FEDERAL UNIVERSITY OF SANTA CATARINA
TECHNOLOGICAL AND SCIENTIFIC CENTER
GRADUATE PROGRAM IN AUTOMATION AND SYSTEMS

Thiago Raulino Dal Pont

Classification of legal documents and prediction of indenization based on Natural Language Processing and Deep Learning

Florianópolis
2021

Thiago Raulino Dal Pont

Classification of legal documents and prediction of indenization based on Natural Language Processing and Deep Learning

Dissertation submitted to the Graduate Program in Automation and Systems at the Federal University of Santa Catarina to obtain the Master's Degree in Automation and Systems Engineering.
Supervisor:: Prof. Jomi Fred Hübner, PhD.
Co-supervisor:: Prof. Aires José Rover, PhD

Florianópolis
2021

Ficha de identificação da obra

A ficha de identificação é elaborada pelo próprio autor.

Orientações em:

http://portalbu.ufsc.br/ficha

Thiago Raulino Dal Pont

Classification of legal documents and prediction of indenization based on Natural Language Processing and Deep Learning

The present work at [master] level was evaluated and approved by an examining board composed of the following members:

Prof.(a) xxxx, Dr(a).
Instituição xxxx

Prof.(a) xxxx, Dr(a).
Instituição xxxx

Prof.(a) xxxx, Dr(a).
Instituição xxxx

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de Master's Degree in Automation and Systems Engineering.

Coordenação do Programa de
Pós-Graduação

Prof. Jomi Fred Hübner, PhD.
Supervisor:

Florianópolis, 2021.

Este trabalho é dedicado aos meus colegas de classe e
aos meus queridos pais.

ACKNOWLEDGEMENTS

Inserir os agradecimentos aos colaboradores à execução do trabalho. Xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx.

"O ótimo é
inimigo
do
bom."
(SOBRENOME do autor da epígrafe, ano)

ABSTRACT

Resumo traduzido para outros idiomas, neste caso, inglês. Segue o formato do resumo feito na língua vernácula. As palavras-chave traduzidas, versão em língua estrangeira, são colocadas abaixo do texto precedidas pela expressão “Keywords”, separadas por ponto.

Keywords: Keyword 1. Keyword 2. Keyword 3.

RESUMO

No resumo são ressaltados o objetivo da pesquisa, o método utilizado, as discussões e os resultados com destaque apenas para os pontos principais. O resumo deve ser significativo, composto de uma sequência de frases concisas, afirmativas, e não de uma enumeração de tópicos. Não deve conter citações. Deve usar o verbo na voz ativa e na terceira pessoa do singular. O texto do resumo deve ser digitado, em um único bloco, sem espaço de parágrafo. O espaçamento entre linhas é simples e o tamanho da fonte é 12. Abaixo do resumo, informar as palavras-chave (palavras ou expressões significativas retiradas do texto) ou, termos retirados de thesaurus da área. Deve conter de 150 a 500 palavras. O resumo é elaborado de acordo com a NBR 6028.

Palavras-chave: Palavra-chave 1. Palavra-chave 2. Palavra-chave 3.

LIST OF FIGURES

LIST OF FRAMES

LIST OF TABLES

CONTENTS

1 INTRODUCTION

According to the last report *Justiça em Números*, published annually by the National Council of Justice (CNJ), by the end of 2019, there was around 77,1 million ongoing processes waiting for a solution in the Brazilian Judiciary. In total, in 2019, 30,2 million lawsuits were filled in all Judiciary, an increase of 6,8% in relation to 2018. From those processes, about 5,2 million were filled in the Special Civel Courts (JECs) (CNJ, 2020).

JECS are Judiciary bodies regulated by Law no. 9,099/1995, which seek to facilitate citizens' access to Justice through simpler and cost-free procedures. As a result, JECs tend to approach the legal problems of ordinary people who find themselves involved in daily conflicts of small economic expression, whether in the purchases they make, in the services they hire or in the accidents they suffer (WATANABE, 1985).

Considering the challenges faced in the judiciary system in Brazil and around the world (CITE), there is an increasing interest (CITE) of the literature on applying Artificial Intelligence (AI) based techniques to solve issues from the lower courts (CITE), such as JECs, to the superior courts (CITE).

AI, according to Russell and Norvig (2020), relates to the study of intelligent agents that receives percepts or inputs from the environment and make actions. The agent uses a function to map the percepts to the appropriate actions and such function can be explicitly defined and/or learned using Machine Learning (ML) techniques (CITE).

TODO: Improve In the legal context, the agent receives as percept (inputs) a the case's textual data and acts (output) setting a decision. Considering the complexity of legal judgments, such decision may be learned using ML and Text Mining (TM) techniques, **TODO:** Define TM.

Regarding the tasks to apply in legal decisions include clustering, which is ..., classification..., regression, ... (?)

TM techniques have been used to solve several tasks regarding textual data, such as classification (CITE DEF), regression (CITE DEF) and clustering (CITE DEF) in several contexts, such as the legal (), medical (), engineering and social media

TODO: Continue...

1.1 PROBLEM DEFINITION

The judicial lawsuits processed at the JECs are decided manually by the judge, and there is no automation in that sense. This leads to slowness and, as a consequence, the large number of processes pending solution (CITE). In addition, these processes are composed of unstructured textual data that, in addition to being represented in natural language, have their own legal vocabulary (CITE).

1.2 RESEARCH QUESTION

"Is it possible to predict the result of a legal case based on its content and predict the amount of compensation for immaterial damage using machine learning and text mining techniques?"

The question can be broke down into three:

- How to translate the complexity of the legal language to a numerical representation?
- Which machine learning techniques for classification can bring an legally acceptable accuracy to the predicted lawsuit result?
- Which machine learning techniques for regression can bring an legally acceptable error to the predicted amount of compensation for immaterial damage?

1.3 OBJECTIVES

In this section we introduce to the reader the main objective and the specific objectives necessary to achieve it.

1.3.1 Main Objective

To evaluate if we can predict with a legally acceptable amount of accuracy the result of a legal case and predict the amount of compensation using machine learning and text mining techniques.

1.3.2 Specific Objectives

To do that we need to:

- Demonstrate that representing the legal cases numerically using word embeddings and BOW can achieve legally acceptable results in the classification and regression tasks.
- Demonstrate that it is possible to predict the lawsuit result using classical and deep machine learning techniques for classification with legally acceptable accuracy.
- Demonstrate that it is possible to predict the amount of compensation using classical and deep machine learning techniques for regression with legally acceptable error.

1.4 JUSTIFICATION AND SUBJECT RELEVANCE

- RSL Results

- AI Regulations and initiatives
- Growth in interest

1.5 METHODOLOGICAL PROCEDURES

- Systematic Review
- Legal Dataset
- Experiments using Python
- Evaluation
- Check Legal Acceptable Accuracy and Errors
- Resources

1.6 CONTRIBUTIONS

During the research, experiments were conducted to answer the research questions. Some of these experiments resulted in publications in journals and conferences. Considering the published works and the experiments applied, the contributions of these work are as follows:

- Pre-trained word embeddings models for Brazilian legal texts, since there was no available representation available before.
- Impact of adjustments in the pipeline for regression and classification.
- As real life application, we would help the Judiciary by helping to end the lawsuits in JEC at the conciliation hearing step.

1.7 DOCUMENT ORGANIZATION

The work is structured in X chapters, beginning with this introduction.

In Chapter **??**, we introduce...

In Chapter **??**, ...

REFERÊNCIAS

CNJ. **Justiça em Números 2020**. Ed. by CNJ. Brasília: CNJ, 2020. P. 236.

RUSSELL, Stuart; NORVIG, Peter. **Artificial Intelligence: A Modern Approach**. [S.l.]: Pearson, 2020. ISBN 978-0-13-604259-4.

WATANABE, Kazuo. In: WATANABE, Kazuo (Ed.). **Juizado Especial de Pequenas Causas: lei n. 7.244/1984**. São Paulo: Revista dos Tribunais, 1985. Filosofia e características básicas do Juizado Especial de Pequenas Causas.

APPENDIX A – DETAILS FROM SYSTEMATIC REVIEWS OF THE LITERATURE

A.1 TEXT REPRESENTATION

A.2 TEXT CLASSIFICATION

A.3 TEXT REGRESSION