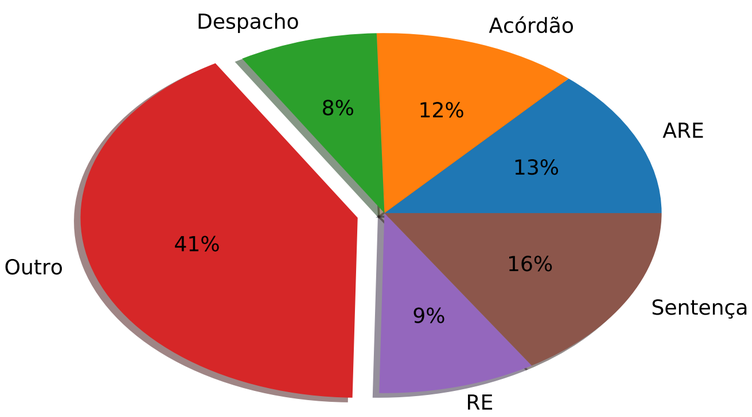
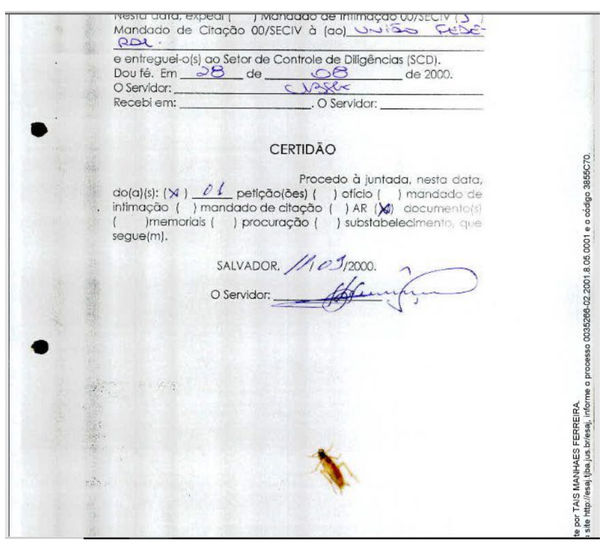
Document classification using a Bi-LSTM to unclog Brazil's supreme court

F.A. Braz, N.C. Silva, T.E. de Campos, F.B. Chaves, M.H.P. Ferreira, A.P.G.S. Almeida,
F.B. Vidal, P.H.G. Inazawa, V.H.D. Coelho, B.P. Sukiennik, D. Alves Bezerra,
D.B. Gusmão, G.G. Ziegler, R.V.C. Fernandes, F. Hartmann Peixoto
{Faculdade Gama, Departamento de Ciência da Computação, Faculdade de Direito}
University of Brasília, Brazil

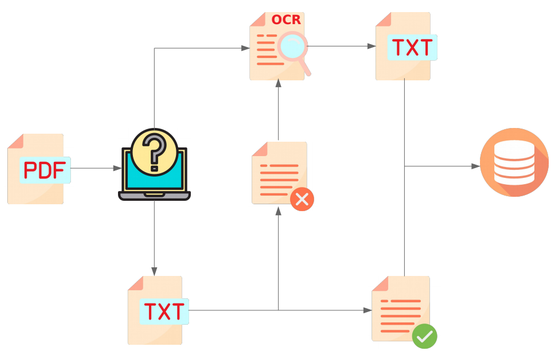
Contribution

- Brazilian court system: the most clogged up judiciary system in the world.
- Thousands of lawsuit cases reach the supreme court every day.
- These cases need to be classified so that they can be allocated to the right team.
- First step: split cases into a set of labeled documents.
- VICTOR dataset (da Silva et al., 2018): 6,813 documents.

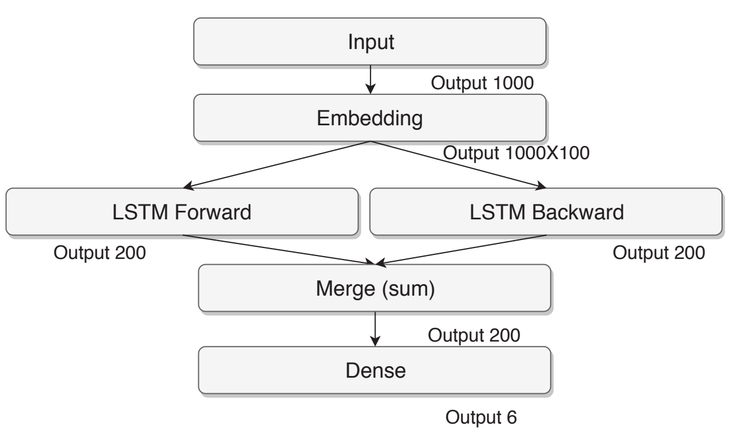




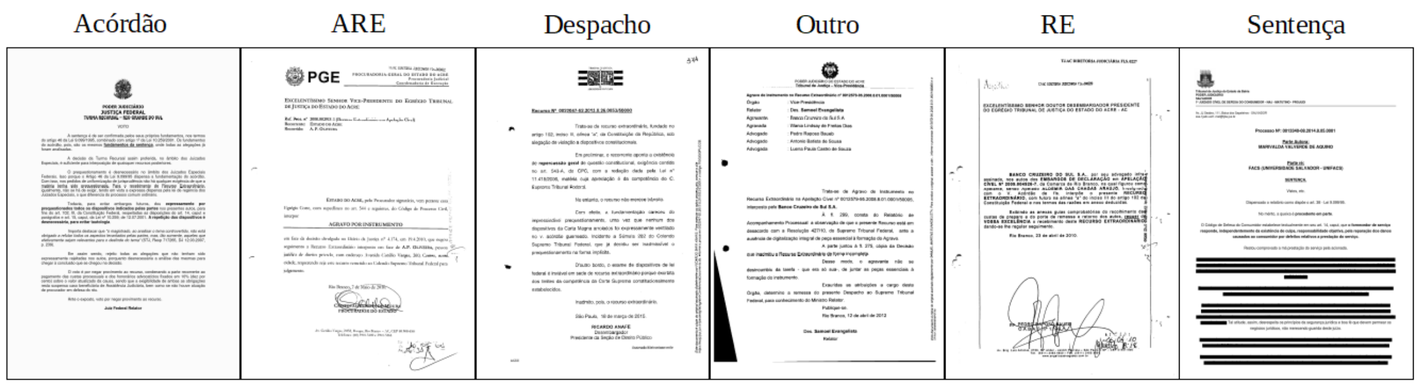
Preprocessing workflow



Network architecture

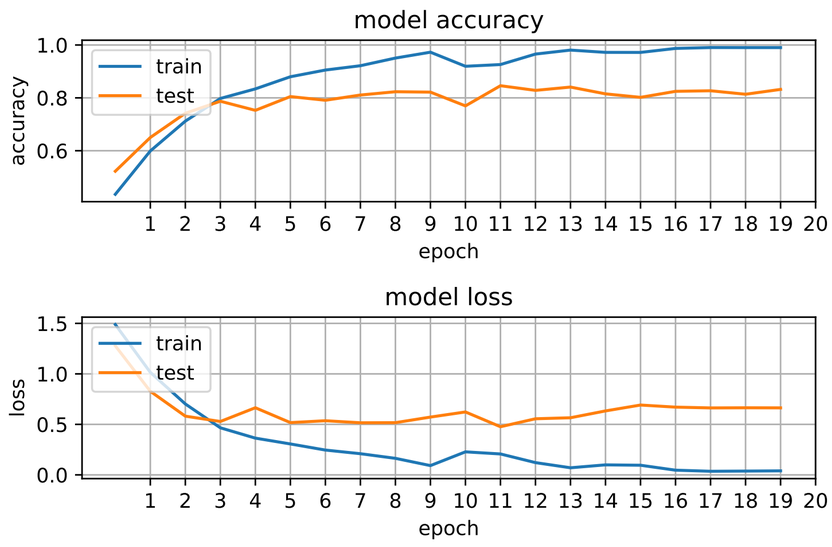


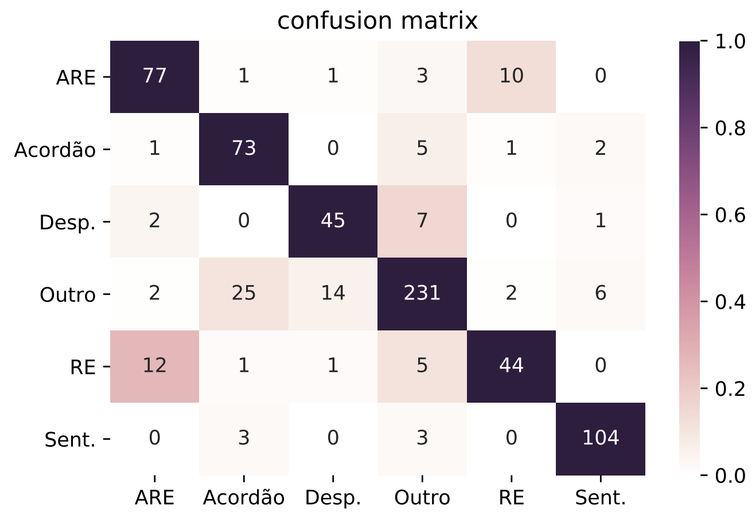
Data samples



Experiments and results

- Our Bi-LSTM model takes inputs of 1000 tokens, which covers most of the contents of one page.
- We only need to run OCR on up to two pages per document.
- The dataset split was: 70% of the samples for training, 20% for validation and 10% for test.
- Result: 84% F_1 score with no pre-processing.





Conclusions

- We proposed a tool to significantly speed up the first steps of the analysis of legal documents that reach Brazil's supreme court (STF).
- The task consists in classifying legal briefs into a set of 6 classes.
- For that we introduced a Bi-LSTM which processes the first 1000 tokens of the documents (first page).
- The model is strong enough to classify documents with an F_1 score of 84%, without using OCR on the remaining pages of the document.

Dataset reference

da Silva, N. C. et al. (2018). 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.

Acknowledgments

SUPREMO TRIBUNAL FEDERAL
CAPES
FINATEC

Contact: teodecampos@unb.br
https://cic.unb.br/~teodecampos/ViP/