



A LSTM Recurrent Neural Network Implementation for Classifying Entities on Brazilian Legal Documents

Rafael Mecheseregian Razeira and Ildeberto Aparecido Rodello^(✉)

University of São Paulo, Ribeirão Preto, SP 14049-900, Brazil
rodello@usp.br

Abstract. Although the use of Natural Language Processing and Named Entity Recognition methods to deal with classification problems in the most diverse areas is something well-established, applications in Law offer challenges due to the specific terminology, broader vocabulary and the presence of more complex semantic and syntactic structures compared to spoken Portuguese. In this short paper, we present a Recurrent Neural Network implementation using LSTM to classify entities such as class, subject, value, individuals, among others items, in lawsuits. The initial dataset focused on a sample of 100 thousand lawsuits of São Paulo state court, in Brazil. The proposed method achieved an accuracy and F1-score of approximately 90% in the tested data. Such preliminary results indicate that it is possible to create a model capable of generalizing such classifications on a large scale even regarding the specifics of Brazilian legal texts terminology.

Keywords: Natural Language Processing · Named entity recognition · Legal documents

1 Introduction

Currently, information has been recognized as a strategic asset for any organization to be effective in decision-making. The data gathered from internal and external sources constitute this strategic asset and are a particularly big challenge for modern Information and Communication Technology (ICT) systems and organizations [1].

According to this perspective, by harnessing the power of computational resources, lawyers can predict from past experiences more accurately how events will unfold in litigation. Thus, lawyers who adopt data-driven decision-making will have a clear advantage over their colleagues who still cling to their outdated instruments [2].

In many ways, artificial neural networks (ANN) have been successfully applied to diverse problems. Among them, it is possible to mention pattern recognition [3]. Regarding Law, it is possible to find citations of ANN usage to aid the measurement of punishment [4] and in fighting cybercrime [5]. Other attempts to use machine learning models in the Law area come across a lack of data, especially in Named Entity Recognition (NER) problems. Even so, some attempts to create a public database for the application

of Law written in Brazilian Portuguese had good results [6]. However, the amount of data made available is a trammel since they are in the hundreds or thousands, which is difficult to train deep learning models with results expressively good to overcome the human performance [6].

In this context, the project “Mediation and Conciliation empirically evaluated: jurimetry for proposing efficient actions” was an empirical research that studied ways to improve the delivery of the judicial provision [7]. It consists of a database with an extensive collection of lawsuits ranging from the years 2015 to 2018 collected from public data courts and the Court of Auditors Justice (TJ) in Brazil. Based on this database, initial analysis was performed on quantitative information about the lawsuits, which are extremely relevant to decision-making and action proposals, such as the demands average duration and the values being negotiated, among others.

In contrast to other areas for NER task applications, legal texts in Brazilian Portuguese have some issues that hinder such application. For instance, non-regular patterns in capitalization, punctuation and the text structure, as well as semantic structure and the jargon of Law. These particularities tend to increase the difficulty in obtaining a good dataset for training the model. In addition, the NER datasets for general purpose, do not have specific tags to Law.

This paper presents an approach for developing a dataset to train a deep learning model in legal texts and an implementation of a Recurrent Neural Network (RNN), applied in Law, to perform some types of classification of legal texts published in official journals, as interest in specific types of elements belonging to the lawsuit, such as its subject, classes, monetary values involved, among others.

The paper is organized as follows. Section 2 describes the data source as well as the tagging procedure. Section 3 presents the methodology, specifically how the model was implemented. The preliminary results are presented and discussed on Sect. 4. Finally, the Sect. 5 briefly presents our conclusions and the future work.

2 The Database

The data source used by this study is the database generated from the research project “Mediation and Conciliation empirically evaluated: jurimetry to propose efficient actions” [7]. The database contains lawsuits referring to some Brazilian states with each sample built up of the following fields: *id*, the database unique identifier. *Dados Fórum*, the judge and court identification. *Conteúdo*, the content of the lawsuit, i. e., a plain text written by the judge for describing it as well as indicating decisions and following, among other information. *Processo*, the lawsuit id according to the CNJ (National Council of Justice). *Assunto*, the subject of the lawsuit. *Classe*, the class of the lawsuit. *Data*, the date of publication. *Autor*, the plaintiffs involved in the lawsuit. *Reu*, the defendants. *Valores*, the monetary values. Table 1 shows a sample of a record from the database.

Among the mentioned fields, *Conteúdo* refers to the source for training the deep learning model for recognizing words that have the meaning of classification to subjects, classes, legal entities, values, etc. The procedure of building the new data set will be further elaborated in the next section.

Table. 1. Sample of a record of the database (names were suppressed).

Field	Description	Content
id	Database ID	5e6135e5a15735ce218a9960
Dados Fórum	Lawsuit forum information	JUÍZO DE DIREITO DA 4ª VARA CÍVEL JUIZ(A) DE DIREITO [REDACTED] [REDACTED] ESCRIVÃ(O) JUDICIAL [REDACTED]
Processo	Lawsuit number	0001543-25.2013.8.26.0344 Processo 0001543-25.2013.8.26.0344 (034.42.0130.001543) - Procedimento Comum - Evicção ou Vício Redibitório - [REDACTED] [REDACTED] - Sobre a contestação apresentada pelo Curador Especial nas fls. 211, manifeste-se o requerente. Prazo: 15 (quinze) dias. - ADV: [REDACTED] [REDACTED]
Assunto	The subject of the lawsuit	Null
Classe	The class of the lawsuit	Null
Data	The date of the lawsuits was published	2017/07/11
Autor	Plaintiffs	'Nome': [REDACTED], 'Gênero': 'Masculino', 'Pessoa': 'Física'
Reu	Defendants	'Nome': [REDACTED], 'Gênero': 'Masculino', 'Pessoa': 'Física', 'Nome': [REDACTED], 'Gênero': 'Masculino', 'Pessoa': 'Física'
Valores	The monetary values in the lawsuit	Null

Note in the sample record of Table 1 that some fields are null or missing. That is one of the reasons to create a model capable of correctly predicting such information, since, among all sample cases in the mentioned before database, which reaches at more than 100 million, only a few of them have this classification. Something which is a hindrance for new studies on this data, for example, if some interested researcher lawyer in study all the cases in which the class is “*Mandato de Segurança*” (Writ of Mandamus) to know what’s the best arguments, or the history line, etc. Without this classification the research will be depreciated or not even possible

2.1 The Dataset Tagging Procedure

The database used for creating a dataset is composed of lawsuits from the state of *São Paulo* court. It contains around 90 million records representing around 19.183.228 unique lawsuits. From that, a sample of one hundred thousand was taken for the training base.

Analyzing the *Conteúdo* field, at first glance there is a possible division by the character ‘ - ’ in order of process number, class, subject, parts and content as also shown in Table 1. However, not all documents followed the same pattern and around 12.080 documents from the sample were not used for constructing the database.

After that, the words were separated and tagged, comparing lists of classes and subjects according to CNJ class or subject label pattern (https://www.cnj.jus.br/sgt/conulta_publica_classes.php). Finally, for the remaining content, it was verified who were the plaintiffs and defendants, as well as their gender, if it were a legal entity, government or legal confidentiality. Such words were searched for in the content and tagged.

3 Methodology

It was used different methods for labeling the tags of each word: 1) IOB (Inside-Outside-Beginning) tagging scheme [8], where “B-” indicates the beginning of the Tag, “I-” indicates that the word is within a Tag and “O” indicates that the word does not belong to any Tag and 2) The joining of IOB tags for only a single classification of the words. The description of each acronym is described in Table 2.

Table 2. The Tags and respective descriptions.

IOB TAG	Description	NOT IOB	Description
O	Does not belong to any tag of interest	O	Does not belong to any tag of interest
B-Cla	Begin of the class	Cla	The word is a class
I-Cla	Inside of the class	–	–
B-Ass	Begin of the subject	Ass	The word is a subject
I-Ass	Inside of the subject	–	–
B-Jur	Begin of legal person	Jur	The word is a legal person
I-Jur	Inside of legal person	–	–
B-FiM	Begin of male person	FiM	The word is a male person
I-FiM	Inside of a male person	–	–
B-FiF	Begin of a female person	FiF	The word is a female person
I-FiF	Inside of a female person	–	–
B-Gov	Begin of a government name	Gov	The word is a government name

(continued)

Table 2. (continued)

IOB TAG	Description	NOT IOB	Description
I-Gov	Inside of a government name	–	–
B-Seg	Begin of a legal confidentiality name person	Seg	The word is a legal confidentiality person
I-Seg	Inside of a legal confidentiality name person	–	–
B-Val	Begin of a value in process	Val	The word is a value in the process
I-Val	Inside of a value in process	–	–

3.1 The Implemented Model

The architecture of the implemented model consists of 3 layers: Embedding layer, Bidirectional LSTM layer and Dense layer. Such a network was implemented using Python 3 and deep learning libraries as Tensorflow [9] and Keras [10].

The size tested of the embedding vector was 50 and 100. The Bidirectional layer groups LSTM neurons [11], with a spatial dropout of 0.5, and finally, a last layer called Dense, is a layer of perceptrons containing 17 neurons, with a softmax activation function to calculate the probability of the output for each word. Each neuron is associated with a tag that words can belong to.



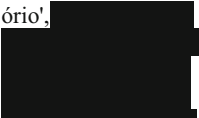
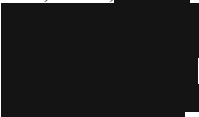

3.2 Pipelines

Before training the model, we replace each word in the documents with numbers that represented them. For instance, the word “Processo” was replaced by the number 16 in that position, the word “Procedimento” was replaced by the number 15679, and so on. In addition, the padding-post technique [12] was used in documents with a fixed length of 150 words. The padding technique consists of adding an arbitrary word to the document if it is less than 150 words, or removing words from the end until we are left with 150. Table 3 exemplifies the pre-processing. In addition, two forms of class balancing were applied: Balanced Under Sample (BUS) [13] and Random Under Sample (RUS) [14]. Each balancing method was tested for both forms of tagging.

The BUS algorithm consists of keeping the context of the interesting tags. In contrast to the RUS algorithm, in which each word is removed randomly, the BUS algorithm maintains a rate of words in the proximity of a tag of interest, while removing the words that are in the middle of the two tags. For instance, if the entire raw text is like in Table 3 (column “Original Document”) that the interesting word tags go from “Procedimento” to “Santos” the final results of a under the sample is a like the phrase starting in “0130” (2 words before) to “contestação” (2 words after), in the column “Word to Numbers with Padding”.

Finally, the dataset was divided into 90% for training and 10% for testing using the metric F1-score and precision to assess the overall performance of the model. During training Adam [15] was used as an optimizer, sparse categorical cross entropy as a loss

Table 3. Example of a padding post method (names were suppressed).

Original Document	Words Extracted	Word to Numbers without Padding	Word to Numbers with Padding
Processo 0001543- 25.2013.8.26.0344(034.4 2.0130.001543) - Procedimento Comum - Evecção ou Vicio Redibitório - Erick  s - Sobre a contestação apresentada pelo Curador Especial nas fls. 211, manifeste- se o requerente. Prazo: 15 (quinze) dias. - ADV: 	['Processo','0001543' , '25','2013','8','26','03 44','034','42','0130','0 01543'],'Procedimen to','Comum','Evecçã o','ou','Vicio','Redibit ório',  , 'a','contestação' , 'apresentada','pelo',' Curador','Especial','n as','fls','211','manifes te','se','o','requerente' , 'Prazo','15','quinze',' dias','ADV', 	[16,45585,417,59,1 5,17,21506,8342,53 8,610,45586,44,231 ,21360,30,21005,21 361,24630,1472,25, 260,120686,1547,1 787,25,260,3210,17 87,25,260,11507,3, 230,638,42,4967,11 1,344,31,4662,478, 8,5,193,880,109,21 50,33,18,15995,109 06,165150,10,7487 0,11,2588,1356,131 6,10,179818,11] 	[6,45585,417,59,1 5,17,21506,8342,5 38,610,45586,44,2 31,21360,30,21005 ,21361,24630,1472 ,25,260,120686,15 47,1787,25,260,32 10,1787,25,260,11 507,3,230,638,42,4 967,111,344,31,46 62,478,8,5,193,880 ,109,2150,33,18,15 995,10906,165150, 10,74870,11,2588, 1356,1316,10,1798 18,11,0,0,0,...,0]

function, with a batch size of 128, during 50 epochs and the test data was separated for validation during training.

In short, given the dataset, all of its words were changed to numbers, a padding was applied, a form of tagging was chosen (IOB or NOT IOB), one of the class balance methods was applied (BUS, RUS or none), a size of embedding was chosen and finally trained the network. Thus, totaling 12 tested methodologies.

4 Preliminary Results

The main metric used to assess the overall performance of the models was F1-score. Also, sensitivity and specificity were calculated for each tag and in the scope of measuring the performance of the methodologies. In addition to the F1-Score, the overall precision of the model was calculated. Such metrics were calculated for both each balancing method tested and only on the test data. Some results are available in Table 4 and Table 5, presenting the performance of tested methodologies.

According to Table 5, the model can predict with an F1-score above 89% and accuracy above 97%. Whatever the methodology tried, there are not significant differences between them. At the most the two undersample methods increase the overall accuracy in NOT IOB methods in 1%.

Table 4. Values of model performance using BUS and IOB Tags with size 50 of embedding.

Tag	Precision	Recall	F1-Score	Samples
O	0.99	0.99	0.99	741782
I-Cla	1.00	1.00	1.00	18093
I-Seg	0.85	0.85	0.85	3539
I-Jur	0.79	0.78	0.79	16520
I-FiM	0.79	0.79	0.79	13174
I-FiF	0.82	0.79	0.80	9960
I-Gov	0.85	0.73	0.78	6246
I-Ass	1.00	1.00	1.00	9746
I-Val	0.95	0.97	0.96	2273
B-Cla	1.00	1.00	1.00	8655
B-Seg	0.92	0.96	0.94	1385
B-Jur	0.83	0.85	0.84	3224
B-FiM	0.81	0.88	0.84	4202
B-FiF	0.85	0.84	0.85	2849
B-Gov	0.83	0.82	0.83	1067
B-Ass	1.00	1.00	1.00	5636
B-Val	0.86	0.84	0.85	1032

Table 5. General performance of the methodologies.

Methodology	Embedding dimensions	F1-Score	Recall	Precision	Accuracy
BUS + IOB Tag	50	0.89	0.89	0.89	0.98
BUS + IOB Tag	100	0.89	0.89	0.89	0.98
BUS + NOT IOB	50	0.89	0.89	0.89	0.98
BUS + NOT IOB	100	0.89	0.90	0.89	0.98
RUS + IOB Tag	50	0.89	0.89	0.88	0.98
RUS + IOB Tag	100	0.89	0.88	0.90	0.98
RUS + NOT IOB	50	0.89	0.89	0.89	0.98
RUS + NOT IOB	100	0.89	0.89	0.89	0.98
NO Undersample + IOB Tag	50	0.89	0.90	0.88	0.98
NO Undersample + IOB Tag	100	0.89	0.89	0.89	0.97

(continued)

Table 5. *(continued)*

Methodology	Embedding dimensions	F1-Score	Recall	Precision	Accuracy
NO Undersample + NOT IOB	50	0.89	0.90	0.88	0.97
NO Undersample + NOT IOB	100	0.89	0.90	0.89	0.97

The point to be observed is that two subject and class tags present both an F1-Score and precision stagnant at 1.00 or 0.99, something that would normally be considered a good level. However, in this case, the words of class and subject appear at the beginning of the text, which provide an inconvenience in case of applying the model to other legal texts that contain its class or subject not in the same region of the text. Probably these words will not be classified correctly.

5 Conclusion and Future Work

The short paper presented preliminary results of a LSTM Recurrent Neural Network implementation for classifying entities on Brazilian legal Documents. In the sample of 100 thousand dataset lawsuits from Brazilian courts, the model achieves good results. These results suggest that the applied technique based on Recurrent Neural Network using LSTM can be successful to classify entities for Brazilian legal documents.

In addition, we intend to slice a 5% of train set to fine-tuning other hyperparameters such as the size of embedding vectors, the learning rate and dropout, among others and tests on the test set previous apart.

Finally, implement the methodology with the best results obtained in a data set using the entire database of 90 million of lawsuits.

Acknowledgments. This project has been partially supported by Huawei do Brasil Telecomunicações Ltda (Fundunesp process # 3123/2020) and by University of São Paulo (Portaria PRP Nº 668, DE 17 DE OUTUBRO DE 2018).

References

1. Laudon, K., Laudon, J.: Sistemas de informações gerenciais. 11a. ed. [s.l.] Pearson/Prentice Hall, Upper Saddle River (2015)
2. Lettieri, N., et al.: Ex machina: analytical platforms, law and the challenges of computational legal science. *Future Internet* **10**(5), 26 (2018)
3. Bishop, C.M.: *Neural Networks for Pattern Recognition*. Oxford University Press (1995). ISBN: 0198538642
4. Grimm, C.: *Dosimetria da pena utilizando redes neurais*. Monografia. Curso de Direito. Universidade Federal do Paraná. Curitiba (2006)
5. Lossio, C.J.B.: *O anticrime nas redes sociais: os algoritmos e a rede neural artificial (RNA) em face do cybercrime* (2017)

6. Luz, P., de Araujo, T., Campos, R., Oliveira, M., Couto, S., Bermejo, P.: Lener-br: a dataset for named entity recognition in brazilian legal text. In: Villavicencio, A., et al. (eds.) PROPOR 2018. LNCS (LNAI), vol. 11122, pp. 313–323. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-99722-3_32
7. Ventura, C.A.A., et al.: Justiça Pesquisa - Mediações e Conciliações Avaliadas Empiricamente. <https://www.cnj.jus.br/wp-content/uploads/2011/02/e1d2138e482686bc5b66d18f0b0f4b16.pdf>>, Accessed 15 Jan 2021
8. Ramshaw, L.A., Marcus, M.P.: Text chunking using transformation-based learning. In: Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., Yarowsky, D. (eds.) Natural language processing using very large corpora, pp. 157–176. Springer Netherlands, Dordrecht (1999). https://doi.org/10.1007/978-94-017-2390-9_10
9. Tensorflow. <https://www.tensorflow.org/>, Accessed 12 Aug 2020
10. Keras. <https://keras.io>, Accessed 12 Aug 2020
11. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **5**(2), 157–166 (1994)
12. Dwarampudi, M., Reddy, N.V.: Effects of padding on LSTMs and CNNs. *arXiv preprint arXiv:1903.07288* (2019)
13. Akkasi, A., Varoğlu, E., Dimililer, N.: Balanced undersampling: a novel sentence-based undersampling method to improve recognition of named entities in chemical and biomedical text. *Appl. Intell.* **48**(8), 1965–1978 (2017)
14. Ganganwar, V.: An overview of classification algorithms for imbalanced datasets. *Int. J. Emerg. Technol. Adv. Eng.* **2**(4), 42–47 (2012)
15. Kingma, D.P., Jimmy, B.A.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)