was split into three parts: 70% of the samples for training, 20% for validation and 10% for test.

Table 1: Set of classes in VICTOR project's dataset of document types (*peças*).

Label (in Portuguese)	Description	Samples
Acórdão	Appelate Decision	82
Recurso Extraordinário (RE)	Extraordinary Appeal	63
Agravo de Recurso Extraordinário (ARE)	Extraordinary Appeal Bill/Review	92
Despacho	Administrative Orders	55
Sentença	Judgement	110
Outros	Other Documents Types	280

This dataset is quite challenging as it has a high level of within class diversity. The documents do not follow a standard, not only in terms of layout but also in terms of scanned image quality and whether or not they embed digital text or have only raster images. Furthermore, a significant part of these documents often contained handwritten annotations, stamps, stains etc. As discussed earlier, the PDF files are often not indexed and some of them include several documents.

4 Experiments and results

Although legal documents can be quite long, the first page is usually the most informative. Furthermore, the files in VICTOR dataset often include multiple documents, so later pages in a PDF file can be unreliable. For these reasons, our Bi-LSTM model was designed to take inputs of 1000 tokens, which usually covers most of the contents of one page. Our evaluations on the validation set show that this was discriminative enough. One major advantage of this w.r.t. using text from the whole document is that the main bottleneck of our system is the OCR system, which takes 1s per processing core to run on each page. By using a method that efficiently exploits the first 1000 tokens, we only need to run the OCR on up to two pages per document.

Since most of the text was obtained by running an OCR raster scanned noisy documents (rather than extracting text from PDFs metadata), many out-of-vocabulary 'words' appeared in the text. One can deal with this problem by using a number of regular expressions and stemming, as done in [2] and/or by running a named entity recognition system such as that of [8]. In this paper, we simply limited the tokenizer's vocabulary to the 100,000 most frequent distinct words. This was enough to include all relevant words as well as specific symbols of the judiciary system, such as law numbers and Latin words.

Our model was trained for 20 epochs with a batch size of 64 samples and a learning rate of 0.001. The total training time was of 120.02s on a NVidia Titan XP, which has 12GB of RAM. In this setup, prediction on is done in 1.47ms per document. In contrast, the CNN model of [2] takes 5.87ms per document, excluding the time to run the OCR (1s per page per CPU core).

Figures 2 and 3 detail our results. Our BI-LSTM model archives a mean precision of 85% and f_1-Score 84% without requiring any preprocessing heuristics and regular expressions. This contrasts with the CNN model of [2], which uses a set of hand crafted rules in the tokenization process as well as regular expressions to remove noisy text.

5 Conclusion

We proposed a tool to significantly speed up the first steps of the analysis of legal documents that reach the *Supremo Tribunal Federal* (STF, from Brazil), which is the world's most clogged up supreme court. The task consists in classifying legal briefs (*peças*) into a set of 6 classes. For that we introduced a Bidirectional Long Short-Term Memory model which processes the first 1000 tokens of the documents, i.e., usually just the first page. This model is strong enough to classify these documents with an F_1 score of 84%, dismissing the need to run an OCR on the remaining pages of the document.

The next step of this project consists in designing a tool that combines information from all documents that compose a legal case in order to aid the decision making in the judgment process.

	prec.	rec.	F_1
ARE	0.82	0.84	0.83
Acordão	0.71	0.89	0.79
Desp.	0.74	0.82	0.78
Outro	0.91	0.82	0.87
RE	0.77	0.70	0.73
Sent.	0.92	0.95	0.93
average	0.85	0.84	0.84

Figure 2: Bi-LSTM results per class.

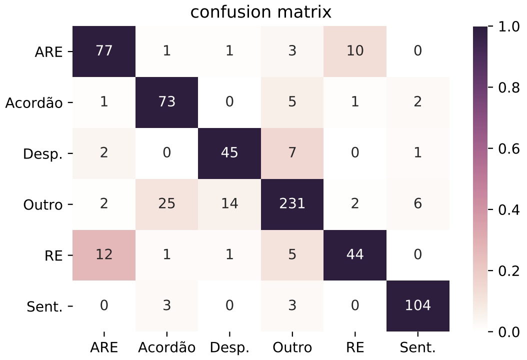

Figure 3: Confusion matrix.

Acknowledgments

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".

References

[1] Conselho Nacional de Justiça. Justiça em números 2017: ano-base 2016. Online, available from `http://www.cnj.jus.br/publicacoes`, Brasilia, Brazil, 2017.

[2] N. Correia da Silva, F. A. Braz, T. E. de Campos, D.B. Gusmao, F.B. Chaves, D.B. Mendes, D.A. Bezerra, G.G. Ziegler, L.H. Horinouchi, M.H.P. Ferreira, G.H.T.A. Carvalho, R. V. C. Fernandes, F. H. Peixoto, M. S. Maia Filho, B. P. Sukiennik, L. S. Rosa, R. Z. M. Silva, and T. A. Junquilho. Document type classification for brazil's supreme court using a convolutional neural network. In *10th International Conference on Forensic Computer Science and Cyber Law (ICoFCS)*, Sao Paulo, Brazil, October 29-30 2018.

[3] Reza Ghaeini, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Yuan Ling, Aaditya Prakash, Xiaoli Z. Fern, and Oladimeji Farri. Dr-bilstm: Dependent reading bidirectional LSTM for natural language inference. *CoRR*, abs/1802.05577, 2018.

[4] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5):602–610, July 2005.

[5] S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, November 1997.

[6] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991, 2015.

[7] Eliyahu Kiperwasser and Yoav Goldberg. Simple and accurate dependency parsing using bidirectional LSTM feature representations. *CoRR*, abs/1603.04351, 2016.

[8] 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.

[9] 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.

[10] 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.

[11] Adithya Rao and Nemanja Spasojevic. Actionable and political text classification using word embeddings and LSTM. Technical report, Lithium Technologies, July 2016. arXiv:1607.02501.

[12] M. Schuster and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, November 1997.

[13] 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.

[14] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *CoRR*, abs/1503.00075, 2015.

[15] UN General Assembly. Transforming our world : the 2030 agenda for sustainable development. Technical Report A/RES/70/1, United Nations, October 21 2015. Available at `http://www.refworld.org/docid/57b6e3e44.html`.