



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.

[illegible]

*“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

Figure 1 – Pipeline for supervised Learning in Texts . . . . .	24
Figure 2 – Classification accuracy for the dataset with cases' result . . . . .	27
Figure 3 – Classification Accuracy for the dataset without cases' results . . . . .	28
Figure 4 – CNN Model for Text Classification (KIM, 2014) . . . . .	30
Figure 5 – Word Embeddings Projection . . . . .	31
Figure 6 – Accuracy for test set from CNN model . . . . .	32
Figure 7 – Macro F1-Score for test set from CNN model . . . . .	33
Figure 8 – Publications on ML and TM applied to the legal domain without year filtering . . . . .	44
Figure 9 – Researches by year for text representation in legal documents . . . . .	50
Figure 10 – Researches by year for text representation in Portuguese documents . . . . .	51

## LIST OF FRAMES

## LIST OF TABLES

Table 1 – Acquired process from Courts for Embeddings Training . . . . .	22
Table 2 – Global context corpora . . . . .	23
Table 3 – Hyperparameters for classification techniques . . . . .	26
Table 4 – Classification accuracy for the dataset with cases’ result . . . . .	26
Table 5 – Classification accuracy for the dataset without cases’ result . . . . .	27
Table 6 – Ten most frequent authors . . . . .	45
Table 7 – Ten most frequent journals and conferences . . . . .	45
Table 8 – Ten most frequent pre-processing techniques . . . . .	46
Table 9 – Ten most frequent representation techniques . . . . .	46
Table 10 – Ten most frequent classification techniques . . . . .	47
Table 11 – Most frequent clustering techniques . . . . .	47
Table 12 – Ten most frequent Evaluation Metrics . . . . .	47
Table 13 – Representations applied to legal texts . . . . .	50
Table 14 – Representation used in Portuguese texts . . . . .	51

## CONTENTS

<b>1</b>	<b>INTRODUCTION . . . . .</b>	<b>15</b>
1.1	PROBLEM DEFINITION . . . . .	16
1.2	RESEARCH QUESTION . . . . .	17
1.3	OBJECTIVES . . . . .	17
1.3.1	<b>Main Objective . . . . .</b>	<b>17</b>
1.3.2	<b>Specific Objectives . . . . .</b>	<b>17</b>
1.4	JUSTIFICATION AND SUBJECT RELEVANCE . . . . .	18
1.5	RESEARCH METHOD AND DELIMITATION . . . . .	19
1.6	DOCUMENT ORGANIZATION . . . . .	19
<b>2</b>	<b>MACHINE LEARNING FOR TEXT . . . . .</b>	<b>20</b>
2.1	MACHINE LEARNING (DEFINIÇÃO) . . . . .	20
2.2	TEXT MINING AND NATURAL LANGUAGE PROCESSING . . . . .	20
2.3	TEXT PRE-PROCESSING . . . . .	20
2.4	TEXT REPRESENTATION . . . . .	20
2.4.1	<b>Bag of Words . . . . .</b>	<b>20</b>
2.4.2	<b>Word Embeddings . . . . .</b>	<b>20</b>
2.4.3	<b>Recent techniques (n usadas; BERT, GPT-3, ...) . . . . .</b>	<b>20</b>
2.5	CLASSIFICATION . . . . .	20
2.5.1	<b>Definition . . . . .</b>	<b>20</b>
2.5.2	<b>Techniques . . . . .</b>	<b>20</b>
2.5.3	<b>Evaluation Methods . . . . .</b>	<b>20</b>
2.6	REGRESSION . . . . .	20
2.6.1	<b>Definition . . . . .</b>	<b>20</b>
2.6.2	<b>Techniques . . . . .</b>	<b>20</b>
2.6.3	<b>Evaluation Methods . . . . .</b>	<b>20</b>
2.7	APPLICATIONS OF MACHINE LEARNING IN TEXT . . . . .	20
2.8	CONCLUSIONS FOR THE CHAPTER . . . . .	20
<b>3</b>	<b>RELATED WORK . . . . .</b>	<b>21</b>
3.1	RELATED WORKS IN TEXT REPRESENTATION . . . . .	21
3.2	RELATED WORKS IN TEXT CLASSIFICATION . . . . .	21
3.3	RELATED WORKS IN TEXT REGRESSION . . . . .	21
3.4	CONCLUSIONS FOR THE CHAPTER . . . . .	21
<b>4</b>	<b>EXPERIMENTS, RESULTS AND DISCUSSION . . . . .</b>	<b>22</b>
4.1	DATASET CONSTRUCTION . . . . .	22
4.1.1	<b>Labeled Dataset from Special Civil Court . . . . .</b>	<b>22</b>
4.1.2	<b>Unlabeled Dataset from Brazilian Higher Courts . . . . .</b>	<b>22</b>
4.2	TEXT CLASSIFICATION IN LEGAL JUDGMENTS . . . . .	23

4.2.1	Pipeline for classification . . . . .	23
4.2.2	Results and Discussion for classification experiments . . . . .	25
4.2.3	Conclusions from the section . . . . .	28
4.3	TEXT REPRESENTATION IN LEGAL JUDGMENTS . . . . .	29
4.3.1	Corpus Processing . . . . .	29
4.3.2	Embeddings Training . . . . .	29
4.3.3	Embeddings Evaluation in Legal Text Classification . . . . .	29
4.3.4	Experimental Results . . . . .	31
4.3.5	Discussion from Context Perspective . . . . .	32
4.3.6	Discussion from Corpus Size Perspective . . . . .	34
4.4	TEXT REGRESSION IN LEGAL JUDGMENTS . . . . .	34
4.5	CONCLUSIONS FOR THE CHAPTER . . . . .	34
5	<b>FINAL REMARKS . . . . .</b>	<b>35</b>
5.1	CONCLUSIONS . . . . .	35
5.2	CONTRIBUTIONS . . . . .	35
5.3	LIMITATIONS . . . . .	36
5.4	FUTURE WORK . . . . .	36
5.5	EXTRA INFORMATION . . . . .	36
5.6	ACKNOWLEDGEMENTS . . . . .	36
	<b>REFERÊNCIAS . . . . .</b>	<b>37</b>
	<b>APPENDIX A – SYSTEMATIC REVIEW OF THE LITERATURE: AI, ML AND LAW . . . . .</b>	<b>42</b>
A.1	DEFINITION OF SEARCH QUESTIONS . . . . .	42
A.2	SEARCH STRATEGIES . . . . .	42
A.3	KNOWLEDGE BASES . . . . .	42
A.4	INCLUSION AND EXCLUSION CRITERIA . . . . .	43
A.5	DATA EXTRACTION PLAN . . . . .	43
A.6	SEARCH EXECUTION AND PRELIMINARY ANALYSIS . . . . .	44
A.7	RESULTS AND ANALYSIS . . . . .	44
	<b>APPENDIX B – SYSTEMATIC REVIEW OF THE LITERATURE: TEXT REPRESENTATION . . . . .</b>	<b>48</b>
B.1	DEFINITION OF SEARCH QUESTIONS . . . . .	48
B.2	SEARCH STRATEGIES . . . . .	48
B.3	KNOWLEDGE BASES . . . . .	48
B.4	INCLUSION AND EXCLUSION CRITERIA . . . . .	49
B.5	DATA EXTRACTION PLAN . . . . .	49
B.6	SEARCH EXECUTION AND PRELIMINARY ANALYSIS . . . . .	49
B.7	RESULTS AND ANALYSIS . . . . .	50

**APPENDIX C – SYSTEMATIC REVIEW OF THE LITERATURE: TEXT**

	<b>REGRESSION . . . . .</b>	<b>52</b>
C.1	DEFINITION OF SEARCH QUESTIONS . . . . .	52
C.2	SEARCH STRATEGIES . . . . .	52
C.3	SEARCH RESOURCES . . . . .	52
C.4	SELECTION CRITERIA . . . . .	52
C.5	DATA EXTRACTION . . . . .	53
C.6	SEARCH EXECUTION AND PRELIMINARY ANALYSIS . . . . .	53

## 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 Civil Courts (JECs) (CNJ, 2020a).

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).

To address the challenges faced in the judiciary systems in Brazil and around the world (SADIKU, 2020), there is an increasing interest of the literature, on applying Machine Learning (ML) and Text Mining (TM) based techniques in the legal area (see Systematic Review of the Literature (SRL) in Appendix A). The papers try to solve a variety of legal problems, such as the information extraction from contracts (HASSAN; LE, 2020), and the prediction of decisions in lower courts and superior courts (ŞULEA et al., 2017; VIRTUCIO et al., 2018).

According to Mitchell (1997), ML focuses on constructing computer systems that learn through experience. Thus, such systems can learn to classify texts, control robots, predict weather and others (SEBASTIANI, 2002; KOBER et al., 2013; SHI et al., 2015). ML based systems can learn using some approaches (SCHMIDHUBER, 2015; CARUANA, 1997), including supervised and unsupervised. In supervised ML, the system maps a relationship between inputs and outputs based on *a priori* knowledge from labeled data. In unsupervised ML, the system tries to find patterns in the data considering the similarities among the many data points (THEODORIDIS; KOUTROUMBAS, 2009).

TM relates, according to Aggarwal and Zhai (2012), to the idea of discovering and analyzing patterns, such as trends and outliers, from textual data. TM also focus on helping users to analyze and digest information towards a better decision making. Thus, the necessity of TM and ML based applications emerges with the large amount of textual data generated by users, companies, universities and so on, and human limitations to analyze it (LECUN et al., 2015; KHAN et al., 2014).

TM and ML have been used to address challenges regarding supervised learning, such as text classification, text regression, and unsupervised learning like text clustering (AGGARWAL; ZHAI, 2012; TRUSOV et al., 2016; MEDVEDEVA et al., 2019; ZHANG; ZHOU, 2019).



## 1.1 PROBLEM DEFINITION

Considering the high level of judicialization in Brazil, where the judiciary is increasingly assigned to solve day-to-day conflict, less cases are decided with agreements between the parties. So they shall wait for a final judgment, which may take from days to years (CURY et al., 2019; MANCUSO, 2020).

If the parties had more information on the possible outcomes, based on the judge's previous decisions, they would achieve a consensus and finish the case, without the need of waiting for a judgment. In this context, the use of ML and TM can help in the conciliation between the parties by predicting the possible outcomes based on the previous decisions.

In the JECs, the cases are decided manually by the judge, and there is no automation in that sense. Thus, this leads to slowness and, as a consequence, an impact on the large number of processes pending solution (CNJ, 2020a).

Moreover, depending on the legal contexts, the judge sets an amount of compensation to be paid by a party to the other. However, there is no explicit rules to define them. Thus, the amount of compensation depends on the judge's interpretation of the case (SADIKU, 2020).

Thus, this work intends to make contributions to the state of the art of ML and TM applied to legal texts by investigating possible solutions to the prediction of legal cases outcomes. To get such achievement, three challenges are addressed. First, the legal text, its complexity, and vocabulary must be represented numerically allowing the use of ML techniques. Second, the prediction of the chances of winning or not a case. And third, the prediction of the amount of compensation when the party wins the case.

The first challenge is addressed in several languages, like Polish, English and Chinese (CHALKIDIS; KAMPAS, 2019; SMYWIŃSKI-POHL et al., 2019). However, from a SRL (see Appendix B), no work explored the representations of legal texts in the Portuguese language using techniques like word embeddings, BERT and others.

The second challenge relates to classification, a supervised ML task which tries to learn a model to map a set of records to one of a set of labels (AGGARWAL; ZHAI, 2012). Based on a SRL (see Appendix A), the classification task has been applied in many legal contexts (CITE 4). However, such work, do not (FINISH, after checking the SRL).

The third challenge can be addressed with the regression task, an supervised ML task which aims to train a model to map the relationship between an input  $x$  to a continuous output  $y$  (DRAPER; SMITH, 1998). According to a SRL (See Appendix C), to the best of our knowledge, it is not addressed in any work in the literature. Thus, there is the possibility of many contributions in this regard.

## 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:

- Which representation techniques can translate the complexity of the legal language to a numerical representation to achieve legal acceptable results in ML tasks?
- 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 whether 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

In order to fulfill the main objective, the following specific objectives must be achieved:

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

There is an increasing interest in the literature on the applications of ML and TM techniques in the legal domain. From that, the concerns about the implications of AI uses in the legal domain and others emerge (BRAZ et al., 2018; DAVIS, 2020). Worldwide, such concerns provoked the debate of regulations on AI (CATH, 2018). In 2020, the National Council of Justice (CNJ) created an ordinance to regulate the uses of AI in the Brazilian Judiciary (CNJ, 2020b). Thus, the application of ML in the legal domain has important theoretical and practical relevance.

Through the construction of three SRLs, detailed in Appendix A, B and C, we observed such increasing research interest on the application of ML and TM in the legal domain in the last years. With the first SRL, we focused on the literature regarding ML and TM in the legal domain. As a result, most of the selected work related to the predictions in legal documents in many languages. However, until the SRL search date, none of the papers address as many classification ML techniques on the same dataset as proposed in this work.

In terms of representation of legal texts, this SRL showed us some relevant work. However, to address the first part of our research question, more research was needed. To serve as a complement, the second SRL focus on the representation of legal texts, and the results showed us that many representation techniques have been explored in the legal domain, including word embeddings and Bidirectional Encoder Representations from Transformers (BERT). However, the available work involving the Portuguese language did not focus on the aspects of representations, that is, just using the existing representations as tools, until the SRL search date.

The third SRL focused on the application of regression in legal texts. However, such research did not find any work in that matter, although the use of many search keywords patterns. A broader literature search, detailed in Appendix C, showed work regarding other contexts, including Financial and Health. Thus, this work's main contributions relates to the use of regression in legal texts with the focus on predicting compensations.

In terms of theoretical relevance, this work brings contributions in terms of representation of legal texts as it trains representation techniques in Brazilian legal documents to be applied in supervised learning tasks. In terms of predicting the legal case result through classification, the amount of published work limits our contributions to the evaluation on how the classical and deep learning techniques behave when applied to cases from JECs. In terms of predicting the amount of compensation through regression, this work bring contributions on impact of several Text Mining and ML techniques in the pipeline as the performance of the ML models in such task.

In terms of practical contributions, the models and pipelines from this work can be applied in real legal conciliation hearings in the JEC at UFSC. A legal expert would

present and explain the predictions to the parties. Thus, we expect to help them on reaching an agreement without the need of waiting for a judgment. In this way, the litigation in the JEC would decrease, contributing to faster and more efficient access for citizens to justice.

## 1.5 RESEARCH METHOD AND DELIMITATION

To answer the main question and its three pieces, the researcher followed a set of steps. Such pieces share common steps.

As the first step, prior to any question definition, is the construction of a Systematic Review of the Literature to evaluate the State of the Art of the application of Machine Learning and Text Mining in the Legal Domain. Two additional SRLs complemented the first as they added specific information on the research of text representation for machine learning tasks and the application of regression in legal texts. From this step, the question and objective were set.

The second step relates to the collection of the textual datasets required to answer each question. As detailed in Chapters X, Y and Z, such datasets include an unlabeled dataset of legal cases from STF, STJ and TJ-SC and legal cases from JEC at UFSC. They also include a set of attributes extracted by a legal expert.

The third step relates to the setup and execution of machine learning experiments according to each part of the question.

The fourth step relates to the evaluation of the achieved results in each experiment.

In terms of resources, this thesis required the use of a high performance computer from the EGOV group at UFSC and a set of open source libraries for ML.

## 1.6 DOCUMENT ORGANIZATION

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

In Chapter 2, we introduce the concepts of Machine Learning applied to text and the base theory to understand the three sub-problems of this work and the techniques we apply. In Chapter 3, we show the relevant work related to each sub-problem. In Chapter 4, we show the pipelines and techniques used to answer the research questions. In Chapter ??, ... In Chapter 5...

## 2 MACHINE LEARNING FOR TEXT

### 2.1 MACHINE LEARNING (DEFINIÇÃO)

### 2.2 TEXT MINING AND NATURAL LANGUAGE PROCESSING

- Definição de Text Mining e NLP
- Como se enquadram no trabalho
- Aplicações

### 2.3 TEXT PRE-PROCESSING

Filtering, Lowercasing, Stopwords, etc.

### 2.4 TEXT REPRESENTATION

#### 2.4.1 Bag of Words

#### 2.4.2 Word Embeddings

#### 2.4.3 Recent techniques (n usadas; BERT, GPT-3, ...)

### 2.5 CLASSIFICATION

#### 2.5.1 Definition

#### 2.5.2 Techniques

#### 2.5.3 Evaluation Methods

### 2.6 REGRESSION

#### 2.6.1 Definition

#### 2.6.2 Techniques

#### 2.6.3 Evaluation Methods

### 2.7 APPLICATIONS OF MACHINE LEARNING IN TEXT

### 2.8 CONCLUSIONS FOR THE CHAPTER

### **3 RELATED WORK**

This chapter presents the related work relevant to this research. Part of the them came from the SRLs and others added as complement by the researcher.

#### **3.1 RELATED WORKS IN TEXT REPRESENTATION**

#### **3.2 RELATED WORKS IN TEXT CLASSIFICATION**

#### **3.3 RELATED WORKS IN TEXT REGRESSION**

#### **3.4 CONCLUSIONS FOR THE CHAPTER**

## 4 EXPERIMENTS, RESULTS AND DISCUSSION

### 4.1 DATASET CONSTRUCTION

#### 4.1.1 Labeled Dataset from Special Civil Court

#### 4.1.2 Unlabeled Dataset from Brazilian Higher Courts

Concerning embeddings training, the first step is to obtain the collection of legal documents from the court web portals, followed by raw text extraction from these documents. To enable us to evaluate the specificity influence of these legal corpora, we divided it into two contexts: related to general legal texts and related to air transport services text.

We also collected texts from other general topics (not related to legal domains) that are already compiled and freely available. Having the corpora for legal and miscellaneous contexts, we applied some processing steps to remove noise from texts. To evaluate the influence of corpus size in embeddings training, we divided these three corpora into smaller pieces based on word count.

To train the embeddings it is required large text corpora to be able to get good embeddings. However, in the Brazilian Portuguese language, we could not find any dataset available on the Internet containing enough legal text corpora for our purposes. Thus, we had to build our legal corpora.

Our main sources of legal text are Brazilian courts platforms. We collected judgments from the webpages of Federal Supreme Court (STF), Superior Court of Justice (STJ) and State Court of Santa Catarina (TJ-SC) (STF, 2020; STJ, 2020; TJSC, 2020). We also collected judgments from the JusBrasil portal containing processes related only to failures on air transport service from all State Courts (TJ) from Brazil (JUSBRASIL, 2020).

Table 1 shows the number of processes acquired and word count for each Tribunal:

Table 1 – Acquired process from Courts for Embeddings Training

Source	Collegial Judgments	Individual Judgments	Subtotal	Word Count
STF	64,779	118,910	183,689	294,937,185
STJ	101,141	0	101,141	312,687,450
TJ-SC	989,964	662,535	1,652,499	3,060,212,814
TJs (JusBrasil)	34,239	0	34,239	78,138,337
<b>TOTAL</b>			1,971,568	3,745,975,786

Source: Adapted from Dal Pont et al. (2020)

After downloading all processes, most of them in PDF and Rich Text Format (RTF) formats, we extracted raw texts from these files. We did not apply Optical Character Recognition (OCR) in scanned PDF documents, due to time limits to finish the experiments, so only digital PDFs were accounted with RTF files in Table 1.

With the extracted texts, we applied some pre-processing steps, as discussed further in this section.

Then we built the legal text corpora containing all the processes related to all law subjects, which we call *general* legal text corpora in this work. Using this base, we created another text corpora whose processes are related only to air transport and consumer law, and we call it *air transport* legal text corpora.

To be able to compare how good embeddings trained with legal texts perform against those created with all kinds of texts, we also created other corpora from a variety of sources. Thus, we searched for free available textual datasets. In this work, we call these texts as *global* context texts. Table 2 shows all the global text datasets used. Then we apply some preprocessing steps, as will be described further in this section.

Table 2 – Global context corpora

Dataset	Documents	Word Count	Source
Wikipedia in Portuguese	1,014,713	303,622,360	Wikipedia (2019)
Brazilian Literature Books	169	37,848,783	Tatman (2017)
Old Newspapers	617,627	26,441,581	Tan (2020)
Folha de São Paulo News	165,641	74,594,367	Marlessonn (2019)
HC News Corpus	494,128	27,170,063	Christensen (2016)
Blogspot Posts	2,181,073	696,657,915	Santos et al. (2018)
Wikihow Instructions	786,283	22,471,312	Chocron and Pareti (2018)
<b>TOTAL</b>	<b>5,259,634</b>	<b>1,188,806,381</b>	

Source: Adapted from Dal Pont et al. (2020)

## 4.2 TEXT CLASSIFICATION IN LEGAL JUDGMENTS

This sections presents the results from the classification experiments involving JEC documents.

### 4.2.1 Pipeline for classification

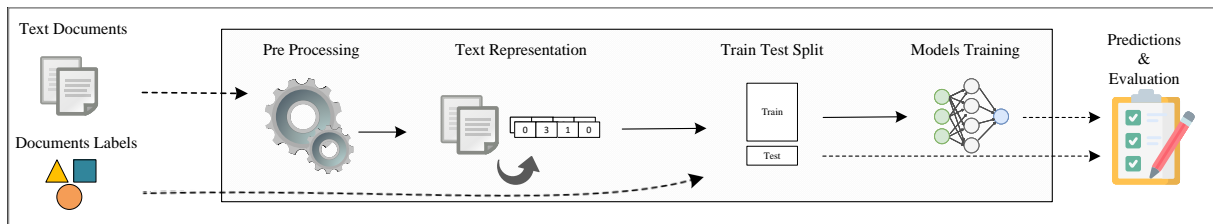
In the experiments with classification of the JEC legal cases, we applied the open-source software Orange 3. Such tool aims at offering a variety of ML and TM



techniques to the user in a simple way, without the need of any programming language. The classification in Orange3 followed the pipeline from Figure 1.

Foram realizados dois experimentos, sendo um com o texto integral da sentença e outro com a retirada da parte dispositiva, isto é, do texto que representa o resultado processual. Essa estratégia foi adotada para possibilitar uma comparação do desempenho das técnicas onde a classe é indicada no corpus textual e onde a classe não é indicada no corpus textual.

Figure 1 – Pipeline for supervised Learning in Texts



The pipeline from Figure 1 represents a simple way to apply ML techniques to texts. It receives two types of input: the texts from legal documents, in a plain text format, and its labels, that is, the cases' results.

In the case of study, the first step in the pipeline is the data pre-processing, consisting on a set of techniques to prepare the data to serve as input to an algorithm (GARCÍA; LUENGO; HERRERA, 2015). Considering the available techniques for pre-processing data described in Chapter 2 and literature in Appendix A, in this work, the following techniques were applied:

- **Transformation:** the conversion to lower case to standardize the spelling of words.
- **Tokenization:** the application of regular expression (`\w+`) to detect the pieces of texts, while removing spaces, symbols, and punctuation.
- **Stemming:** To reduce the variability of similar words, we applied Porter Stemmer (CITE), a simple and efficient stemming algorithm to the Portuguese language. However, it may make errors such as contracting the word *morais* to *morai*, instead of *moral* (Portuguese words for singular and plural of *moral*, respectively).
- **Filtering:** Removing stopwords, such as prepositions and articles to keep only the meaningful words. Orange makes available a list of stopwords in the Portuguese language that include words like *de* (of/from), *para* (for), *alguns* (some).

The next step in the pipeline relates to the extraction of N-Grams which detects sequences of two or more words that appear together consistently in the text. Examples of such sequences include *text mining*, *air transport* and others. In this research, the

limit of the length of N-grams was two. Bigger numbers of N-grams would lead to unreasonably large textual representations.

Nas etapas de stemização e filtro foi necessário a especificação de um dicionário da língua portuguesa. Essas etapas de pré-processamento têm como resultado a geração de uma bag of words (bolsa de palavras), na qual cada documento é representado como um vetor de palavras que ocorrem no documento. Nela é efetuada a contagem dos termos e o cálculo da frequência em que cada termo aparece no documento (MATSUBARA; MARTINS; MONARD, 2003).

After N-Grams Extraction, the numerical representations of the documents are created using the algorithm Bag of Words (BOW). In this work, we used the Term Frequency to set the values of the BOW model.

Having the representation of the texts, the next step consists on dividing the dataset in two subsets: train and test sets. As described in Chapter 2 (CHECK), among the methods to do so, there is the cross validation, which splits the dataset in  $k$  folds, where  $k-1$  are used for training and 1 for testing the models. In  $k$  steps, cross-validation alternates folds in such a way that each of them is used to train and test the ML models. In this research,  $k$  was set to 10, that is, 90% of the dataset used for training and 10% for testing the models.

The next step consists on training the models to predict the result of the cases from JEC according to the possible outcomes, as described in Section 4.1.

In this research, we tested the following techniques: k-Nearest Neighbors (kNN), Support Vector Machines (SVM), Random Forest (RF), feed-forward Neural Networks (NN), Naïve Bayes (NB) and Logistic Regression (LR). Table 3 presents the hyperparameters applied to each technique, based on the values suggested by Orange 3.

After the ten steps of the cross validation, there is the evaluation step, which takes as input the trained models and the test set, and evaluate how well the models perform on classifying unseen legal cases. To estimate the performance, we use the Accuracy, detailed in 2.

The following section presents the results for the two experiments using the pipeline from Figure 1.

#### 4.2.2 Results and Discussion for classification experiments

The results obtained after applying to the pipeline from Figure 1 are presented in the following paragraphs.

Table 4 shows the accuracy obtained for each algorithm on each label using the full cases' text.

On the one hand, there is the highest performance of LR, followed by NN, for the *Well Founded*, *Partially Founded* and *Not founded* labels. On the other hand, there is a significantly lower performance from the SVM for the *Well Founded* and *Partially*

Table 3 – Hyperparameters for classification techniques

Technique	Hyper-parameters
kNN	Number of Neighbors: 4; Distance Metric: Euclidean; Weight: Uniform;
LR	Regularization type: Ridge (L2); C (strength): 1
NB	–
NN	Hidden Layers: 2 Neurons in each layer: 100, 50; Activation Function: tanh; Solver: Stochastic Gradient descent (SGD)
RF	Number of Trees: 10; Minimum subset size: 5
SVM	C (cost): 1.0; $\epsilon$ (Regression loss): 0.1; Kernel: Radial Basis Function (RBF); Iteration Limit: 100

Table 4 – Classification accuracy for the dataset with cases' result

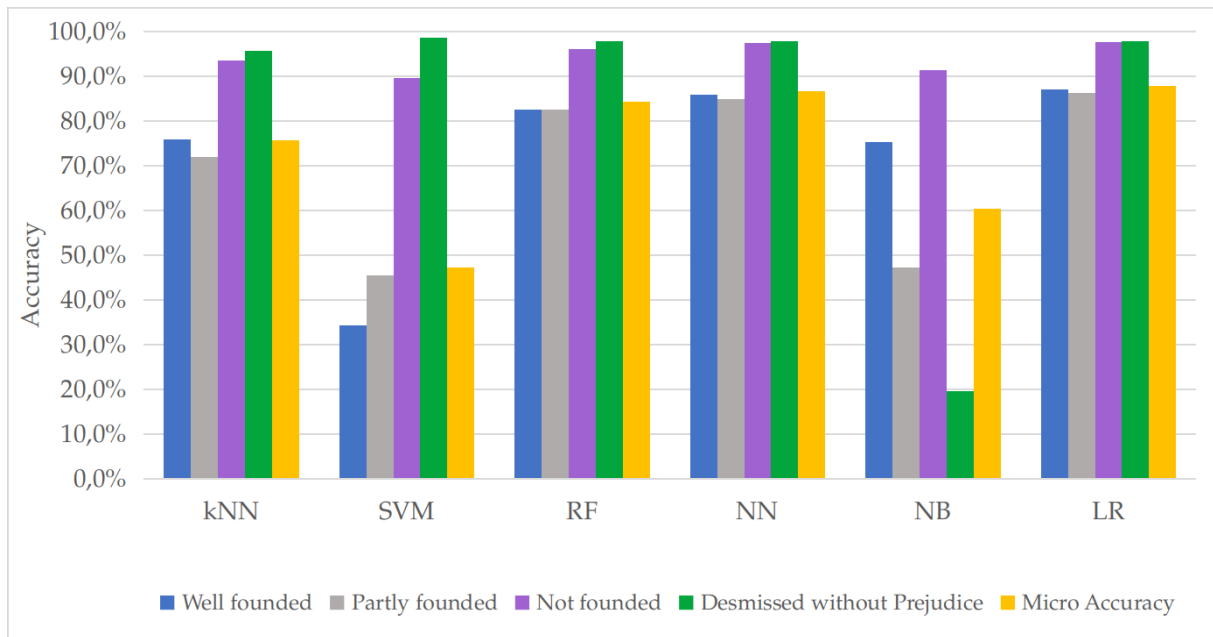
Technique	Well Founded	Partly founded	Not founded	Dismissed without prejudice	Micro Accuracy
kNN	75,9%	71,9%	93,5%	95,7%	75,8%
SVM	34,3%	45,5%	89,6%	98,5%	47,3%
RF	82,6%	82,6%	96,1%	97,9%	84,2%
NN	85,9%	84,9%	97,5%	97,9%	86,7%
NB	75,2%	47,3%	91,4%	19,6%	60,3%
LR	87,1%	86,2%	97,6%	97,9%	87,8%

Source: Adapted from Sabo et al. (2019)

*Founded* labels, while in NB for the *Partially Founded* label. However, for the *Dismissed without prejudice* label, whose sample is smaller, the SVM achieved the highest performance, while the Naïve Bayes significantly achieved the lowest. The accuracy achieved by each classifier can be better compared in Figure 2.

Table 5 contains the accuracy achieved by each algorithm in each of the four labels using the text of the cases without result of the process.

Figure 2 – Classification accuracy for the dataset with cases' result



Source: Adapted from Sabo et al. (2019)

Table 5 – Classification accuracy for the dataset without cases' result

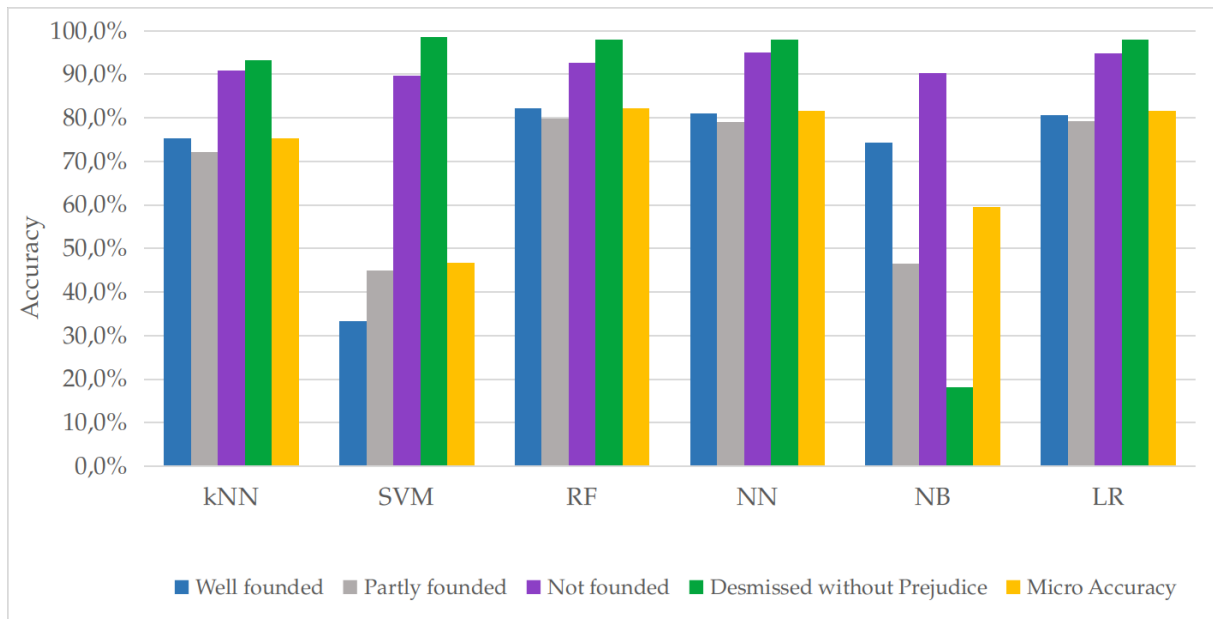
Technique	Well founded	Partly founded	Not founded	Dismissed without prejudice	Micro Accuracy
kNN	75,2%	72,1%	90,9%	93,2%	75,4%
LR	80,7%	79,2%	94,9%	97,9%	81,6%
NB	74,3%	46,5%	90,2%	18,1%	59,5%
NN	81,0%	79,0%	95,1%	97,9%	81,6%
RF	82,2%	79,9%	92,7%	97,9%	82,2%
SVM	33,4%	45,0%	89,6%	98,5%	46,7%

Source: Adapted from Sabo et al. (2019)

On the one hand, there is the highest performance of RF, followed by NN and LR, for the *Well founded*, *Partially founded* and *Not founded* classes. On the other hand, there is a significantly lower performance of the SVM for the *Well founded* and *Partially founded* labels, and NB for the *Partially founded* labels. However, for the *Dismissed without prejudice* label, whose sample is smaller, again the SVM achieved the highest performance, while the NB obtained a significantly lower result. The accuracy achieved by each classifier can be better compared in Figure 3.

Table 3 shows the difference in accuracy between the experiments, indicating an average of the accuracy obtained by each classifier in the four labels.

Figure 3 – Classification Accuracy for the dataset without cases' results



Source: Adapted from Sabo et al. (2019)

In general, it is inferred from the averages that the accuracy in the experiment with the removal of the cases' result (part of the text indicating the label to which it belongs) suffered a minimal reduction, which demonstrates that the classifiers only maintained their performance with the text in which the facts narrated by the parties to the process are reported and the legal grounds applicable to the case. This becomes a prerequisite for carrying out experiments with texts from legal proceedings in which there is still no sentence, that is, in which the judge has not yet decided the result.

It is also observed that the greatest difference was obtained with the LR for the *partly founded* label, whose sample is larger, while the difference obtained by the kNN for the same class was negative, which means that this classifier performed better on the text of cases without the result.

#### 4.2.3 Conclusions from the section

The experiments from this section dealt with initial experiments carried out in the sentences of the JEC/UFSC, indicating only four possible classes to which the texts belong. In general, it was possible to obtain an overview of how the several ML techniques behave in the face of legal texts (specific on Consumer Law and failure in air transport service), evaluating which classification models reached higher and lower accuracy.

### 4.3 TEXT REPRESENTATION IN LEGAL JUDGMENTS

#### 4.3.1 Corpus Processing

After text extraction from the documents, we applied some pre-processing steps, which are required before training the embeddings or text classification. The first of them was the conversion to lower case. Then punctuation marks were removed, as well as special characters and some symbol characters. We removed stopwords neither apply stemming or lemmatization, following the literature (MIKOLOV et al., 2013; PENNINGTON et al., 2014).

In relation to our three corpora used in embeddings training, which comprising 3.7 billion, 100 million and 1.19 billion words for *general*, *air transport* and *global* corpora, respectively, we created others based on them according to the following smaller corpora sizes, considering the word count: 1,000; 10,000; 50,000; 100,000; 200,000; 500,000; 1,000,000; 5,000,000; 10,000,000; 25,000,000; 100,000,000; 500,000,000; 750,000,000 and 1,000,000,000.

We choose these corpora sizes to be able to compare the variation on evaluation metrics while increasing corpora size. For the air transport context, we could not embrace all these sizes due to limited corpora available. The largest sub-base had 100 million words for this context.

Finally, each of these smaller corpora was used to train one different word embeddings representation.

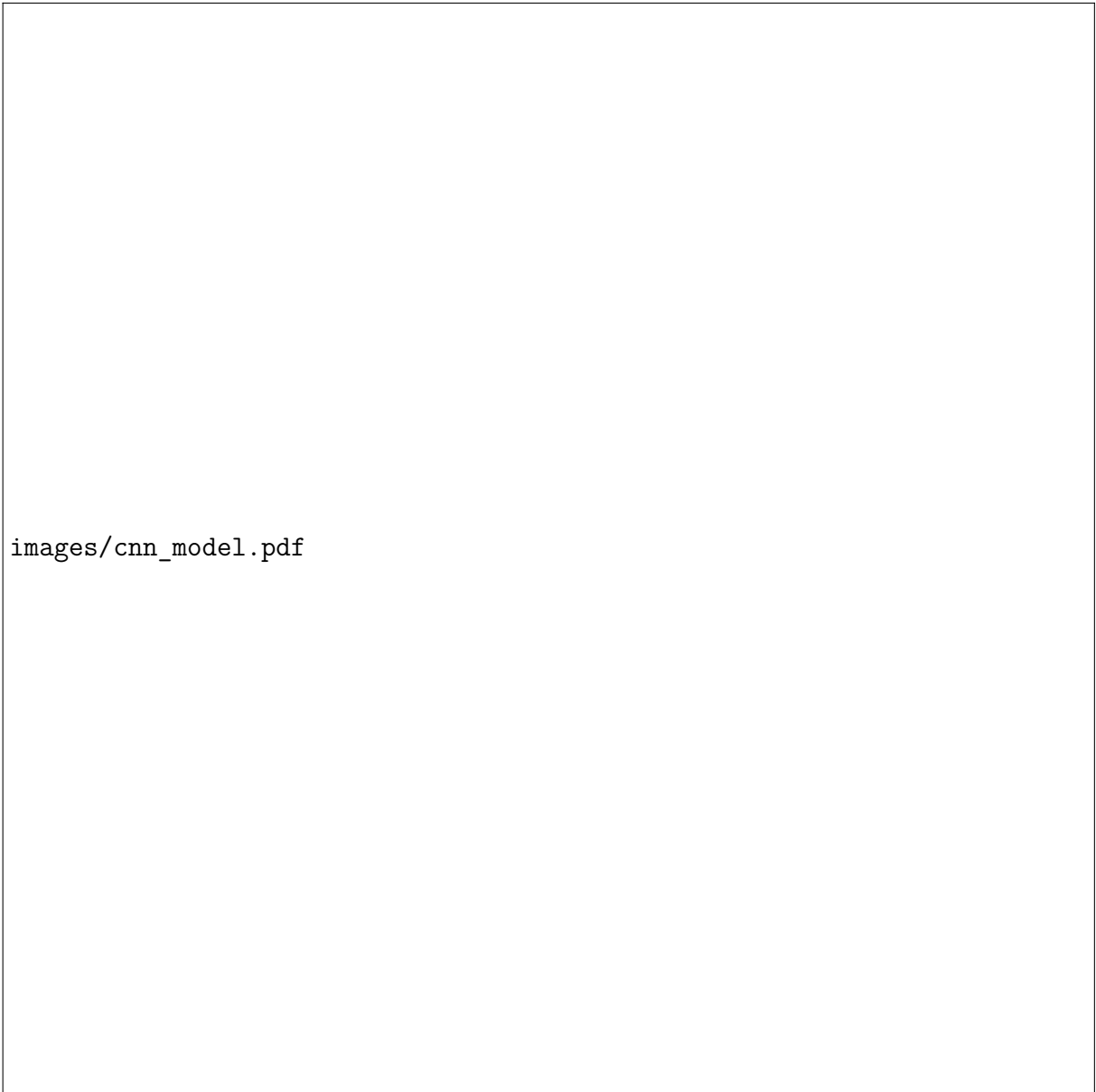
#### 4.3.2 Embeddings Training

In this work, we chose GloVe representation due to its good results in many NLP tasks, including text classification, and also for its training time which is significantly less than other techniques like Word2Vec and FastText (PENNINGTON et al., 2014). In terms of GloVe parameters, we kept most of the default values, except for windows size, training iterations, and vector size, which were set to 5, 100, and 100, respectively. With these values, we achieved better results in text classification.

Considering the corpus sizes described in Section ?? and the parameters above described, we trained 15 representations for *general* and *global* contexts bases. For *air transport* context base, we trained 11 embeddings.

#### 4.3.3 Embeddings Evaluation in Legal Text Classification

To evaluate the GloVe embeddings representations, we applied each of them to the task of text classification on judgments from JEC/UFSC. Also, we used Convolutional Neural Networks as a classification model based on the literature (KIM, 2014). Fig. 4 illustrates this model.



images/cnn\_model.pdf

Figure 4 – CNN Model for Text Classification (KIM, 2014)

This CNN takes into account the order of the words by stacking the corresponding embeddings for each word as they occur in the text. Then it applies multiple convolutional masks with different dimensions that correspond to the red and yellow contours in Fig. 4. Mask widths are equal to word embedding size while the heights can vary. In this context, mask height can be related to the idea of N-Grams, since they embrace multiple embeddings at the same time. In the original model, these heights were set to 3, 4, and 5. We added one more mask of height 2, which increased classification metrics. Also, we set to 10 the number of masks for each of these sizes, without affecting our results, but decreasing the required training time.

In this work, we applied each of the embeddings trained in conjunction with the CNN described in the classification of JEC/UFSC judgments, where Out of Vocabulary

(OOV) words are replaced by a vector of random values. Thus, we trained and tested 41 models. Furthermore, due to the stochastic nature of neural networks training methods (COHEN, 1995), each of these models was trained and tested 200 times and the resulting evaluation metrics were averaged.

Finally, we compare the performance in classification using Accuracy and Macro F1-Score.

#### 4.3.4 Experimental Results

Following the steps presented in section ??, we trained all 41 word embeddings representations for GloVe.

To illustrate how these embeddings behave, in Fig. 5, we used Principal Component Analysis (PCA) to create a projection in two dimensions of a set of words from *general* context embedding trained with 1 billion words.



Figure 5 – Word Embeddings Projection



Using each embedding, we trained and tested CNNs for text classification in JEC/UFSC judgments. These two steps were repeated 200 times, and the evaluation metrics were averaged for each group of repetitions.

In Fig. 6 and 7, we present the results, for accuracy and F1-Score, respectively, from test data applied to each CNN model. These results are related to embeddings trained with *general*, *air transport*, *global* texts. The x-axis denotes the corpus sizes used to train the embeddings, while the y-axis represents accuracy or F1-Score. Each data point represents the average of the evaluation metric, after 200 train and test repetitions using each specific embedding.

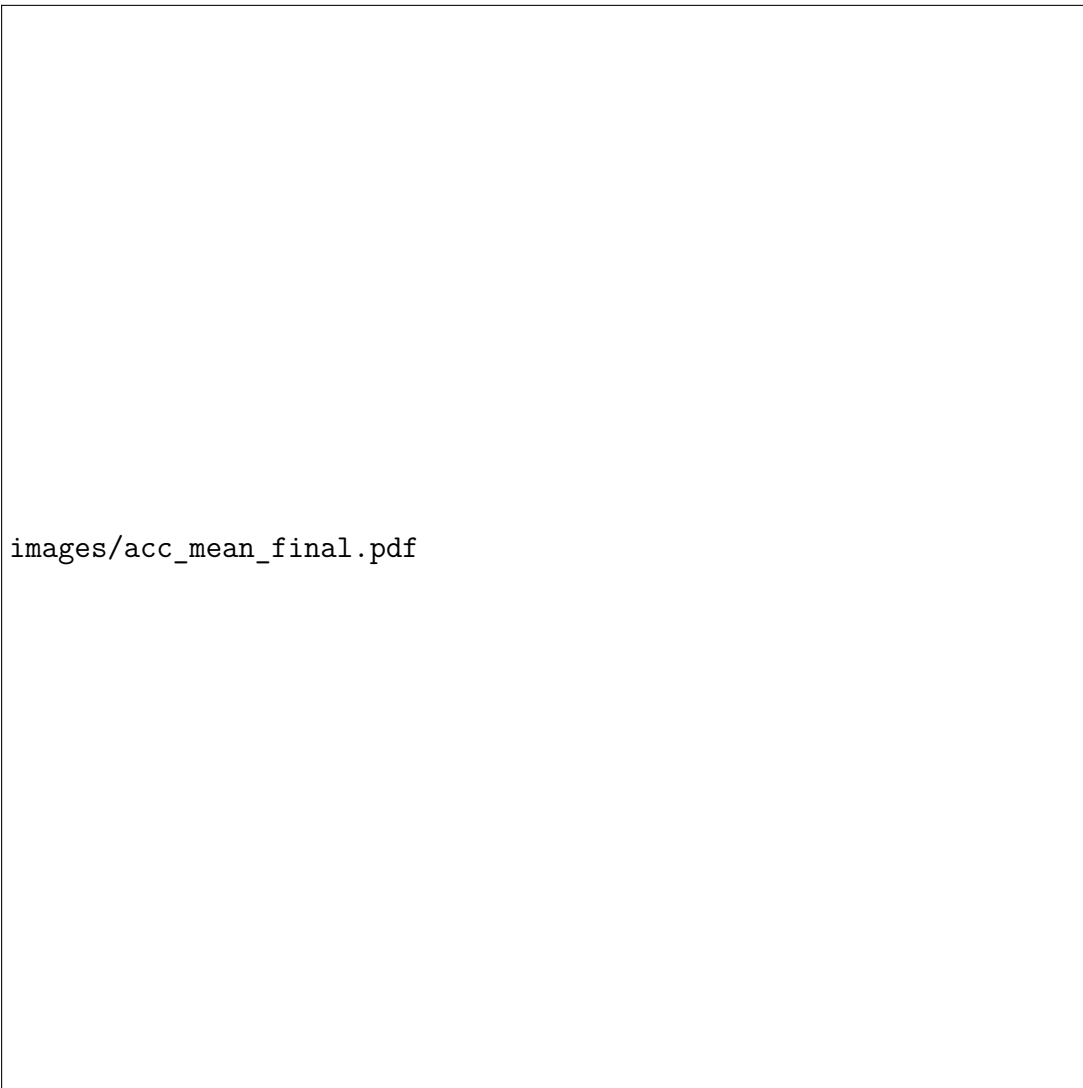
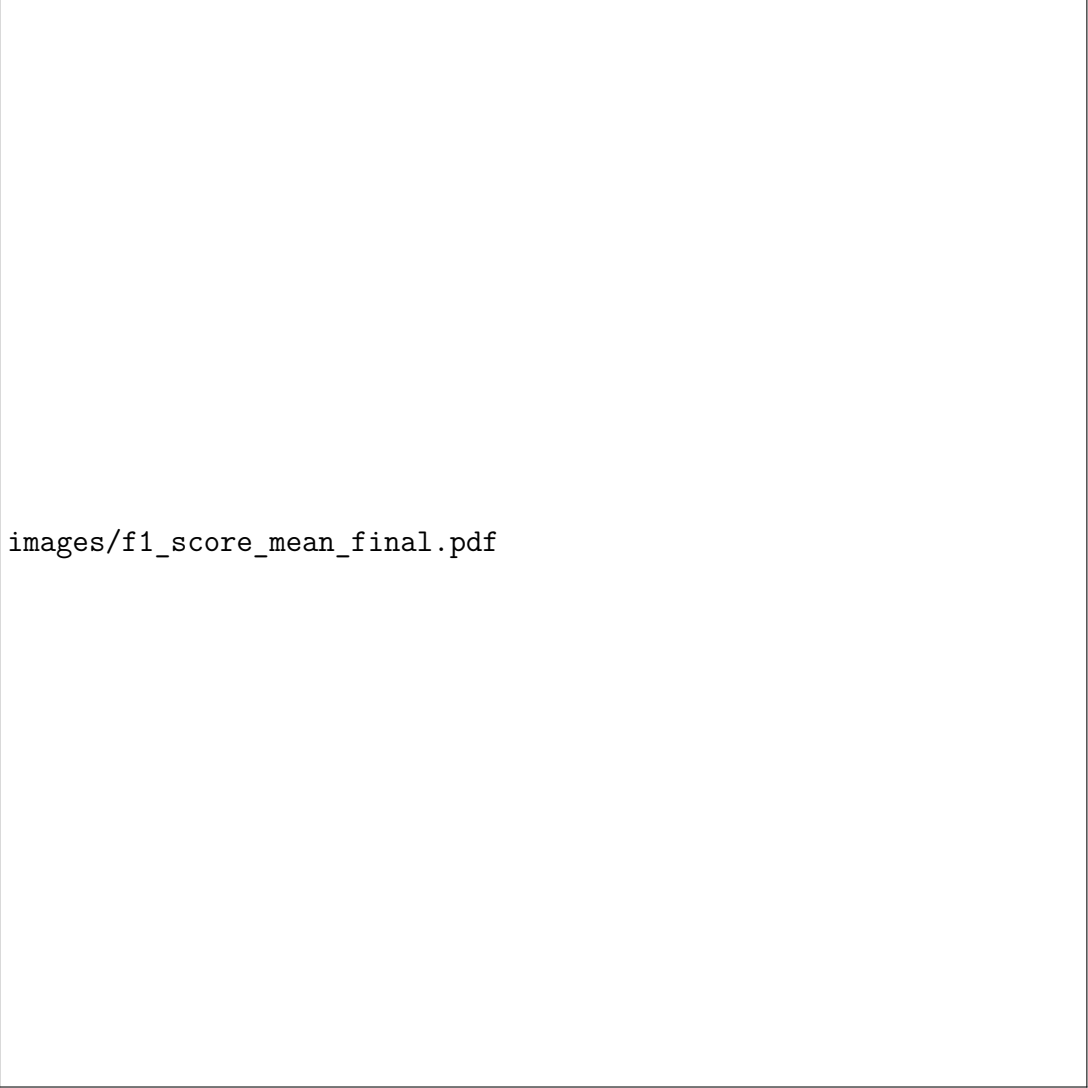


Figure 6 – Accuracy for test set from CNN model

#### 4.3.5 Discussion from Context Perspective

In this section, we will consider the first part of our research question: Does the specificity of the corpora in embeddings training contribute to the quality of the classification?



images/f1\_score\_mean\_final.pdf

Figure 7 – Macro F1-Score for test set from CNN model

In terms of accuracy, when we compare *global* against others (Fig. 6), we have that higher text specificity leads to better results, for most of the corpus sizes used for embeddings training. Furthermore, when comparing *general* and *air transport* curves, there is a significant difference in accuracy only for the lowest and highest x-values. However, in terms of F1-Score, as shown in Fig. 7, our observations change, once *general* and *air transport* curves have a similar shape. Also, for the highest corpus sizes, *general* and *global* curves converge to similar values of F1-Score. We believe that these differences in accuracy and F1-Score emerge from the fact that our dataset to text classification is imbalanced, once the former does not take this fact into account, while the latter does. However, this result still requires further investigation. In general, we can note that for smaller corpora size for embeddings training, text specificity has a more impact than for large sizes.

### 4.3.6 Discussion from Corpus Size Perspective

In this section, we will consider the second part of our research question: How does the corpus size contribute to the embedding quality?

When we observe both accuracy and F1-Score measures from Fig. 6 and 7, it is clear the tendency for improvement while increasing corpus size. However, the metrics converge with the largest corpus sizes. There are two exceptions. The first one occurs with smaller values of corpus sizes for *global* curve, as it decreases in F1-Score measures. The second corresponds to the last data point in *air transport* curves. The former can happen when the classifier performs poorly for some classes while gets better in others. The latter may indicate that those curves could improve if we had more significant corpus sizes related to that context.

In general, we can note that the greater the corpus size in embeddings training, the better are the results. However, this impact decreases as the corpus size increases until a point where more words in the corpus have little impact on the results.

## 4.4 TEXT REGRESSION IN LEGAL JUDGMENTS

## 4.5 CONCLUSIONS FOR THE CHAPTER

## 5 FINAL REMARKS

### 5.1 CONCLUSIONS

### 5.2 CONTRIBUTIONS

During the research, experiments were conducted to answer the research questions. Some of these experiments resulted in publications in journals and conferences, as follows:

SABO, Isabela Cristina; DAL PONT, Thiago Raulino; ROVER, Aires José; HÜBNER, Jomi Fred. Classificação de sentenças de Juizado Especial Cível utilizando aprendizado de máquina. **Revista Democracia Digital e Governo Eletrônico**, v. 1, n. 18, p. 94–106, 2019.

DAL PONT, Thiago Raulino; SABO, Isabela Cristina; HÜBNER, Jomi Fred; ROVER, Aires José. Impact of Text Specificity and Size on Word Embeddings Performance: An Empirical Evaluation in Brazilian Legal Domain. In: CERRI, Ricardo; PRATI, Ronaldo C. (Eds.). Cham: Springer International Publishing, 2020. v. 12319. (Lecture Notes in Computer Science). P. 521–535. DOI:10.1007/978-3-030-61377-8\_36.

SABO, Isabela Cristina; DAL PONT, Thiago Raulino; WILTON, Pablo Ernesto Vigneaux; ROVER, Aires José; HÜBNER, Jomi Fred. Clustering of Brazilian legal judgments about failures in air transport service: an evaluation of different approaches. **Artificial Intelligence and Law**, Springer Netherlands, n. 0123456789, p. 1–37, Apr. 2021. ISSN 0924-8463. DOI:10.1007/s10506-021-09287-3.

DAL PONT, Thiago Raulino; SABO, Isabela Cristina; H, Jomi Fred. Regression in Brazilian Legal Judgments to Predict Compensation for Immaterial Damage. **Natural Language Engineering**, 2021. (Prelo)

Considering the published works and the experiments applied, the contributions to the state of the art in ML applied to legal texts are as follows:

- Pre-trained word embeddings models for Brazilian legal texts, since there was no available representation available before.
- New application of regression in legal textual data.
- Impact of adjustments in the pipeline for regression and classification.

In terms of practical contributions, the models and pipelines from this work will be applied in real legal conciliation hearings in the JEC at UFSC. A legal expert will present and explain the predictions to the parties. Thus, we expect to help them on reaching an agreement without the need of waiting for a judgment. In this way, the

litigation in the JEC would decrease, contributing to faster and more efficient access for citizens to justice.

### 5.3 LIMITATIONS

### 5.4 FUTURE WORK

### 5.5 EXTRA INFORMATION

In the following, the repositories with the codes used in this research are listed:

- Text Representation ([https://github.com/thiagordp/embeddings\\_in\\_law\\_paper](https://github.com/thiagordp/embeddings_in_law_paper))
- Text Classification ()
- Text Regression ([https://github.com/thiagordp/text\\_regression\\_in\\_law\\_judgments](https://github.com/thiagordp/text_regression_in_law_judgments))
- Clustering ([https://github.com/thiagordp/clustering\\_jec](https://github.com/thiagordp/clustering_jec))

### 5.6 ACKNOWLEDGEMENTS

This research was subsidiada by the CAPES in the first 25 months in the CAPES-PROEX. The other 5 by Petrobras.

## REFERÊNCIAS

- AGGARWAL, Charu C.; ZHAI, Cheng Xiang. **Mining Text Data**. Ed. by Charu C. Aggarwal and ChengXiang Zhai. Boston, MA: Springer US, 2012. v. 9781461432, p. 1–522. ISBN 978-1-4614-3222-7. DOI: 10.1007/978-1-4614-3223-4.
- BRAZ, Fabricio Ataide et al. Document classification using a Bi-LSTM to unclog Brazil's supreme court, 2018. arXiv: 1811.11569.
- CARUANA, Rich. Multitask Learning. **Machine Learning**, v. 28, n. 1, p. 41–75, 1997. ISSN 08856125. DOI: 10.1023/A:1007379606734.
- CATH, Corinne. Governing artificial intelligence: ethical, legal and technical opportunities and challenges. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 376, n. 2133, p. 20180080, Nov. 2018. ISSN 1364-503X. DOI: 10.1098/rsta.2018.0080.
- CHALKIDIS, Ilias; KAMPAS, Dimitrios. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. **Artif Intell Law**, Springer Netherlands, v. 27, n. 2, p. 171–198, 2019. ISSN 15728382.
- CHOCRON, Paula; PARETI, Paolo. Vocabulary Alignment for Collaborative Agents: a Study with Real-World Multilingual How-to Instructions. In: PROCEEDINGS of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-18. [S.l.]: International Joint Conferences on Artificial Intelligence Organization, July 2018. P. 159–165.
- CHRISTENSEN, Hans. **HC Corpora**. [S.l.]: Kaggle, 2016. Available from: <https://web.archive.org/web/20161021044006/http://corpora.heliohost.org/>.
- CNJ. **Justiça em Números 2020**. Ed. by CNJ. Brasília: CNJ, 2020a. P. 236.
- CNJ. **Portaria N 271**. Brasília: [s.n.], 2020b. Available from: <https://atos.cnj.jus.br/atos/detalhar/3613>.
- COHEN, Paul R. **Empirical Methods for Artificial Intelligence**. Cambridge, MA, USA: MIT Press, 1995. ISBN 0262032252.

CURY, Augusto et al. **Soluções Pacíficas de Conflitos: Para um Brasil Moderno**. 1. ed. [S.l.]: Forense, 2019. P. 250. ISBN 978-8530982089.

DAL PONT, Thiago Raulino; SABO, Isabela Cristina; HÜBNER, Jomi Fred; ROVER, Aires José. Impact of Text Specificity and Size on Word Embeddings Performance: An Empirical Evaluation in Brazilian Legal Domain. In: CERRI, Ricardo; PRATI, Ronaldo C. (Eds.). Cham: Springer International Publishing, 2020. v. 12319. (Lecture Notes in Computer Science). P. 521–535. ISBN 978-3-030-61376-1. DOI: 10.1007/978-3-030-61377-8\_36.

DAVIS, Anthony E. The Future of Law Firms (and Lawyers) in the Age of Artificial Intelligence. **Revista Direito GV**, v. 16, n. 1, 1dummt, 2020. ISSN 2317-6172. DOI: 10.1590/2317-6172201945.

DRAPER, Norman Richard; SMITH, Harry. **Applied regression analysis**. 3. ed. New York: Wiley, 1998. (Wiley series in probability and mathematical statistics). ISBN 0471221708.

HASSAN, Fahad Ul; LE, Tuyen. Automated Requirements Identification from Construction Contract Documents Using Natural Language Processing. **Journal of Legal Affairs and Dispute Resolution in Engineering and Construction**, v. 12, n. 2, p. 04520009, May 2020. ISSN 1943-4162. DOI: 10.1061/(ASCE)LA.1943-4170.0000379.

JUSBRAZIL. **JusBrasil. Conectando pessoas à justiça**. [S.l.], 2020. Available from: <https://www.jusbrasil.com.br/home>.

KHAN, Nawsher; YAQOOB, Ibrar; HASHEM, Ibrahim Abaker Targio; INAYAT, Zakira; MAHMOUD ALI, Waleed Kamaleldin; ALAM, Muhammad; SHIRAZ, Muhammad; GANI, Abdullah. Big data: Survey, technologies, opportunities, and challenges. **Scientific World Journal**, v. 2014, January 2018, 2014. ISSN 1537744X. DOI: 10.1155/2014/712826.

KIM, Yoon. Convolutional Neural Networks for Sentence Classification. In: PROCEEDINGS of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, Sept. 2014. P. 1746–1751.

KOBER, Jens; BAGNELL, J. Andrew; PETERS, Jan. Reinforcement learning in robotics: A survey. **International Journal of Robotics Research**, v. 32, n. 11, p. 1238–1274, 2013. ISSN 02783649. DOI: 10.1177/0278364913495721.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, v. 521, n. 7553, p. 436–444, May 2015. ISSN 0028-0836. DOI: 10.1038/nature14539.

MANCUSO, Rodolfo de Camargo. **A Resolução dos Conflitos e a Função Judicial no Contemporâneo Estado de Direito**. 3. ed. Belo Horizonte: Juspodivm, 2020. P. 912.

MARLESSONN. **News of the Brazilian Newspaper**. [S.l.]: Kaggle, 2019. Available from: <https://www.kaggle.com/marlesson/news-of-the-site-folhauol>.

MEDVEDEVA, Masha; VOLS, Michel; WIELING, Martijn. Using machine learning to predict decisions of the European Court of Human Rights. **Artificial Intelligence and Law**, 2019. ISSN 15728382. DOI: 10.1007/s10506-019-09255-y.

MIKOLOV, Tomas; CHEN, Kai; CORRADO, Greg; DEAN, Jeffrey. Efficient Estimation of Word Representations in Vector Space. **1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings**, p. 1–12, Jan. 2013. arXiv: 1301.3781.

MITCHELL, Thomas M. **Machine Learning**. 1. ed. [S.l.]: McGraw-Hill Education, 1997. ISBN 9780070428072.

PENNINGTON, Jeffrey; SOCHER, Richard; MANNING, Christopher. Glove: Global Vectors for Word Representation. In: 5. PROCEEDINGS of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2014. P. 1532–1543.

SABO, Isabela Cristina; DAL PONT, Thiago Raulino; ROVER, Aires José; HÜBNER, Jomi Fred. Classificação de sentenças de Juizado Especial Cível utilizando aprendizado de máquina. **Revista Democracia Digital e Governo Eletrônico**, v. 1, n. 18, p. 94–106, 2019.

SADIKU, Asmir. Immaterial Damage and Some Types of its Compensation. **Prizren Social Science Journal**, Prizren Social Science Journal, v. 4, n. 1, p. 50–56, Apr. 2020. DOI: 10.32936/pssj.v4i1.142.



SANTOS, Henrique; WOLOSZYN, Vinicius; VIEIRA, Renata. BlogSet-BR: A Brazilian Portuguese Blog Corpus. In: PROCEEDINGS of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), 2018.

SCHMIDHUBER, Jürgen. Deep learning in neural networks: An overview. **Neural Networks**, Elsevier Ltd, v. 61, p. 85–117, Jan. 2015. ISSN 08936080. DOI: 10.1016/j.neunet.2014.09.003. arXiv: 1404.7828.

SEBASTIANI, Fabrizio. Machine learning in automated text categorization. **ACM Computing Surveys**, v. 34, n. 1, p. 1–47, Mar. 2002. ISSN 0360-0300. DOI: 10.1145/505282.505283. eprint: 0110053 (cs).

SHI, Xingjian; CHEN, Zhourong; WANG, Hao; YEUNG, Dit-Yan; WONG, Wai-kin; WOO, Wang-chun. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. **Advances in Neural Information Processing Systems**, 2015-Janua, p. 802–810, June 2015. ISSN 10495258. arXiv: 1506.04214.

SMYWIŃSKI-POHL, Aleksander; LASOCKI, Karol; WRÓBEL, Krzysztof; STRZAŁTA, Marek. Automatic Construction of a Polish Legal Dictionary with Mappings to Extra-Legal Terms Established via Word Embeddings. In: PROCEEDINGS of the Seventeenth International Conference on Artificial Intelligence and Law - ICAIL '19. [S.l.]: ACM Press, 2019.

STF. **Supremo Tribunal Federal**. [S.l.], 2020. Available from: <http://portal.stf.jus.br/>.

STJ. **STJ - Jurisprudência do STJ**. [S.l.], 2020. Available from: <https://scon.stj.jus.br/SCON/>.

ŞULEA, Octavia Maria; ZAMPIERI, Marcos; VELA, Mihaela; VAN GENABITH, Josef. Predicting the law area and decisions of French supreme court cases. In: INTERNATIONAL Conference Recent Advances in Natural Language Processing, RANLP. [S.l.: s.n.], 2017. P. 716–722. DOI: 10.26615/978-954-452-049-6-092. arXiv: 1708.01681.

TAN, Liling. **Old Newspapers**. [S.l.]: Kaggle, 2020. Available from: <https://www.kaggle.com/alvations/old-newspapers>.

TATMAN, Rachael. **Brazilian Literature Books**. [S.l.]: Kaggle, 2017. Available from: <https://www.kaggle.com/rtatman/brazilian-portuguese-literature-corpus>.

THEODORIDIS, Sergios; KOUTROUMBAS, Konstantinos. **Pattern Recognition**. 4. ed. Burlington: Academic Press, 2009. ISBN 9781597492720.

TJSC. **Jurisprudência Catarinense - TJSC**. [S.l.], 2020. Available from: <http://busca.tjsc.jus.br/jurisprudencia/>.

TRUSOV, Roman; NATEKIN, Alexey; KAL AidIN, Pavel; OVCHARENKO, Sergey; KNOLL, Alois; FAZYLOVA, Aida. Multi-representation approach to text regression of financial risks. **Proceedings of Artificial Intelligence and Natural Language and Information Extraction, Social Media and Web Search FRUCT Conference, AINL-ISMW FRUCT 2015**, v. 7, p. 110–117, 2016. DOI: 10.1109/AINL-ISMW-FRUCT.2015.7382979.

VIRTUCIO, Michael Benedict L. et al. Predicting Decisions of the Philippine Supreme Court Using Natural Language Processing and Machine Learning. In: 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC). [S.l.]: IEEE, July 2018. P. 130–135. DOI: 10.1109/COMPSAC.2018.10348.

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.

WIKIPEDIA. **PT Wiki dump progress on 20191120**. [S.l.]: Wikipedia, 2019. Available from: <http://wikipedia.c3sl.ufpr.br/ptwiki/20191120/>.

ZHANG, Haoyang; ZHOU, Liang. Similarity Judgment of Civil Aviation Regulations Based on Doc2Vec Deep Learning Algorithm. In: 2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI). [S.l.]: IEEE, Oct. 2019. P. 1–8. DOI: 10.1109/CISP-BMEI48845.2019.8965709.

## APPENDIX A – SYSTEMATIC REVIEW OF THE LITERATURE: AI, ML AND LAW

The Systematic Review of the Literature, detailed in this section, focused on finding works related to the applications of TM and ML in the legal domain.

### A.1 DEFINITION OF SEARCH QUESTIONS

The main question in this SRL is: “Which are the Text Mining Techniques applied in the legal domain?”

As secondary questions, there is:

- Which is the legal application?
- What are the pre-processing techniques used?
- What are the representation techniques used?
- What are the feature extraction techniques used?
- What are the classification techniques used?
- What are the clustering techniques used?
- What are the regression techniques used?
- What are the evaluation techniques used?

### A.2 SEARCH STRATEGIES

In order to have an overview of the publications regarding TM and legal domain, we searched in February 29, 2020, for works published on conferences or journals from 2010 to 2020. The search used the following search string:

```
("text mining" OR "natural language processing" OR "nlp" OR
"language processing") AND ("deep learning" OR "machine learning")
AND ("classification" OR "cluster*" OR "regression" OR
"categorization" OR "embedding*" OR "representation" OR "predict*")
AND ("text" OR "document") AND ("law" OR "legal" OR "judicial" OR
"justice" OR "court" OR "legislation" OR "juridical" OR "lawful")
```

In this search, we also added synonymous for Machine Learning, Text Mining, and ML Tasks, and the legal domain.

The search embraced the papers’ titles, abstracts and keywords.

### A.3 KNOWLEDGE BASES

In this SRL, we included bases predominantly related to computing as well as interdisciplinary basis. The list of knowledge bases follows:

- Scopus
- IEEE Xplore
- Web of Science
- ACM Digital Library

#### A.4 INCLUSION AND EXCLUSION CRITERIA

Following the questions of this research, we defined a set of inclusion and exclusion criteria. The process of selection embraced the reading of title, abstract and keywords and the accordance with the criteria:

The following is the list of inclusion criteria:

- Published in journal or conference
- Involves legal processes, or texts with juridical language;
- Involves Machine Learning or Text Mining;
- The techniques used are named;
- Empirical Works.
- Published from 2010 and February 2020.

And the following is the list of exclusion criteria:

- Not written in Portuguese or English;
- Published over 10 years ago;
- Does not involve Machine Learning or Text Mining;
- Does not involve the legal domain;
- Theoretical works.

#### A.5 DATA EXTRACTION PLAN

Information to extract from the papers:

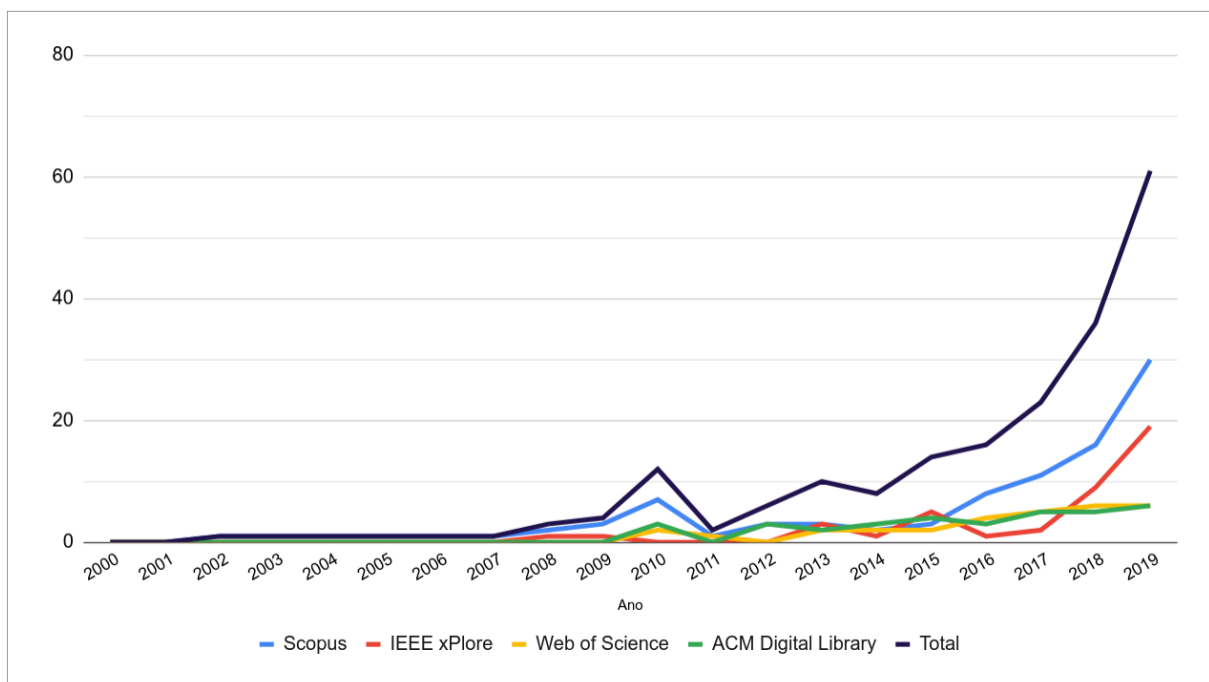
- Title
- Keywords
- Abstract
- Authors
- Year of publication
- Journal or Conference
- Authors affiliation
- Pre-processing techniques
- Representation / Feature Extraction Techniques
- Classification / Clustering / Regression Techniques

- Other techniques
- Model Evaluation Techniques

## A.6 SEARCH EXECUTION AND PRELIMINARY ANALYSIS

After applying the search strings to the knowledge bases on February 29, 2020 without filters on year of publication, we obtained the number of publications shown in Figure 8.

Figure 8 – Publications on ML and TM applied to the legal domain without year filtering



After applying the search strings to the knowledge bases with time filtering on February 29, 2020, the search returned 195. After duplicates removal, the number of works reduced to 147. Finally, from the reading of title, abstract and keywords and the application of the selection criteria, the number of works selected for full reading reduced to 46.

The selected works were used as references in the Introduction chapter, in the Related Works and in the Background.

## A.7 RESULTS AND ANALYSIS

In this section we show a sequence of quantitative data in terms of the tens most frequent authors, journals, text mining techniques and others.

In the following tables, there is ten most frequent authors and conferences found in the SRL and the most frequent techniques for pre-processing, representation, classification, clustering and evaluation.

Table 6 – Ten most frequent authors

<b>Authors</b>	<b>Papers</b>
Matthes, Florian	5
Glaser, Ingo	4
Chalkidis, Ilias	3
Scepankova, Elena	3
Quaresma, Paulo	2
Gonçalves, Teresa	2
Galgani, Filippo	2
Compton, Paul	2
Hoffmann, Achim	2
Androutsopoulos, Ion	2

Table 7 – Ten most frequent journals and conferences

<b>Journal / Conference</b>	<b>Papers</b>
Lecture Notes in Computer Science	8
CEUR Workshop Proceedings	4
Frontiers in Artificial Intelligence and Applications	3
Artificial Intelligence and Law	2
2010 6th International Conference on Wireless Communications, Networking and Mobile Computing, WiCOM 2010	1
Proceedings of the ACM Conference on Computer and Communications Security	1
Foundations and Trends in Information Retrieval	1
Conference on Legal Knowledge and Information Systems	1
Expert Systems with Applications	1
International Conference on Cloud Computing and Services Science	1

Table 8 – Ten most frequent pre-processing techniques

<b>Preprocessing</b>	<b>Papers</b>
Stop words removal	15
Lemmatization	8
Stemming	7
Tokenization	4
Lowercase	4
POS Tagging	3
Normalization	3
Remove punctuation	3
Remove noise	1
Regularization	1

Table 9 – Ten most frequent representation techniques

<b>Representation</b>	<b>Papers</b>
TF-IDF	19
Word2Vec	14
Bag of Words	10
N-Gram	7
Part-of-Speech Tag	6
Named Entity Recognition	5
Doc2Vec	4
Word Embeddings	3
FastText	3
BERT	2

Table 10 – Ten most frequent classification techniques

Classification Tech	Papers
Support Vector Machine	24
Convolution Neural Network	19
Naïve Bayes	18
Decision Tree	17
k Nearest Neighbors	14
Recurrent neural network	14
Random Forest	13
Long Short Term Memory	13
Logistic Regression	11
Conditional Random Field	7

Table 11 – Most frequent clustering techniques

Clustering Tech	Papers
Hierarchical Clustering	2
Fuzzy C-Means	1
Hierarchical LDA	1
k-Means	1

Table 12 – Ten most frequent Evaluation Metrics

Evaluation Metric	Papers
F1-score	23
Accuracy	21
Precision	20
Recall	19
Cross-validation	6
Rouge	2
Area under curve ROC	2
ROC	1
BLEU	1

This SRL did not find works that applied regression techniques in the legal domain.



## APPENDIX B – SYSTEMATIC REVIEW OF THE LITERATURE: TEXT REPRESENTATION

In this SRL, we tried to find work related to the application of text representation techniques on legal texts written in the Portuguese language. However, such task did not succeed due to the absence of work in that sense. Thus, two smaller SRLs were conducted to find papers with broader searches. The first focused on the representation of legal texts in any language and the second in the representation of general texts in Portuguese.

### B.1 DEFINITION OF SEARCH QUESTIONS

The question of the first search: “What are the text representation techniques applied to texts from the legal domain?”

The question of the second search: “What are the text representation techniques applied to texts written in Portuguese?”

### B.2 SEARCH STRATEGIES

Search for representation of legal texts in any languages:

```
("legal" OR "law" OR "court" OR "justice") AND ("embedding*" OR
"language model*" OR "machine learning" OR "deep learning" OR
"natural language processing" OR "text mining") AND ("doc2vec"
OR "paragraph2vec" OR "word2vec" OR "glove" OR "wang2vec" OR
"fasttext" OR "bert" OR "elmo" OR "law2vec")
```

Search for representation of general texts in Portuguese:

```
("portuguese" OR "brazil*") AND ("embedding*" OR "deep learning" OR
"machine learning" OR "natural language processing" OR "text mining")
AND ("doc2vec" OR "paragraph2vec" OR "word2vec" OR "glove" OR
"wang2vec" OR "fasttext" OR "bert" OR "elmo" OR "law2vec" )
```

### B.3 KNOWLEDGE BASES

In this SRL, we searched on the following bases:

- Scopus
- ACM Digital Library
- IEEE Xplore
- Web of Science

## B.4 INCLUSION AND EXCLUSION CRITERIA

Following the questions of this research, we defined a set of inclusion and exclusion criteria. The process of selection embraced the reading of title, abstract and keywords and the accordance with the criteria:

The following is the list of inclusion criteria:

- Published in journal or conference
- Involves legal texts in any languages or general texts in Portuguese;
- Involves Machine Learning or Text Mining;
- The techniques used are named;
- The work evaluate or train representations
- Empirical Work;
- Published from 2010 and May 2020.

And the following is the list of exclusion criteria:

- Work not written in Portuguese or English;
- Published over 10 years ago;
- Does not involve legal texts in any language neither general texts in Portuguese;
- Does not involve Machine Learning or Text Mining;
- Theoretic work

## B.5 DATA EXTRACTION PLAN

In the SRL from this section, we focused on just retrieving the representation techniques used and the application.

## B.6 SEARCH EXECUTION AND PRELIMINARY ANALYSIS

After applying the first search string to the knowledge base on May 7, 2020 for researches published from 2010 to May 2020, the search returned 52 documents. After reading title, abstract and keywords and applying the selection criteria, the number of papers reduced to 12.

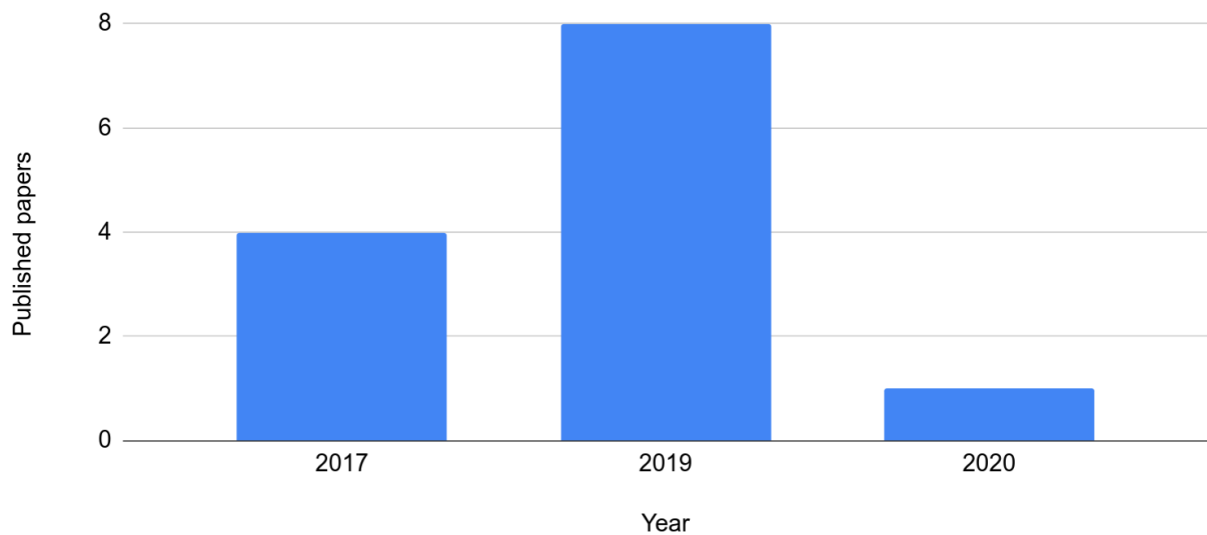
In terms of the second search string, after applying to the knowledge base on May 7, 2020 with the same time filtering, the search returned 136 documents. After reading title, abstract and keywords and applying the selection criteria, the number of papers reduced to 20.

## B.7 RESULTS AND ANALYSIS

In this section, we show the results and analysis in terms of representation techniques for the first part of this SRL related to legal texts from many languages and general texts in Portuguese.

In the first part of this research, we focused on the representation techniques applied in the legal domain. In Figure 9, one can see the distribution by year of the research interest, on the area of representation of legal documents for ML tasks, considering our selection criteria. Although we set the interval to ten years, the selected work only embraced three distinct years.

Figure 9 – Researches by year for text representation in legal documents



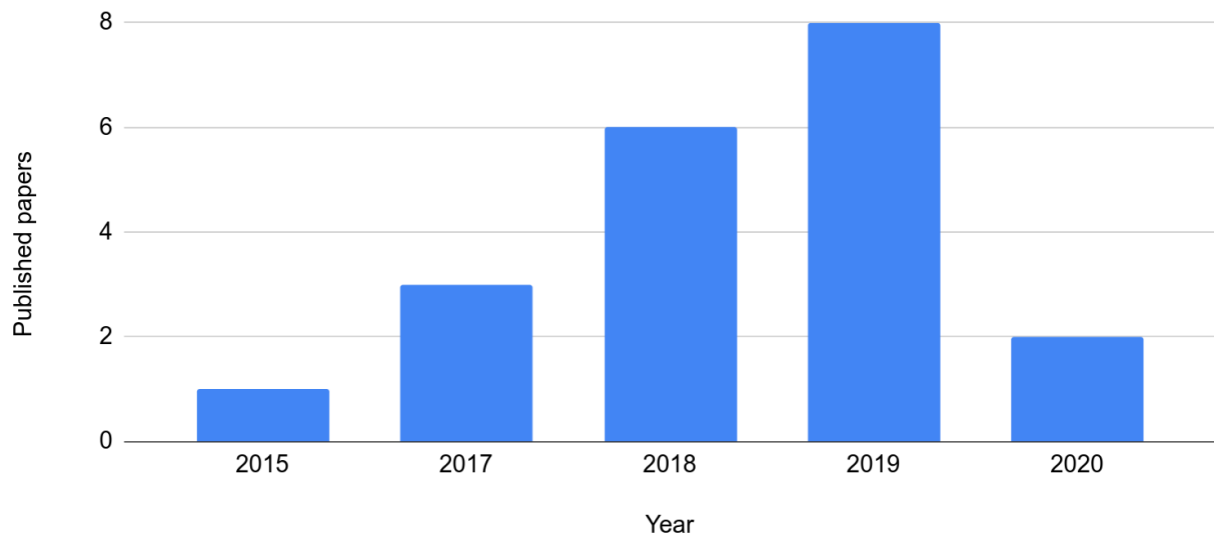
In Table 13, there is the list of representation techniques used in the selected work. Note that, many papers reported using more than one representation techniques in their experiments.

Table 13 – Representations applied to legal texts

Representation	Papers
Word2Vec	8
Doc2Vec	2
WordVec CBOW	2
Glove	2
FastText	2
ELMo	1
Bag of Words	1
Law2Vec	1

In the second part of this research, we focused on the representation techniques applied in general texts written in the Portuguese language. In Figure 10, one can see the distribution by year of the research interest on the area, considering our selection criteria. Although, the SRL embraced the last ten years the selected work embraced the last five.

Figure 10 – Researches by year for text representation in Portuguese documents



In Table 14, there is the list of representation techniques used in the selected work. Note that, many papers reported using more than one representation techniques in their experiments.

Table 14 – Representation used in Portuguese texts

Representation	Papers
Word2Vec	7
Word2Vec Skipgram	7
Glove	6
TF-IDF	3
Wang2Vec Skipgram	3
Bag of Words	2
ELMo	2
FastText	2
FastText Skipgram	2
LDA	2

As mentioned, we do not find, until the date of the SRLs, any work related to evaluation or training of representations of legal texts in the Portuguese language.

## APPENDIX C – SYSTEMATIC REVIEW OF THE LITERATURE: TEXT REGRESSION

In this SRL, we focused on find papers related to the application of regression techniques on legal texts written in Portuguese language. However, such task did not succeed due to the lack of work in that matter. Thus, we applied a broader search regarding the application of regression in texts without context or language limitations.

### C.1 DEFINITION OF SEARCH QUESTIONS

The question of this search is: “What are the regression techniques applied to texts considering any applications and languages?”

### C.2 SEARCH STRATEGIES

The search used the following search string:

```
( "regression on text*" OR "text* regression" OR "regression text*"
OR "regression from text*" OR "regression for text*" )
AND NOT "logistic regression"
```

### C.3 SEARCH RESOURCES

In this SRL, we searched on the following bases:

- Scopus
- ACM Digital Library
- IEEE Xplore
- Web of Science

### C.4 SELECTION CRITERIA

Following the questions of this research, we defined a set of inclusion and exclusion criteria. The process of selection embraced the reading of title, abstract and keywords and the accordance with the criteria:

The following is the list of inclusion criteria:

- Published in journal or conference
- Involves regression applied to textual data;
- Involves Machine Learning or Text Mining;
- The techniques used are named;
- Empirical Work;
- Published from 2010 and December 2020.

And the following is the list of exclusion criteria:

- Work not written in Portuguese or English;
- Published over 10 years ago;
- Does not involve regression applied to textual data;
- Does not involve Machine Learning or Text Mining;
- Theoretic work.

## C.5 DATA EXTRACTION

In the SRL from this section, we focused on just retrieving the regression techniques used and the application.

## C.6 SEARCH EXECUTION AND PRELIMINARY ANALYSIS

After applying the first search string to the knowledge base on December 1, 2020 for researches published from 2010 to December 2020, the search returned 124 documents. After reading title, abstract and keywords and applying the selection criteria, the number of papers reduced to 23.