Pontifícia Universidade Católica
do Rio de Janeiro

Gustavo Martins Campos Coelho

Information Extraction from Legal Opinions in Brazilian Portuguese

Dissertação de Mestrado

Dissertation presented to the Programa de Pós–graduação em Informática, do Departamento de Informática da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática.

Advisor: Prof. Marco Antonio Casanova

Rio de Janeiro
July 2022

Pontifícia Universidade Católica
do Rio de Janeiro

Gustavo Martins Campos Coelho

Information Extraction from Legal Opinions in Brazilian Portuguese

Dissertation presented to the Programa de Pós–graduação em Informática da PUC-Rio in partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the Examination Committee:

Prof. Marco Antonio Casanova
Advisor
Departamento de Informática – PUC-Rio

Profa. Melissa Lemos Cavaliere
PUC-Rio

Prof. Luiz Andre Portes Paes Leme
UFF

Rio de Janeiro, July 22nd, 2022

Gustavo Martins Campos Coelho

Graduated in Mechatronics Engineering by the University of São Paulo (USP).

Bibliographic data

Coelho, Gustavo M. C.

Information Extraction from Legal Opinions in Brazilian Portuguese / Gustavo Martins Campos Coelho; advisor: Marco Antonio Casanova. – 2022.

44 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2022.

Inclui bibliografia

1. Informática – Teses. 2. Processamento de Linguagem Natural. 3. Extração de Informação. 4. Extração de Variáveis em Textos. 5. Classificação de Textos. 6. Reconhecimento de Entidades Nomeadas. I. Casanova, Marco A.. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

I would like to express my gratitude to my Advisor, Prof. Marco Antonio Casanova, who brightly guided me through this research, to the Tecgraf team, with a special acknowledgment to Profa. Melissa Lemos Cavaliere, who welcomed me with open arms into her team, Alimed Celecia and Jefferson de Sousa for their contributions to this research.

I would also like to thank my partner, Nicole Caus, who has shared with me all the tough and bright moments during this journey, my parents, Virgínia Martins Coelho and Décio Jonas Coelho, and my brother, Fernando Martins Campos Coelho, for their long-term support.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Abstract

Coelho, Gustavo M. C.; Casanova, Marco A. (Advisor). **Information Extraction from Legal Opinions in Brazilian Portuguese**. Rio de Janeiro, 2022. 44p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Information Extraction is an important task in the legal domain. While the presence of structured and machine-processable data is scarce, unstructured data in the form of legal documents, such as legal opinions, is largely available. If properly processed, such documents can provide valuable information with regards to past lawsuits, allowing better assessment by legal professionals and supporting data-driven applications. This study addresses Information Extraction in the legal domain by extracting value from legal opinions related to consumer complaints. More specifically, the extraction of categorical provisions is addressed by classification, where six models based on different frameworks are analyzed. Moreover, the extraction of monetary values related to moral damage compensations is addressed by a Named Entity Recognition (NER) model. For evaluation, a dataset was constructed, containing 964 manually annotated legal opinions (written in Brazilian Portuguese) enacted by lower court judges. The results show an average of approximately 97% of accuracy when extracting categorical provisions, and 98.9% when applying NER for the extraction of moral damage compensations.

Keywords

Natural Language Processing; Information Extraction; Text Feature Extraction; Text Classification; Named Entity Recognition.

Resumo

Coelho, Gustavo M. C.; Casanova, Marco A.. **Extração de Informações de Sentenças Judiciais em Português**. Rio de Janeiro, 2022. 44p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A Extração de Informação é uma tarefa importante no domínio jurídico. Embora a presença de dados estruturados seja escassa, dados não estruturados na forma de documentos jurídicos, como sentenças, estão amplamente disponíveis. Se processados adequadamente, tais documentos podem fornecer informações valiosas sobre processos judiciais anteriores, permitindo uma melhor avaliação por profissionais do direito e apoiando aplicativos baseados em dados. Este estudo aborda a Extração de Informação no domínio jurídico, extraindo valor de sentenças relacionados a reclamações de consumidores. Mais especificamente, a extração de cláusulas categóricas é abordada através de classificação, onde seis modelos baseados em diferentes estruturas são analisados. Complementarmente, a extração de valores monetários relacionados a indenizações por danos morais é abordada por um modelo de Reconhecimento de Entidade Nomeada. Para avaliação, um conjunto de dados foi criado, contendo 964 sentenças anotados manualmente (escritas em português) emitidas por juízes de primeira instância. Os resultados mostram uma média de aproximadamente 97% de acurácia na extração de cláusulas categóricas, e 98,9% na aplicação de NER para a extração de indenizações por danos morais.

Palavras-chave

Processamento de Linguagem Natural; Extração de Informação; Extração de Variáveis em Textos; Classificação de Textos; Reconhecimento de Entidades Nomeadas.

Table of contents

List of figures

List of Abreviations

BiLSTM – Bidirectional Long Short-Term Memory
BoW – Bag-of-Words CBOW – Continuous Bag-of-Words
CNN – Convolutional Neural Network
CRF – Conditional Random Field
DBOW – Distributed Bag-of-Words
DM – Distributed Memory
IDF – Inverse Document Frequency
LSTM – Long Short-Term Memory
MuDEC – Multi-step Document Embedding-based Classifier
NER – Named Entity Recognition
NLP – Natural Language Processing
RNN – Recurrent Neural Network
SIF – Smooth Inverse Frequency
SMOTE – Synthetic Minority Oversampling Technique
TF – Term Frequency
TF-IDF – Term Frequency-Inverse Document Frequency

1
Introduction

The excessive duration of a legal case in the Brazilian courts imposes a challenge to legal professionals and the general society. The high volume of new legal cases entering yearly at the courts, combined with the existing backlog, adds more complexity to this matter, encouraging the automation of processes in this context. Between January and April of 2022, the lower court of the State of Rio de Janeiro reported more than 369 thousand new cases, while the mean duration of a case is estimated at 1,518 days [1], indicating the need for improving efficiency in the Brazilian legal system.

Over the past years, efforts have been made to address this issue with the use of Artificial Intelligence as a tool for increasing court efficiency and switching the approach from knowledge-representation techniques toward machine-learning-based approaches. Like in most data-driven methods, this approach requires high-quality, structured, machine-processable data, which is generally scarce in the legal domain (SURDEN, 2018). On the other hand, unstructured data in the form of legal documents, such as legal opinions, is largely available. If properly processed, such documents can provide valuable structured information that can be used to describe each legal case. The description of legal cases by a structured and interpretable dataset can be further used in a variety of applications, such as Similar Case Matching (XIAO et al., 2019), Legal Judgment Prediction (ZHONG et al., 2018), Recommendation Systems and other data-driven applications. Thus, Information Extraction in the Legal Domain becomes an important task for converting unstructured data, such as raw text documents, into structured formats, allowing the use of different machine-learning-based models in this domain.

Information Extraction from text is an important task in text mining. The general goal of Information Extraction is to discover structured information from unstructured or semi-structured text (JIANG, 2012). This task is then divided into multiple sub-tasks which aim at extracting different information types in documents. For instance, Information Extraction based on Natural Language Processing (NLP) models can be addressed by sub-tasks such as Relation Extraction, Element Detection, Word and Document embedding, and others (ZHONG et al., 2020). In this work, Information Extraction from text is addressed by two different approaches: (i) Text Classification and (ii) Named Entity Recognition (NER).

[1] https://painel-estatistica.stg.cloud.cnj.jus.br/estatisticas.html

More specifically, this work is positioned in the context of the automated analysis of legal opinions related to consumer complaints. A *legal opinion* is "a written explanation by a judge or group of judges that accompanies an order or ruling in a case, laying out the rationale and legal principles for the ruling"[2]. A *consumer complaint* is "an expression of dissatisfaction on a consumer's behalf to a responsible party"[3]. In such cases, the legal opinion contains specific provisions referring to the plaintiff's claim, such as moral damage, material damage, and legal fees due by the defeated party. The use of the term *legal opinion* is restricted to this particular context in what follows.

We address two related problems:

P1. Classification of legal opinions.

P2. Extraction of monetary values from legal opinions.

To address the first problem, we analyze several models to classify legal opinions in pre-defined sets of classes, according to relevant features in the given context, such as the case ruling (Accepted, Partially Accepted, Rejected or Dismissed), the type of monetary restitution (Simple, Doubled or None), and other context-specific features. To address the extraction of numerical features (such as the moral damage value), we propose a NER approach, where the monetary values related to moral damages are treated as entities, and further post-processed for the conversion to structured features.

For the evaluation of the proposed models, we created a dataset containing 964 manually annotated legal opinions (in Brazilian Portuguese) enacted by lower court judges in the State Court of Rio de Janeiro in the context of consumer complaints involving electric power companies

The rest of the dissertation is organized as follows. Chapter 2 introduces background concepts and summarizes related work. Chapter 3 describes the models for Text Feature Extraction used in the experiments. Chapter 4 describes Information Extraction from text with a focus on Text Classification and NER. Chapter 5 describes the experiments and compares the results. Finally, Chapter 6 presents the conclusions and directions for future research.

[2]https://en.wikipedia.org/wiki/Legal_opinion
[3]https://en.wikipedia.org/wiki/Consumer_complaint

2
Background and Related Work

This chapter overviews recent approaches related to the extraction of information in the legal domain. More precisely, it focuses on the extraction of information from legal documents by using NLP-based models to address the tasks of Text Classification and NER, where this dissertation is more specifically positioned.

2.1
Background

As a broad research topic, NLP-based information extraction from text is addressed by multiple approaches, depending on the volume of available training instances, quality of annotations, specific goals and interpretation requirements. Zhong et al. (2020) summarizes NLP methods in the legal domain in two main groups (Symbolic-based and Embedding-based Methods), which are applied to multiple applications, such as judgment prediction, question answering and text summarization (figure 2.1). Symbolic-based Methods apply interpretable handcrafted symbols to legal texts (such as the extraction of entities and relations), while embedding-based methods aim at designing efficient neural models to achieve better performance (such as character, word and concept embeddings).

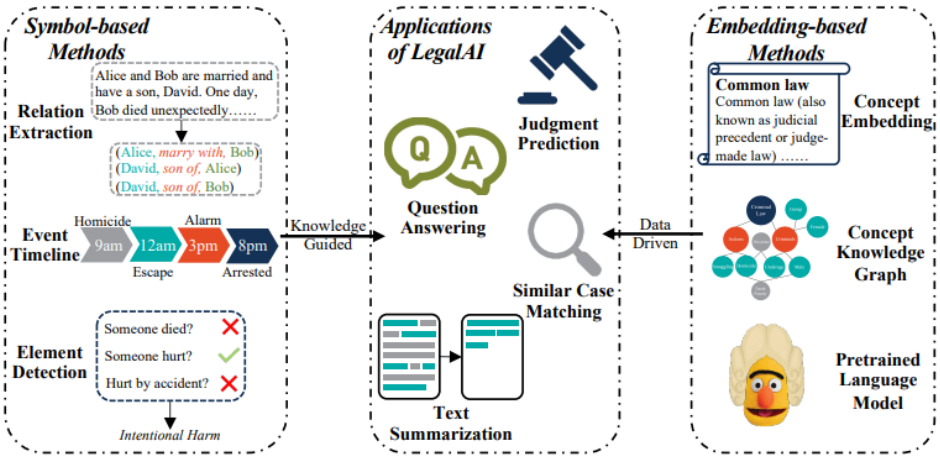


Figure 2.1: Overview of NLP applications and methods in the Legal Domain (ZHONG et al., 2020).

The two described groups of methods have opposite main challenges. While interpretable symbolic models lack effectiveness, embedding-based models lack interpretability, which may bring ethical issues to the legal system such

as gender bias and racial discrimination. Thus, the combination of these methods might be suitable for mitigating their limitations and potentializing their advantages. For instance, NER models can take advantage of word embeddings for the extraction of handcrafted and interpretable entities. Similarly, text classification can benefit from document embeddings for classifying documents into pre-defined legal-related categories.

Chalkidis and Kampas (2019) explores the use of deep learning in the Law area by dividing the research into three categories. In the first category, relevant works related to Text Classification are listed, where textual units such as sentences, paragraphs and documents are categorized. The second category addresses Information Extraction in the context of sequence labeling tasks, such as chunking and Named Entity Recognition. The third category focuses on question-answering tasks by tackling the problem of retrieving articles of interest out of a collection of legal documents or articles that entail a query. The study addressed by this dissertation falls under the first and second categories, expanding the models to include traditional machine learning models for text classification, in addition to deep learning approaches.

2.2 Text classification in the Legal Domain

Text Classification is an important task in Information Extraction. Sulea et al. (2017) argues that the use of text classification, as in various other domains, can benefit legal professionals by providing a decision support system or at least a sanity check system. The proposed framework uses word unigrams and word bigrams as features and an Ensemble classifier as the classification model. The goal is to assign each legal document to one class, from a predefined set of classes. The results show a 98% average F1-score in predicting a case ruling, 96% for predicting the law area of a case, and 87.07% for estimating the date of a ruling.

Minaee et al. (2021) explores the use of deep learning in text classification by listing more than 150 deep learning-based models. The list includes feed-forward networks, RNN and CNN-based models, graph neural networks and hybrid models. The survey shows that deep learning-based models surpass classical machine learning-based approaches, improving the state of the art on various text classification tasks.

Wei et al. (2018) provides an empirical study on Text Classification in the legal industry by comparing traditional machine learning classifiers such as logistic regression and support vector machines with deep learning models. The experiments are based on four different datasets, each of them containing

millions of records. The results show that models based on Convolutional Neural Network (CNN) can perform better in the given scenario, especially when considering high volumes of training data. However, the complexity related to deep learning model optimization limits their applications, once they require longer training intervals, and more advanced setups, with the use of GPUs during implementation.

In the Brazilian Legal Domain, Araujo et al. (2020) introduces a dataset built from Brazil's Supreme Court digitalized legal documents, composed of more than 45 thousand appeals, which includes roughly 692 thousand documents. The documents contain labels related to the document type and lawsuit theme. The presented baseline is composed of bag-of-words models, CNNs, Recurrent Neural Networks (RNNs) and boosting algorithms. The results show that CNN and Bidirectional Long Short-Term Memory (BiLSTM) outperforms the remaining models in all categories, emphasizing the potential of deep learning approaches in this task.

2.3 Named Entity Recognition in the legal domain

The extraction of named entities is a frequent approach for information extraction in the legal domain, where usual entities such as person, organization and location are combined with entities related to the legal context. Leitner, Rehm and Moreno-Schneider (2019) addresses this task by extracting several fine-grained semantic entities, such as company, institution, court and regulation. Models based on BiLSTM and Conditional Random Field (CRF) are applied to the task, with character embedding. The results of both model families demonstrate that BiLSTMs models outperform CRF with an F1-score of 95.46% for the fine-grained classes and 95.95% for the coarse-grained classes.

In the Brazilian legal context, Araujo et al. (2018) introduces a dataset for NER in Brazilian legal documents. Along with the dataset, a baseline model is proposed with a combination of LSTM and CRF. The models presented competitive performance when applied to the extraction of legislation and legal case entities, achieving 97.04% and 88.82% of F1-scores, respectively. Correia et al. (2022) introduces a corpus composed of 594 decisions, where law students annotate two levels of nested legal entities. The nested entities are defined by four coarser and twenty-four fine-grained legal named entities, such as legal procedure number, judgment date and others.

Fernandes et al. (2022) proposes a set of NER models to extract information from legal opinions enacted by lower and Appellate Courts. More specifically, three datasets are built to identify legal entities such as the moral

and material damage values, the legal fee due by the defeated party and others. Five models are proposed based on different combinations of word and character embeddings, RNNs and CRFs. The optimal results reached by the models range from 68.42% and 90.43% depending on the dataset.

Fernandes et al. (2020) extracts modifications proposed by the Brazilian upper Court to lower Court judge's decisions. The task is performed by firstly defining six entities that correspond to the most popular legal categories that are modified by the Appellate Court in a specific legal domain. The extraction of these entities is evaluated by five models based on different combinations of RNNs and CRFs, and the best performance is reached by a combination of a BiLSTM and a CRF layer.

2.4 Chapter Conclusion

Text Classification and NER are common NLP tasks related to Information Extraction in the Legal Domain. Their outcome provides valuable information to the legal industry, enabling and providing legal reasoning to other AI applications such as Judgment Prediction.

The current state-of-the-art methods for both tasks are based on deep learning architectures. While NER models usually perform better with a combination of LSTM and CRF, Text Classification models based on CNN and LSTM tend to outperform other classic Machine Learning-based models. However, the experiments listed are evaluated by relatively large datasets, usually composed of thousands or even millions of instances, which is suitable for deep learning models. The scarcity of large legal-related datasets in Brazilian Portuguese is a challenge in this context.

3
Text Feature Extraction

Text Feature Extraction is the task of representing text instances in a structured set of numerical features Kowsari et al. (2019). Such representations are an essential data source for many Machine Learning-based models, which require structured datasets for optimization and inference processes, turning this task into a core step for Information Extraction from text. This chapter addresses some of the main Feature Extraction models, along with comparisons referred to the quality of representations and computational complexity between models.

3.1
Bag of Words

One of the simplest methods for the extraction of features from text is the Bag of Words model (BoW) (HARRIS, 1954). This model is based on a one-hot-encoded representation of words in a vocabulary, where each unique word is assigned to a specific index i. In a vocabulary of size n, a word is represented by a sparse vector filled with one at the i-th position, and with zeros at the remaining $n - 1$ positions.

Based on the BoW approach, text instances composed of a set of words are similarly represented by a vector of size n, where each i-th position in the vector is filled with a statistic measure associated with the corresponding word. A commonly used measure is the Term Frequency-Inverse Document Frequency (TF-IDF) (JONES, 1972), which combines a measure of a word's frequency in the text and how common or rare the word is compared to a set of documents.

To calculate the first component of TF-IDF, the Term Frequency (TF) is measured by dividing the absolute frequency $f(i, d)$ of a word i in the document d by the sum of frequencies of each word in the document (equation 3-1). The second component, or the Inverse Document Frequency (IDF) is calculated based on the uniqueness of a word i in a set of documents D of size N, according to equation 3-2. Lastly, the Term Frequency-Inverse Document Frequency (TF-IDF) is calculated by simply multiplying both values (equation 3-3).

$$TF(i,d) = \frac{f(i,d)}{\sum_{i' \in d} f(i',d)} \tag{3-1}$$

$$IDF(i,D) = \log \frac{N}{|d \in D : i \in d|} \tag{3-2}$$

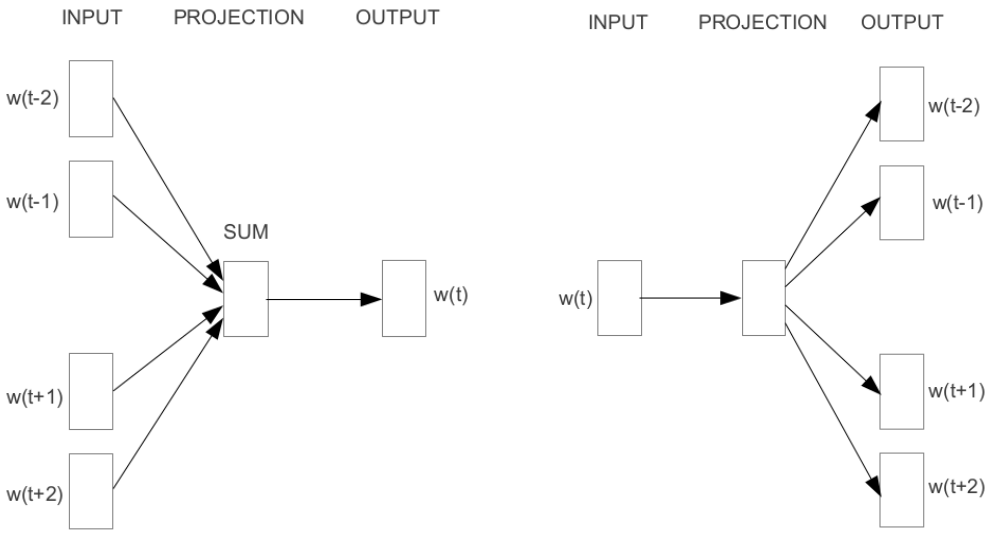
$$TF - IDF(i,d,D) = TF(i,d) \cdot IDF(i,D) \tag{3-3}$$

The Bag of Words approach, while being an intuitive method and simplistic in its implementation, presents three main practical limitations. The first is the lack of semantic information since every word is equally distant from each other on its one-hot-encoded representation. This leads to words with similar semantic meanings having the same similarity as two words with distant semantic meanings. The second is the resulting excessive high dimensional representations, once the dimension of each word vector is equal to the size of the vocabulary, which is frequently composed of thousands or even millions of words. In addition, once only the index which corresponds to the word position is filled with a non-null value, the method is characterized by the inefficient use of computational resources. The third main limitation is the lack of contextual information related to the position of the words in the text. Sentences composed of the same words in different order frequently represent different meanings (e.g. "This is my car" and "Is this my car"). By mapping simple features such as the term frequency, the resulting text representation does not address these contextual differences.

3.2 Word2vec

As a more complex and robust representation of words when compared to BoW, Mikolov et al. (2013) introduces a continuous vector representation of words commonly referred to as Word2vec. This approach is based on the optimization of a neural network classifier, with a softmax activation function at the last layer, for the prediction of a target word based on its context (Continuous Bag-of-Words) or the predictions of context words based on a target word (Continuous Skip-gram), as shown in figure 3.1.

The resulting models provide weights that are used to represent the words in a continuous space, also called word embeddings. In opposition to BoW, the word embeddings express similarities between words with similar semantic meanings. The problem of high dimensional and sparse vectors is also addressed since the size of the vector is defined by the model's dimensionality and treated as a hyperparameter. In addition, since there is no need for labeled data, the



(a) Continuous Bag-of-Words (CBOW). (b) Skip-gram.

Figure 3.1: Models approached by Word2vec.

word vectors can be trained in a large corpus, and made available for different applications.

Although its advantages, the use of Word2vec for Text Classification adds a new challenge related to the representation of text, in oppose to the representation of individual words. Traditional Machine Learning based methods require fixed-length vectors for classification, while a text is composed of a variable number of words. Thus, when used for text representation, Word2vec requires additional approaches to represent a set of word embeddings of variable sizes.

3.3 Smooth Inverse Frequency

Smooth Inverse Frequency (SIF) (ARORA; LIANG; MA, 2017) is a method for text representation based on the simple approach of word embedding averages. The original text is firstly converted into a set of d-dimensional word embeddings by using Word2vec or any other similar method such as Glove (PENNINGTON; SOCHER; MANNING, 2014). Once the text is represented by a set of word embeddings, the weighted average is computed. The weight w of a word is given by equation 3-4, where a is a hyperparameter and $p(w)$ is the word's estimated frequency. Thus, the result for a corpus composed of n text instances is a matrix of size n x d.

$$w = \frac{a}{a + p(w)} \tag{3-4}$$

In addition, the projection of the resulting averages on their first singular vector is removed. The intuition behind this operation is the removal of syntactic information of common words, which are not relevant for distinguishing the given text instances.

One clear advantage of SIF is the simplicity of its implementation and low computational cost. Once pre-trained word embeddings can be used, the remaining operations (averaging, single value decomposition and optional fine-tuning of embeddings) can be performed at a relatively low cost. Although its simplicity, this method is claimed by the authors to outperform sophisticated supervised methods, including RNNs and LSTMs.

As a downside, SIF doesn't take word positions into account, once the averaging operation disregards the embeddings order. Similar to the Bag of Words approach in this aspect, a piece of contextual information is missed.

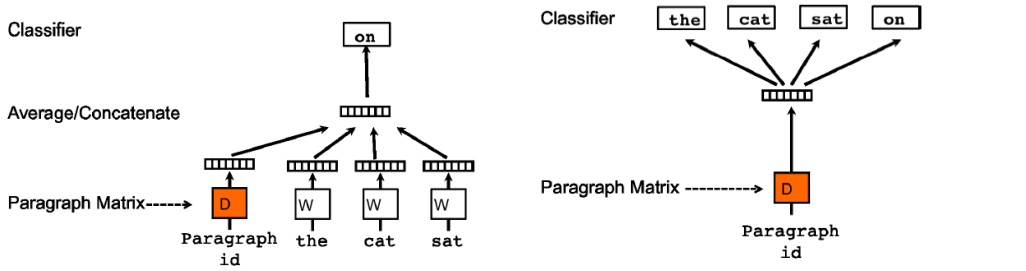
3.4 Distributed Representation of Documents

In a similar approach when compared to Word2vec, the Distributed Representation of Documents (LE; MIKOLOV, 2014), commonly referred to as Paragraph Vector or simply Doc2vec, represents a text instance by a dense vector based on the prediction of words in the text.

Just like Word2vec is divided into two approaches (CBOW and Ski-gram), Doc2vec is similarly divided into the following models: Distributed Memory (DM) and Distributed Bag of Words (DBOW). In the first model (DM), illustrated by figure 3.2(a), a paragraph (or text) ID is added to the context words for the prediction of the target word. The ID acts as a memory that remembers what is missing from the context, or the text topic. At the inference of a new text instance, the word vectors and the softmax layer are fixed and the new text vector is inferred by gradient descent.

In the DBOW approach, illustrated by figure 3.2(b), the paragraph ID is used as the only input during the classification, while the output word is randomly sampled from the context window. This model requires less storage, since only the softmax weights are stored, in oppose to DM, where the word vectors are stored in addition.

The main advantages of Doc2vec are the use of word semantics by the dense word vectors and the capacity to capture contextual information in word



(a) Distributed Memory (DM). (b) Distributed Bag of Words (DBOW).

Figure 3.2: The two approaches used by Doc2vec.

positions, in opposite to SIF. In fact, the results shown by Le and Mikolov (2014) indicate superior results for several text classification and sentiment analysis tasks when compared to other methods based on vector averaging and BoW.

3.5 Chapter Conclusion

Models for Text Feature Extraction vary in their quality of representations, as well as their complexity. In word-level representations, the BoW model is a simple approach that lacks semantics, along with inefficient use of computational resources, while Word2vec offers a more robust and efficient model, enabling semantic representation of words.

In text-level representations, TF-IDF and SIF are intuitive and simple methods during implementation, but present relevant conceptual limitations, such as the lack of sensitivity to word positions. On the other hand, Doc2vec is a fairly more complex model, capturing both the semantics of words and the general context.

In the proposed task, text feature extraction is an essential step for word-level and text-level representations, which can be used in different Information Extraction tasks, impacting their overall results.

4
Information Extraction from Text

In this chapter, some of the models applied for Information Extraction are presented. More specifically, the models are divided into two different groups, according to their main tasks. In the first group, information is extracted based on Text Classification, while the second group aims at the task of Named Entity Recognition (NER). Lastly, the two approaches are compared in the face of the specific task of provisions extraction from legal opinions, establishing a pattern for the selection of models based on the provision to be extracted.

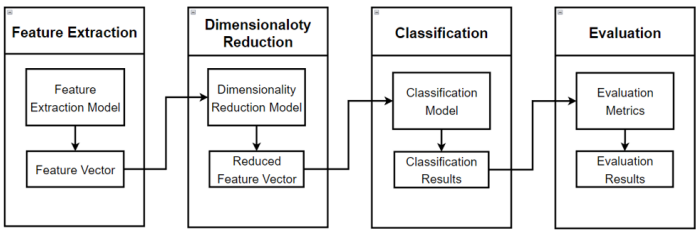
4.1
Text Classification

Text classification is a popular Natural Language Processing (NLP) task that aims at predicting the categorical values associated with textual instances. Such instances might be composed of phrases, paragraphs, or even entire documents, naturally increasing the task's complexity.

Kowsari et al. (2019) summarizes most text classification systems as a four-step procedure, illustrated by figure 4.1. The first step converts textual units into fixed-length numerical vectors by using a feature extraction model, as detailed in chapter 3. The second step covers an optional dimensionality reduction over the results of the first step, which is potentially high dimensional, depending on the feature extraction model applied. Ideally, the dimensionality reduction should minimize the model complexity while preserving effective information related to the classes of interest. The third step follows with a classification model, such as Naïve Bayes, support vector machines (SVM), gradient boosting trees and random forests. In this step, each reduced feature vector referred to a document is classified in one of the pre-defined classes. Finally, the fourth step evaluates the results in comparison with the true values, and the applicable metrics are computed.

4.1.1
Multi-step Document Embedding-based Classifier

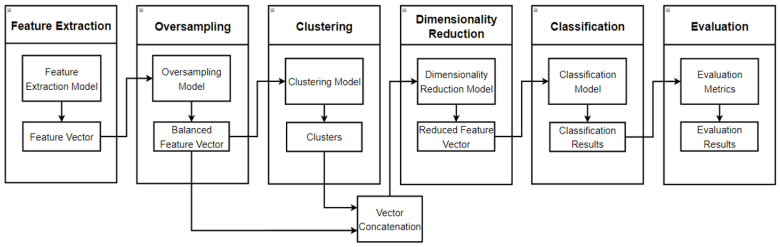
Coelho. et al. (2022) argues that this framework can benefit from additional steps, addressing different aspects to optimize the classification performance. The resulting framework is named Multi-step Document Embedding-based Classifier (MuDEC), illustrated in figure 4.2. In comparison with the standard framework, MuDEC adds the oversampling and clustering steps. The



Models				
TF-IDF Classifier	TF-IDF	-	Logistic Regression	Precision, Recall, Accuracy, F1-Score.
SIF Classifier	SIF			
Doc2vec Classifier	Doc2vec			

Figure 4.1: Standard Text Classification framework.

oversampling step aims at balancing the dataset distribution, when one or more categories are less frequent than the remaining. The use of this approach is justified by the possible negative impact of imbalanced datasets in classification algorithms, leading to a model bias towards the majority classes. In the original implementation, SMOTE is chosen as the oversampling model. This method creates synthetic instances by using a k nearest neighbors approach, where k instances features are randomly chosen from the minority classes, and a synthetic instance is created along the line segments joining them (CHAWLA et al., 2002). This procedure is repeated until the dataset is evenly distributed.



Model						
MuDEC	Doc2vec	SMOTE	K-Means	-	Logistic Regression	Precision, Recall, Accuracy, F1-Score.

Figure 4.2: MuDEC framework.

The clustering step addresses the identification of textual patterns. The main idea behind this step is based on the existence of sets of instances with a similar format (e.g., legal opinions containing a similar writing style). In the original implementation, k-Means is used for the identification of clusters based on the feature vector representations, and the cluster to which each instance belongs is turned into a new one-hot-encoded categorical feature and concatenated to the related feature vector. Intuitively, this task provides a feature that indicates relations between instances of the same cluster, possibly contributing to the classification.

The results demonstrate the positive impact of the additional steps, with exception of the dimensionality reduction, which slightly reduces the overall accuracy, while increasing the computational efficiency in terms of running time.

4.1.2 C-LSTM

In a different framework, Zhou et al. (2015) introduces a neural network approach for text representation and classification named C-LSTM. The strategy used by this model combines CNN and LSTM layers, as shown in figure 4.3. First, each token in the text is converted to dense word vectors by applying a word embedding model. The result is an L x d matrix, where L is the number of tokens in the text and d is the word embedding dimension. A CNN is then applied to the matrix, where n filters of fixed size generate n feature maps with w windows. The resulting vectors are rearranged as feature representations for each window, and the result is fed to an LSTM layer. Lastly, the LSTM layer outputs a fixed-length text representation, which is connected to a softmax layer for classification.

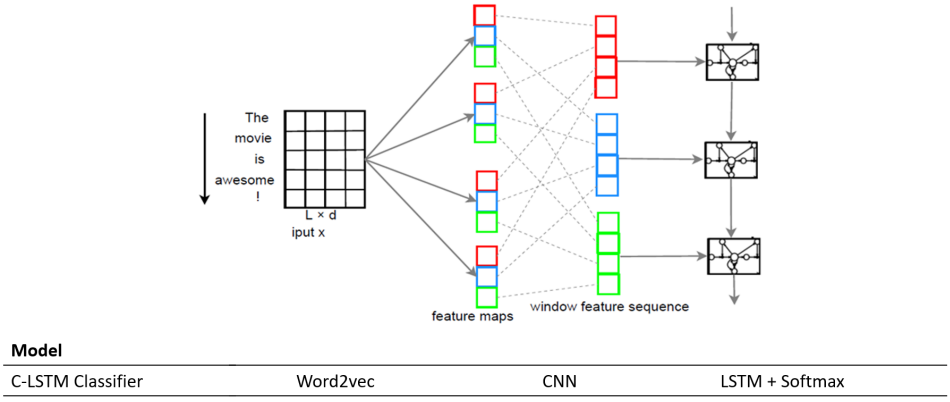


Figure 4.3: C-LSTM framework.

Since CNNs and LSTMs adopt different ways of understanding natural language, they work in different roles inside this framework. While the CNN layer is used to capture a sequence of higher-level phrase representations, the LSTM layer captures global and temporal semantics. Thus, C-LSTM can map both word semantics (with the use of word embeddings) and local and global contextual information from text instances.

Despite its sophistication in terms of text representation and further classification, the C-LSTM framework can potentially require larger datasets for competitive performance, due to its complexity inherited from deep neural

networks, where the number of parameters to be optimized can easily result in overfitting when small training sets are used.

4.2 BERT-CRF for NER

Named Entity Recognition is the task of identifying named entities in text and classifying them into predefined categories, such as a person, location or any other class of interest. NER systems are used in many NLP applications, such as question answering, topic modeling, and information retrieval. Although being a well-known technique for decades, there has been a recent change of approach from handcrafted rules, lexicons, orthographic features and ontologies to deep neural network architectures, resulting in very accurate models when provided with enough train instances (YADAV; BETHARD, 2019).

Souza, Nogueira and Lotufo (2019) proposes a BERT model for a Portuguese NER task. The model, illustrated by figure 4.4, receives plain text as input. In the example, the input sentence "David Gilmour performs live at Pompeii" is composed of two entities: "David Gilmour" (Person) and "Pompeii" (Location).

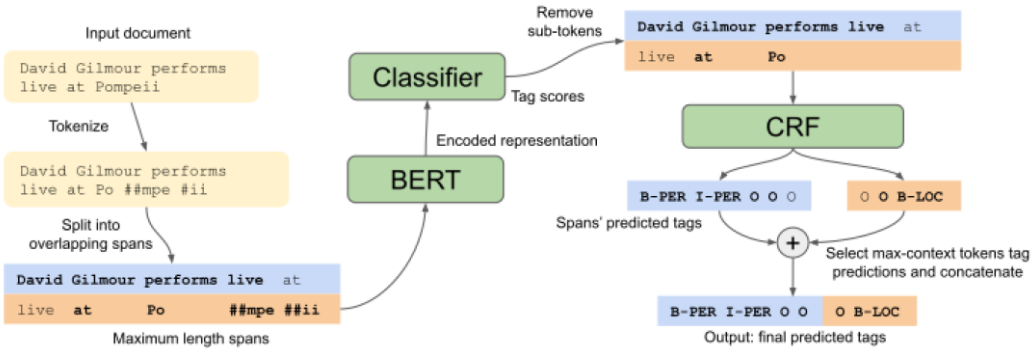


Figure 4.4: The architecture of the BERT model for NER.

The sentence is first word-level tokenized, and the text is split into overlapping spans of pre-defined length. For each span, the words are converted into BERT embeddings (DEVLIN et al., 2018). These features are fed to a neural network classifier with a softmax output layer, which returns the scores for each token. The scores indicate the likelihood of a token being classified to each entity, based on the semantic representation given by the embeddings. The scores are then fed to a Linear-Chain CRF, which classifies each token into its corresponding entity, taking into account their context, given by the sequence of scores in each span. As an alternative, the CRF layer can be removed from this architecture, extracting the end results directly from the classifier.

The model's training process can vary in two main approaches. In the Fine-tuning approach, a linear layer is used as the classifier and all weights are optimized jointly during training, including BERT, classifier and CRF weights. The Feature-based approach uses a 1-layer BiLSTM model as the classifier. This approach freezes the BERT weights during training, while the classifier and CRF are optimized.

The BERT-CRF approach for NER has shown competitive results when applied to tasks consisting of 5 and 10 entity classes in the Portuguese language. According to the original paper, the results reached more than 80% in precision, recall and F1 score.

4.3 Chapter Conclusion

Text Classification and NER are two important tasks positioned in the context of Information Extraction from Text. While Text Classification models are applied for the extraction of categorical target variables from entire pieces of text, NER extracts specific named entities by identifying their positions in the text.

The differences between these two approaches impact their usage in two main aspects. First, the annotation process for Text Classification is significantly less complex, once each text unit is simply assigned to a single categorical value, while NER requires part-of-speech (POS) tagging, including the entity type, beginning and end positions in the text. Furthermore, Text Classification is limited to categorical target variables, hindering the extraction of numerical values, which can assume an infinite number of unique values. Since NER extracts pieces of text, numerical variables can be extracted as entities and converted to suitable formats, such as integers or float variables. These differences are highly important in the context of extracting provisions from legal opinions, since both categorical and numerical target variables are present.

5
Experiments

In this chapter, the experiments conducted for the extraction of provisions from legal opinions are addressed. Firstly, the experimental setup is detailed, along with the dataset used for the model's training and evaluation, and the optimization of hyperparameters. Lastly, the main results are presented for the extraction of categorical and numerical provisions, as well as a performance comparison between models.

5.1
Experimental Setup

The selection of models applied to each provision is addressed according to the provision type. In the current context, a legal opinion can take only one class for each categorical provision (e.g. a Case Ruling can either be Accepted, Partially Accepted, Rejected or Dismissed). For this reason, the extraction of a categorical provision is treated as a Text Classification task. Additionally, the extraction of the single numeric provision present in the dataset is treated as a NER task, by extracting the values directly from the text and converting them to integer variables.

5.1.1
Datasets

The experiments are structured based on a dataset containing 964 manually annotated legal opinions (in Brazilian Portuguese) enacted by lower court judges in the State Court of Rio de Janeiro in the context of consumer complaints involving electric power companies. Each legal opinion in the dataset was manually analyzed by legal professionals to locate six types of provisions. Table 5.1 describes the categorical provisions, along with their related categories.

Categorical provisions	Categories
Case ruling	Accepted, Partially Accepted, Rejected, Dismissed
Cancellation of TOI	True, False
Restoration of supply	True, False
Restitution	None, Simple, Doubled
Removal from credit downgrade score list	True, False

Table 5.1: Categorical provisions and unique values.

In this context, a *case ruling* is "a court's decision on a matter presented in a lawsuit"[1]. The term *TOI* refers to a report issued by an electrical company resulting from an irregularity allegedly present in the existing home electric supply, found during an inspection by one of its employees. In this context, the *Cancellation of TOI* refers to the illegality of this procedure and the cancellation of its effects, usually related to monetary penalties to the consumer. The *restoration of supply* indicated if the electric company should reestablish the plaintiff's supply. The *restitution* establishes if the electric company should refund excessive monetary charges paid by the plaintiff, and whether the refund value should be doubled or not. Lastly, the *removal from the credit downgrade score list* indicates if the inclusion of the plaintiff in a credit downgrade score list by the electric company should be reverted.

Along with the described categorical provisions, numerical provisions are also present in the legal opinions. Examples of these provisions are the moral and material damage compensations and the legal fee due by the defeated party. In this study, the moral damage compensation is the only numerical provision addressed, leaving the remaining for future experiments. Along with the value associated with the given moral damage, each document was POS tagged, denoting the position in text where the values are expressed.

Figure 5.1 illustrates the presence of the provisions in a legal opinion, where the expressions referred to case ruling, cancellation of TOI, restitution and moral damage values are highlighted. Section 5.1.2 presents more details with regards to legal opinions structure.

Isso posto, julgo PROCEDENTE EM PARTE OS PEDIDOS, para confirmar a tutela antecipada deferida, e declarar o cancelamento do TOI nº[redacted]; condeno a parte ré a restituir, na forma simples, todos os valores pagos a título do mencionado TOI, corrigidos monetariamente do desembolso e com juros de mora de um por cento ao mês desde a citação; condeno a parte ré a compensar a autora no valor de R$ 1.500,00 (mil e quinhentos reais) corrigidos monetariamente desta data e com incidência de juros de mora de um por cento ao mês desde a citação. Condeno a parte ré nas custas processuais e em honorários advocatícios no percentual de dez por cento do proveito econômico auferido com a presente sentença.

Registrada digitalmente. Publique-se. Intimem-se.

Figure 5.1: Example of the operative part of a legal opinion.

The annotated dataset distribution for each provision is illustrated by figure 5.2.

[1]https://www.law.cornell.edu/wex/ruling

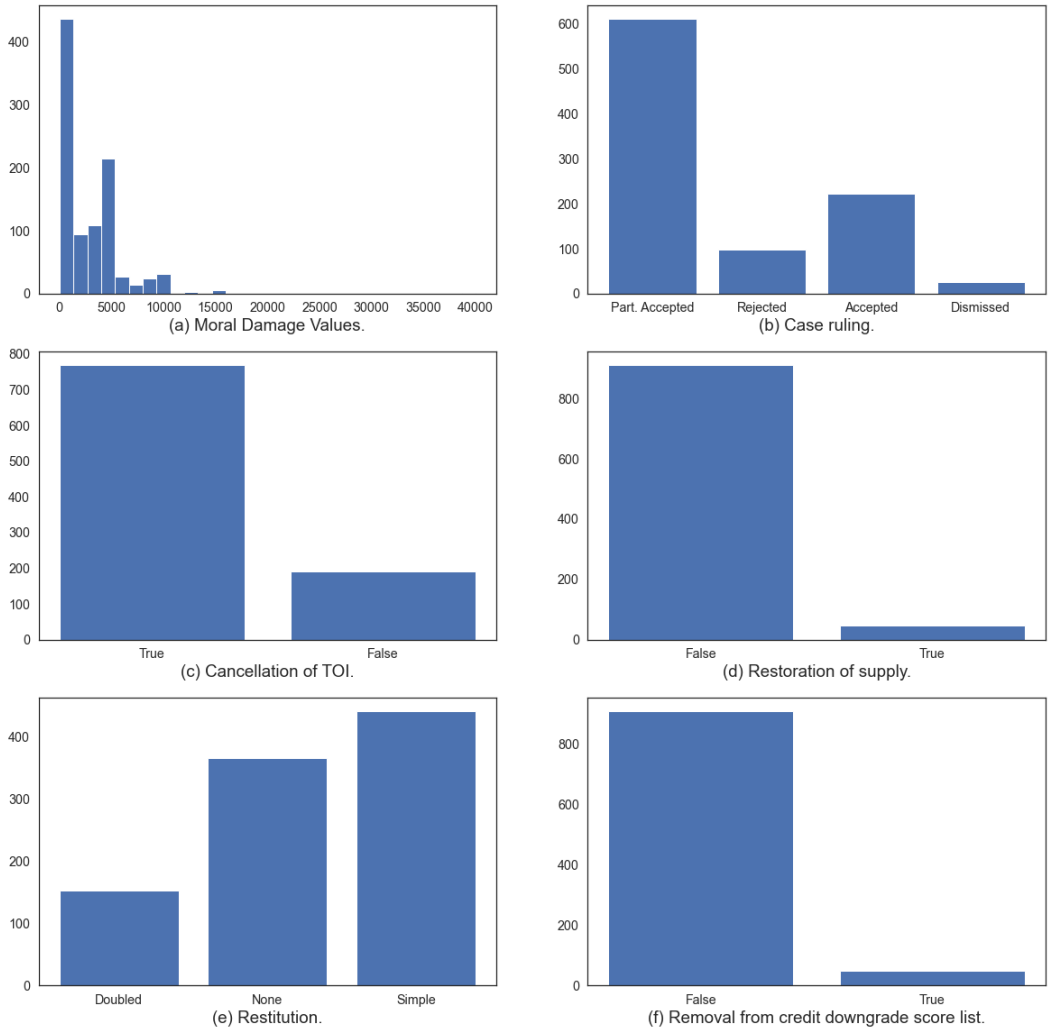


Figure 5.2: Dataset distribution.

5.1.2 Models

For Text Classification, six models are addressed based on different architectures and feature extraction methods. The TF-IDF Classifier, SIF Classifier and Doc2vec Classifier are based on the standard Text Classification framework illustrated by figure 4.1, using respectively TF-IDF, SIF and Doc2vec as the Feature Extraction models and Logistic Regression as the classifier. A MuDEC-based model (figure 4.2) is added to the experiments, applying Doc2vec for feature extraction, SMOTE for oversampling, k-Means for clustering and Logistic Regression for classification. The dimensionality reduction step was removed in both the standard and MuDEC frameworks. In addition, a C-LSTM Classifier is added to the experiments, following the original implementation described by figure 4.3.

Lastly, an Ensemble model based on MuDEC and C-LSTM Classifier is evaluated. Its inference process is based on the prediction probabilities.

The end prediction of a given instance is chosen as the one with the highest probability given by these models, as illustrated by figure 5.3. By doing so, the Ensemble model is expected to combine different approaches to optimize the classification task.

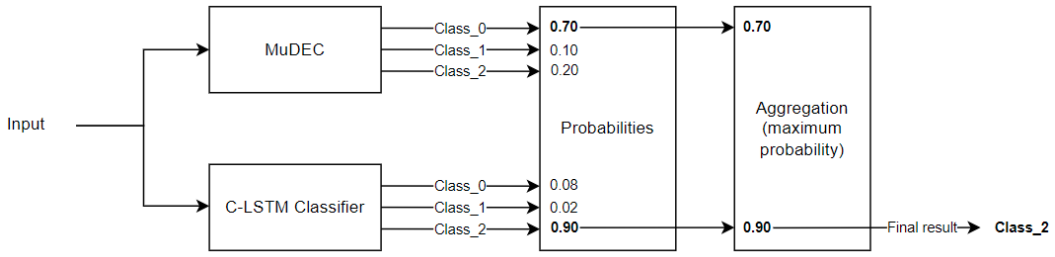


Figure 5.3: Ensemble Classifier inference process.

As a prior step for each model, common pre-processing routines are applied. This is a specially important step considering that legal opinions are structured differently from other domains. In this step, the text is lowercased and the punctuation, line breakers, and excessive spaces are removed. In addition, to minimize the task's complexity, the document is filtered to contain only the operative part of the judgment, where the lower or Appellate Court judge presents the judicial solution to the lawsuit, as illustrated by figure 5.4.

Figure 5.4: Illustration of a legal opinion structure (FERNANDES et al., 2022).

To identify this section, a list of regular expressions was used. The list is composed of expressions such as "given these considerations" and "in the face of the above" as a reference for splitting the document and removing the first section. If the filter fails to identify these expressions, the end section is obtained by considering the document's last m characters. For our experimental setup, $m = 2,500$ was empirically found to be the most suitable

value to comprise most of the end sections. After the identification of the end section and the removal of the remaining document, the stopwords are removed and the words are tokenized. Lastly, the tokens are stemmed to decrease the variety of expressions with different suffixes, resulting in a more simple representation of the text. The pre-processing step is therefore highly dependent on the type of legal document in question and must be adjusted accordingly for other contexts.

For the extraction of the numerical provisions (i.e. the moral damage value), NER models were implemented based on the framework described in section 2.3, exploring different options related to the use of CRF and the training approaches (Feature-based and Fine-tuning) and treating the numerical provision as an entity. The entities are extracted from the text, post-processed for the conversion to numerical values and evaluated based on the dataset annotations.

5.1.3 Optimization of hyperparameters

The models addressed for text classification were optimized based on their main hyperparameters, as shown in table 5.2.

Models	Hyperparameters
TF-IDF Classifier	Logistic Regression penalty, Logistic Regression optimization algorithm.
SIF Classifier	Vector dimension, embedding training epochs, Logistic Regression penalty, Logistic Regression optimization algorithm.
Doc2vec Classifier	Vector dimension, embedding training epochs, Logistic Regression penalty, Logistic Regression optimization algorithm.
MuDEC	Vector dimension, embedding training epochs, number of SMOTE nearest neighbours, number of clusters, Logistic Regression penalty, Logistic Regression optimization algorithm.
C-LSTM	Vector dimension, number of CNN filters, LSTM dimension, feature window size, classifier training epochs, embedding training epochs.

Table 5.2: Models hyperparameters

For the optimization process, a Bayesian Optimizer (SNOEK; LAROCHELLE; ADAMS, 2012) was used, where each model's hyperparameters are defined as input search dimensions for the optimization of the objective function. To improve the statistical significance of each evaluation

round, the objective function is defined based on a 10-fold stratified cross-validation setup, by averaging the negative F1-Scores weighted by support (the number of true instances for each label).

The number of iterations during optimization is chosen based on the search dimension complexity. SIF and Doc2vec bayesian optimization run for 200 iterations, while MuDEC, where clustering and SMOTE parameters are added, runs for 500 iterations. Although holding the same search dimension complexity compared to MuDEC, the optimization of the C-LSTM Classifier runs for only 100 iterations, once its high running time inhibits the usage of longer optimizations.

The NER models were implemented with default parameters taken from the original implementation (SOUZA; NOGUEIRA; LOTUFO, 2019).

5.2 Results

In this section, the main results related to the proposed models and experimental setup are detailed. The results are divided into two subsections, where the categorical features and the numerical features are addressed separately.

5.2.1 Extraction of categorical provisions

After the hyperparameters optimization, the models were applied to the dataset to evaluate their main metrics in a 10-fold stratified cross-validation setup. For comparison, the precision, recall, F1-score and accuracy were extracted, and their mean results are shown in Table 5.3, divided per each provision. The highlighted lines represent the best models per provision by their mean F1-Score.

When analyzing the results related to the case ruling provision, as illustrated by Figure 5.5, we can conclude that the C-LSTM model has outperformed TF-IDF, SIF, Doc2vec and MuDEC by a significant margin, reaching an F1-Score of 97.0%.

The concentration of errors between the partially accepted and accepted categories can be explained by a particularity of this provision in the text. The case ruling can be expressed as a summary of each plaintiff's claim. For instance, a rejection of the moral damage claim doesn't necessarily translates the entire case as rejected, since the remaining claims might have been accepted, turning the case ruling into partially accepted. The sequential and more complex representation of these expressions given by the C-LSTM model

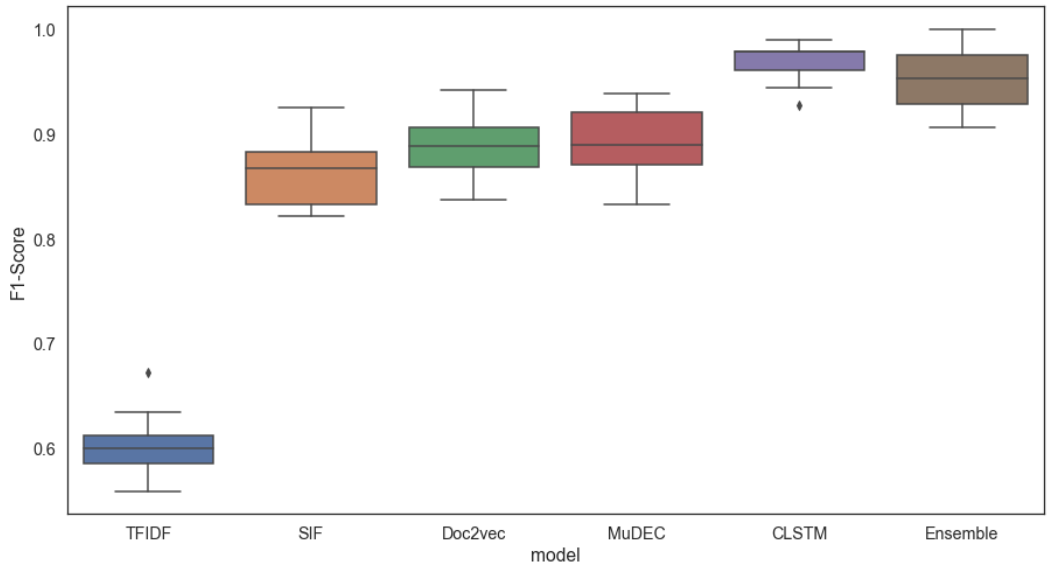


Figure 5.5: Cross-validation results per model for case ruling.

has shown to be a better fit for capturing the correct context in this case. The performance difference between C-LSTM and MuDEC is also expressed in the Ensemble results, where the addition of MuDEC to the C-LSTM model deteriorates the performance when compared to the C-LSTM itself.

The summarized test results in the cross-validation setup for the C-LSTM model are illustrated by the confusion matrix shown in figure 5.6, where the errors are spread around the "Partially Accepted" class, reinforcing the challenge of mapping individual claims rulings to the overall case ruling, as illustrated by figure 5.7.

Figure 5.6: C-LSTM summarized cross-validation confusion matrix for Case Ruling.

Ante o exposto, confirmo a decisão que deferiu o pedido de tutela e JULGO: a) PROCEDENTE O PEDIDO e DECLARO inexistente o parcelamento do TOI [redacted], cálculo foi realizado com base no art. 130, III, da Resolução nº 414/2010 da ANEEL, e faculto à Ré a possibilidade de emissão de novas faturas em substituição àquelas que foram emitidas com o parcelamento supramencionado, no prazo de 30 (trinta) dias, sob pena de perdimento do crédito, com vencimento em datas futuras, sendo que com um intervalo de 30 (trinta) dias entre os vencimentos de uma e outra para viabilizar o pagamento por parte do Demandante, devendo tais faturas serem emitidas com o novo parcelamento do TOI [redacted], cujo valor total do débito deverá ser de 284,5 kWh, observando os artigos 115, II e 113, I da Resolução nº 414/2010 da ANEEL, sem juros e correção monetária, corrigidos monetariamente, a contar do desembolso, e com juros legais de mora a contar da citação; b) PROCEDENTE O PEDIDO para condenar a Ré a devolver, em dobro, todos os valores comprovadamente pagos pela Demandante referentes ao parcelamento do TOI [redacted], que ultrapassaram o débito a ser recuperado de 284,5 kWh, corrigidos monetariamente, a contar do desembolso, e com juros legais de mora a contar da citação; c) IMPROCEDENTE O PEDIDO DE DANO MORAL. Custas processuais pela parte ré. Condeno ainda, a título de honorários advocatícios, em 10% do valor da condenação. Transitada em julgado, dê-se baixa e arquivem-se

Figure 5.7: Example of case ruling expressed individually by provisions.

The extraction of the provision "Cancellation of TOI" is once again best performed by the C-LSTM model, as shown in figure 5.8, achieving an F1-Score of 97.0%.

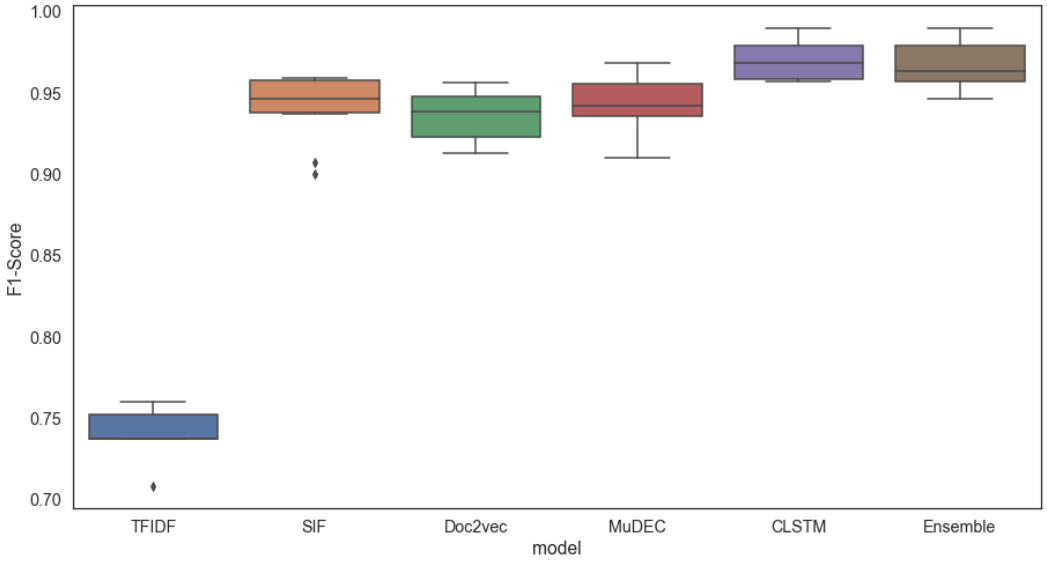


Figure 5.8: Cross-validation results per model for the cancellation of TOI.

Although C-LSTM has best performed in this task, the results are not as distant to the remaining models, in comparison with the previous provision. Different from the case ruling, the extraction of the cancellation of TOI doesn't require the summarizing of individual results for the final prediction, and the errors are concentrated in cases where the cancellation of TOI is expressed in more complex and ambiguous patterns, such as illustrated by figure 5.10. Despite this fact, C-LSTM is still able to best map the data complexity when compared to MuDEC, for instance. Similar to the previous provision, the Ensemble model was not able to take advantage of MuDEC and C-LSTM, mostly preserving the C-LSTM performance.

The summarized test results in the cross-validation setup for the C-LSTM model are illustrated by the confusion matrix shown in figure 5.9, which shows

a well-balanced error pattern.

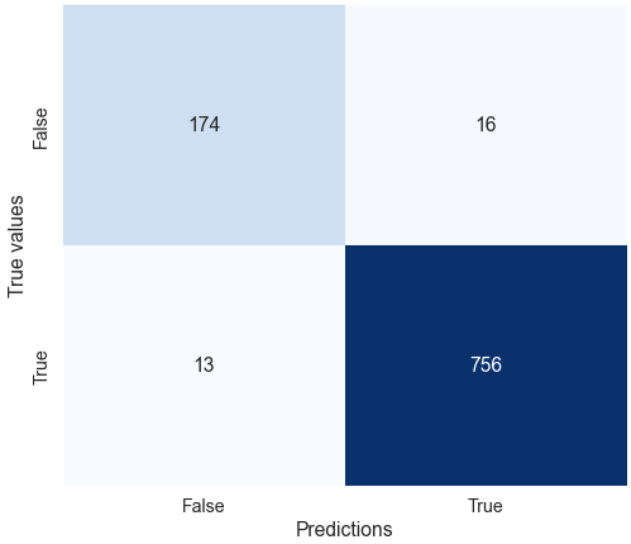


Figure 5.9: C-LSTM summarized cross-validation confusion matrix for the cancellation of TOI.

ISSO POSTO, julgo procedente a pretensão inicial para: a) anular o contrato de confissão de dívida objeto da lide (fls. 14); b) conceder a tutela antecipada na sentença, determinando o restabelecimento do serviço, no prazo de 48 (quarenta e oito) horas, sob pena de multa diária de R$200,00 (duzentos reais), ficando os efeitos da tutela antecipada condicionado a causa da interrupção como proveniente unicamente do inadimplemento do contrato de confissão de dívida desconstituído pela sentença, nos termos da fundamentação acima. Intime-se preferencialmente pela via eletrônica ou por OJA, conforme o caso, para cumprimento da tutela antecipada; c)

Pelo exposto, JULGO PROCEDENTE, para confirmar os efeitos da decisão de tutela antecipada, e desconstruir todo e qualquer valor aplicado referente ao TOI [redacted], bem como a restituir os valores pagos pela autora, com juros e correção monetária a contar do pagamento e a pagar danos morais equivalentes a R$ 6.000,00, com juros a contar de citação e correção monetária a contar da publicação desta sentença. Condeno a ré a pagar as despesas dos processos e os honorários advocatícios de 10% sobre o valor da condenação.

Figure 5.10: Example of complex expressions of cancellation of TOI.

The extraction of the provision "Restoration of Supply" is challenging due to its extreme imbalance. Although the Ensemble model has reached the highest F1-score, as shown in figure 5.11, the F1-Score of 96.7% can be misleading.

By analyzing the confusion matrix associated with the extraction of this provision by the Ensemble model (figure 5.12), it becomes clear that the errors are concentrated in the minority class. In summary, the model miss-classifies over 50% of the instances in this class, indicating the need for a large volume of training instances for effective implementation.

The results of the extraction of the provision "Restitution" (figure 5.13) show the best performance by the Ensemble model with an F1-Score of 97.4%, followed closely by C-LSTM. Once more, the C-LSTM seems to outperform MuDEC, Doc2vec, SIF and TF-IDF by at least 5%.

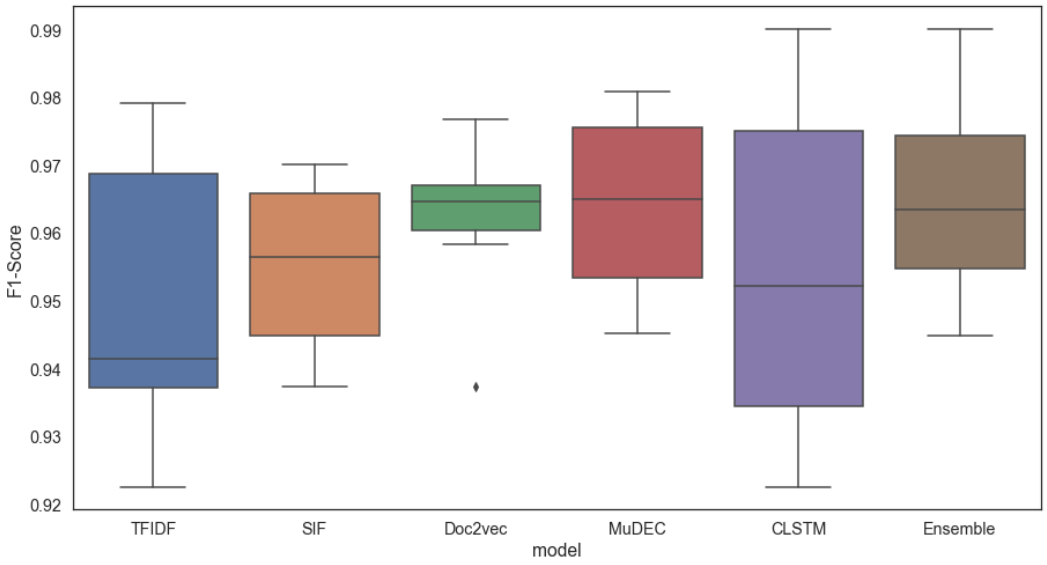


Figure 5.11: Cross-validation results per model for the restoration of supply.

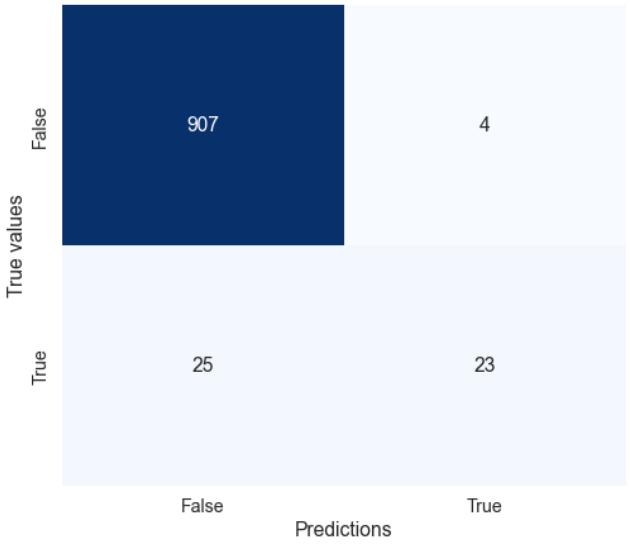


Figure 5.12: Ensemble summarized cross-validation confusion matrix for the restoration of supply.

In opposition to the restoration of supply, the provision "Restitution" presents a balanced class distribution, which results in better error distribution, as shown in figure 5.14.

The results related to the extraction of "Removal from Credit Downgrade Score List" show similar performance within the set of models (figure 5.15). The best F1-Score is achieved by MuDEC (97.7%), followed closely by the Ensemble and Doc2vec.

Similar to the restoration of supply, this provision presents an unbalanced class distribution, leading to poor results when analyzed by class, as shown in figure 5.16. Although the results are considerably better when compared

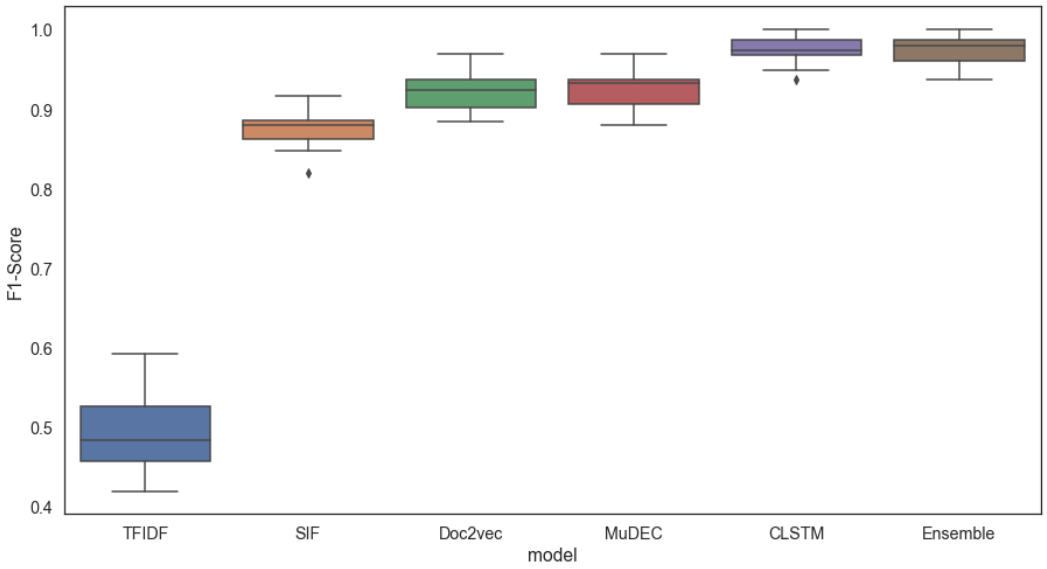


Figure 5.13: Cross-validation results per model for the restitution.

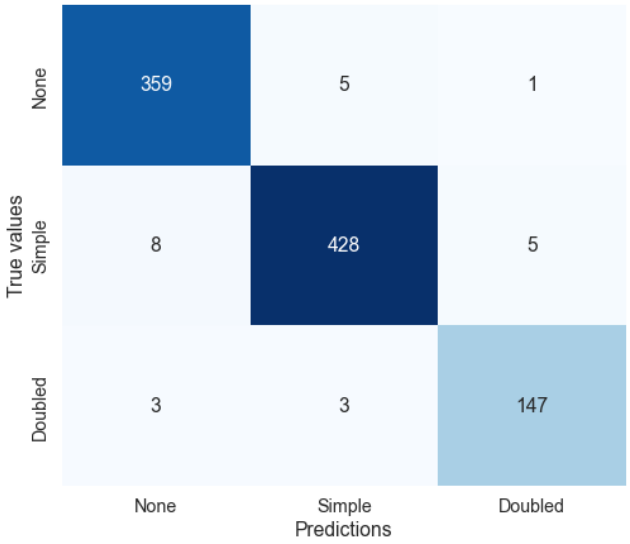


Figure 5.14: Ensemble summarized cross-validation confusion matrix for the restitution.

to Restoration of Supply, the confusion matrix indicates the need for more training instances from the minority class for more effective results.

5.2.2 Extraction of numerical provision

The extraction of the numerical provision (moral damage compensation) by the BERT model for NER was implemented by four different approaches. The first, named BERT, uses a Fine-tuning approach (i.e. updating all weights jointly and using a linear layer as classifier) without the use of a CRF layer. The second (BERT-CRF) is similar to the first, with the addition of a CRF

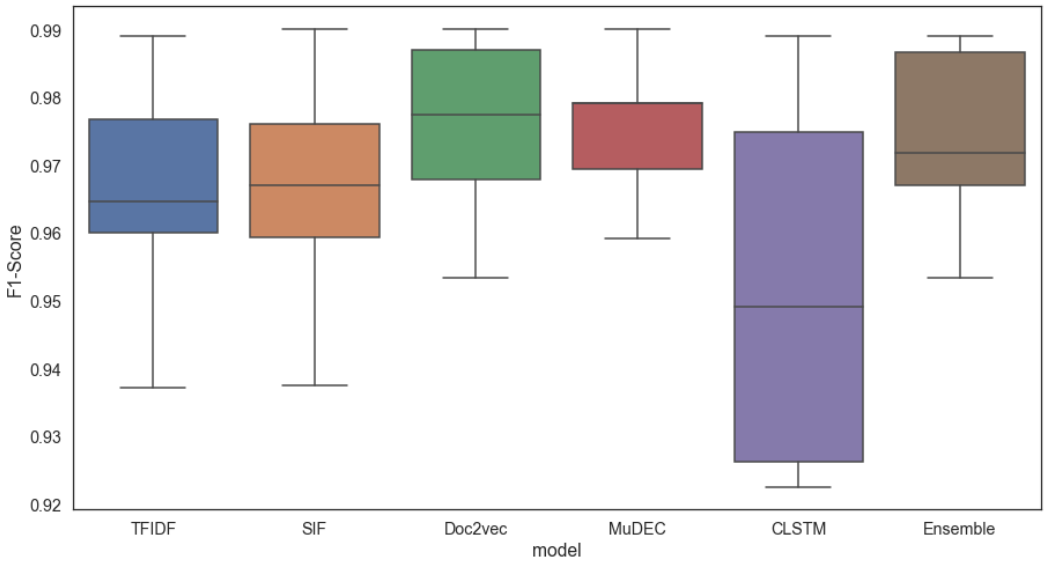


Figure 5.15: Cross-validation results per model for the removal from credit downgrade score list.

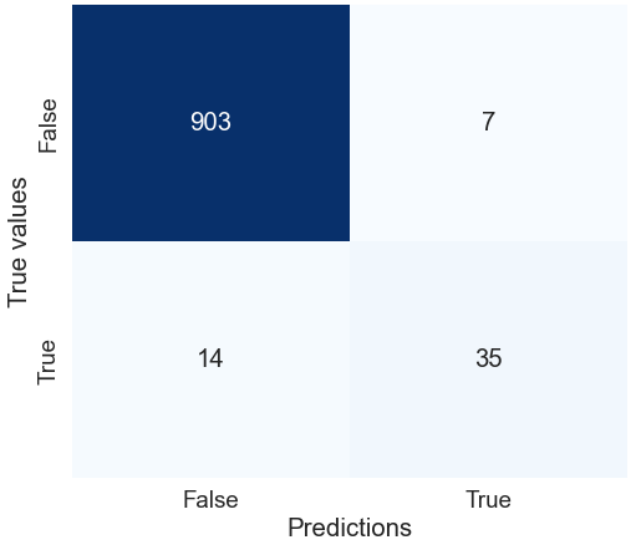


Figure 5.16: Ensemble summarized cross-validation confusion matrix for the removal from credit downgrade Score list.

layer. The third model, named BERT-LSTM uses a Feature-based approach (i.e freezing the BERT weights and using a BiLSTM as the classifier) without a CRF layer, which is added to the fourth model (BERT-LSTM-CRF).

The 10-fold cross-validation accuracies related to each model are illustrated by figure 5.17. Table 5.4 shows the mean results when evaluating both accuracy and Root Mean Square Error (RMSE).

The results demonstrate the effectiveness of this approach for the extraction of the Moral Damage value from legal opinions. The mean accuracies range from 96.2% and 98.9% within the different model approaches. As a natural result, the use of the CRF layer enhances the performance by around

Provision	Model	Precision	Recall	F1-score	Accuracy
Case ruling	TF-IDF	0.707	0.691	0.604	0.691
	SIF	0.872	0.870	0.863	0.870
	Doc2vec	0.893	0.893	0.890	0.893
	MuDEC	0.911	0.907	0.907	0.907
	C-LSTM	**0.969**	**0.970**	**0.968**	**0.970**
	Ensemble	0.955	0.953	0.953	0.953
Cancellation of TOI	TF-IDF	0.829	0.813	0.741	0.813
	SIF	0.942	0.942	0.941	0.942
	Doc2vec	0.939	0.938	0.936	0.938
	MuDEC	0.944	0.943	0.943	0.943
	C-LSTM	**0.970**	**0.970**	**0.970**	**0.970**
	Ensemble	0.968	0.968	0.967	0.968
Restoration of supply	TF-IDF	0.946	0.961	0.949	0.961
	SIF	0.959	0.963	0.955	0.963
	Doc2vec	0.964	0.969	0.963	0.969
	MuDEC	0.967	0.965	0.964	0.965
	C-LSTM	0.953	0.961	0.954	0.961
	Ensemble	**0.969**	**0.968**	**0.967**	**0.968**
Restitution	TF-IDF	0.618	0.556	0.492	0.556
	SIF	0.887	0.875	0.874	0.875
	Doc2vec	0.929	0.925	0.925	0.925
	MuDEC	0.930	0.924	0.925	0.924
	C-LSTM	0.974	0.973	0.973	0.973
	Ensemble	**0.975**	**0.974**	**0.974**	**0.974**
Removal from credit downgrade score list	TF-IDF	0.969	0.971	0.966	0.971
	SIF	0.968	0.969	0.968	0.969
	Doc2vec	0.977	0.977	0.976	0.977
	MuDEC	**0.978**	**0.978**	**0.977**	**0.978**
	C-LSTM	0.945	0.965	0.952	0.965
	Ensemble	0.975	0.976	0.974	0.976

Table 5.3: Results for the classification of categorical provisions.

Model	Training approach	Accuracy	RMSE
BERT	Fine-tuning	0.962	825.8
BERT-CRF	Fine-tuning	0.982	639.8
BERT-LSTM	Feature-based	0.954	738.6
BERT-LSTM-CRF	**Feature-based**	**0.989**	**277.2**

Table 5.4: Results for the extraction of moral damage compensations.

2%, indicating the effectiveness of the contextual information captured by the CRF. Interestingly, the Feature-based approach outperforms the Fine-tuning approach by 0.75% on average. This result possibly indicates the quality of

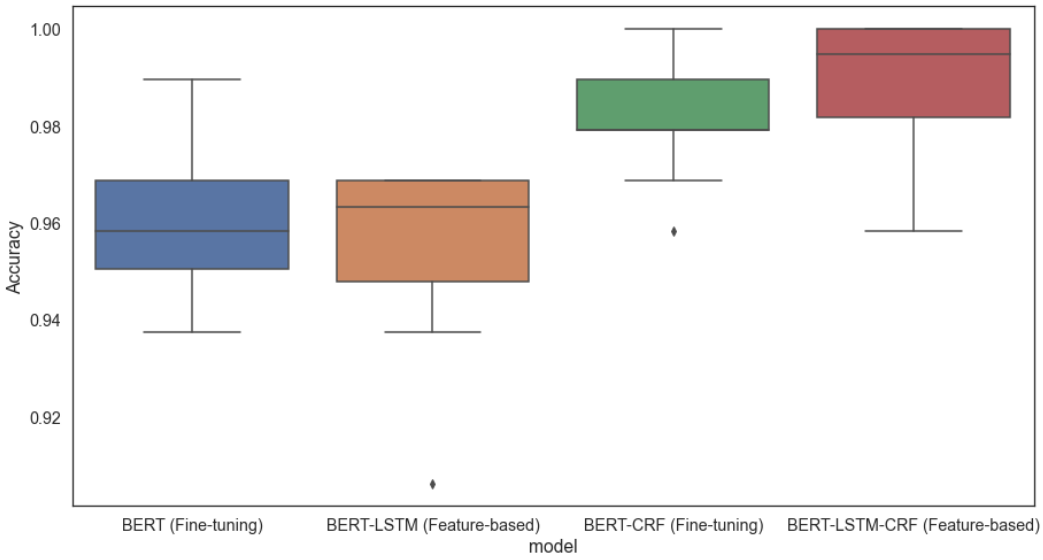


Figure 5.17: Cross-validation results per model for moral damage compensations.

BERT embeddings, achieving better results when their weights are frozen, which is unexpected, given the specificity of the Legal Domain context.

Figure 5.18 shows the summarized confusion matrix for the BERT-LSTM-CRF model. For simplicity, the results are rounded to the nearest thousand (in Brazilian reais).

With an accuracy of 98.9% and an RMSE of 277.2 Brazilian reais, the extraction of the moral damage compensation by the BERT-LSTM-CRF model in a Feature-based approach reaches promising results for the given task.

In fact, the only errors are related to the value zero, which is assigned by the model only if no moral damage entity is found in the text. These errors result from two main sources: (i) The identification of other numeric entities (such as the legal fees) as the moral damage compensations when no moral damage is present in the legal opinion (figure 5.19) and (ii) The failure to identify any value at all when the moral damage is present in the legal opinion (figure 5.20).

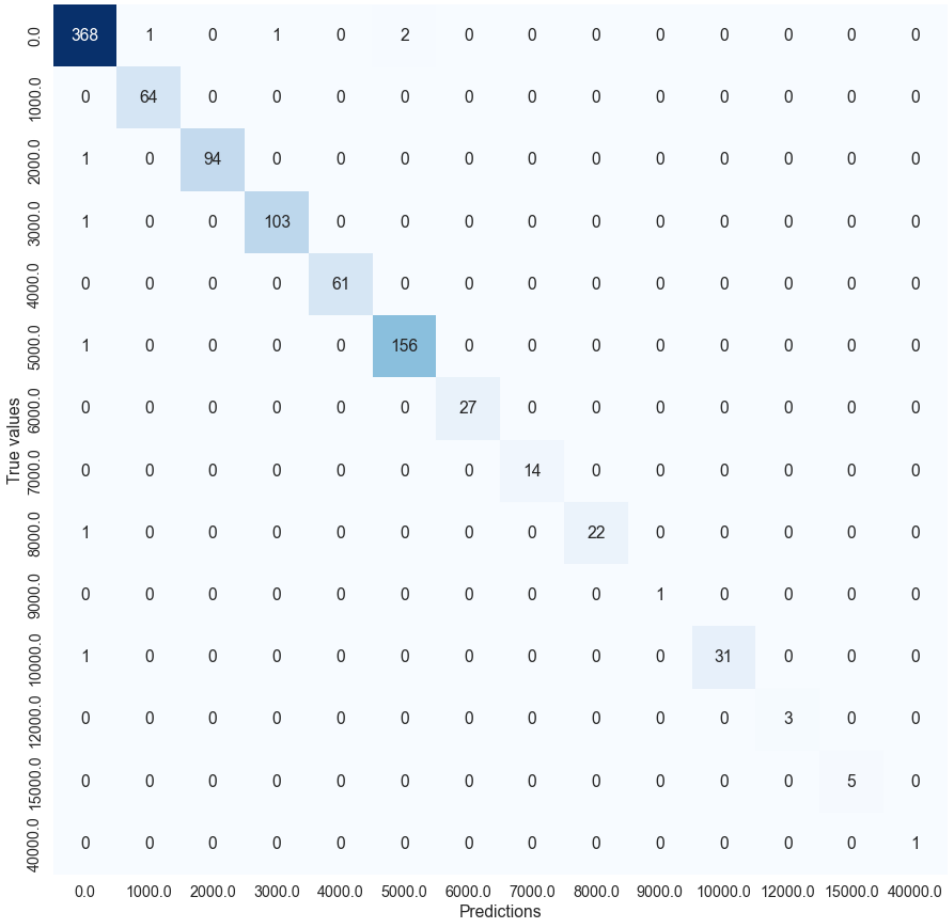


Figure 5.18: Summarized cross-validation confusion matrix for the moral damage compensation.

em R$500,00 (quinhentos reais), na forma do art. 85, §8°, do CPC. Condeno a parte ré nos honorários advocatícios devidos ao patrono da parte autora, fixados em R$1.000,00 (um mil reais), na forma do art. 85, §8°, do CPC. Na cobrança das despesas processuais e honorários advocatícios deverá ser observada a gratuidade de justiça deferida à parte autora (fl. 44). P.R.I. Transitada em julgado e sem incidentes, dê-se baixa e arquivem-se os autos, observando o art. 229 A da Consolidação Normativa.

Figure 5.19: Example of wrong assignment of moral damage value by the model.

Isto posto JULGO PROCEDENTE O PEDIDO, ao teor do artigo 487 I do CPC, desconstituo o TOI de número [redacted] condeno a ré a indenizar a autora a título de dano moral na importância de 8.000,00 , quantia esta devidamente corrigida a contar desta data e acrescida de juros de 1% ao mês contados da citação. Condeno a ré a se abster de suspender o fornecimento de energia elétrica no domicilio da autora em razão do TOI que ora se desconstitui. Condeno a ré em se abster em inserir o nome da autora no cadastro restritivo de credito autora em razão do TOI que ora se desconstitui. Quanto as obrigações de não fazer fixo multa diária de 500,00. Condeno a ré nas despesas processuais e nos honorários advocatícios que fixo em 10% sobre o valor da condenação.

Figure 5.20: Example of model's fail to identify the moral damage value.

6
Conclusion and Future Work

In this work, we explored the field of Information Extraction in the Brazilian Legal Domain by addressing the extraction of provisions from legal opinions in Brazilian Portuguese. Furthermore, we have divided this task into two sub-tasks: (i) Text Classification for extraction of categorical provisions and (ii) Named Entity Recognition for the extraction of numerical provisions. In the first, we explored multiple models, by using different frameworks and Text Feature Extraction methods. In the second, we used a NER model based on BERT embeddings, and analyzed the impact of the CRF layer, as well as different training approaches (Feature-based and Fine-tuning).

For evaluation, we constructed a dataset containing 964 manually annotated legal opinions, and a Bayesian Optimizer was used for hyperparameter optimization by defining the objective function as the negative cross-validation mean F1-Score. By choosing the best models found during evaluation, we achieved mean accuracies higher than 96% at the extraction of each provision. More specifically, we achieved an accuracy of 97.0% for the extraction of the case ruling, 97.0% for the cancellation of TOI, 96.8% for the restoration of supply, 97.4% for the restitution, 97.8% for the removal from credit downgrade score list and 98.9% for the moral damage compensation.

When analyzing the Text Classification methods, the C-LSTM model has outperformed at the extraction of two provisions, as well as the Ensemble, while MuDEC had the best performance in one. The analysis of NER models applied to the extraction of the moral damage compensations shows the positive impact of the CRF layer, which increases accuracy by around 2.0% on average. The Feature-based approach achieves a slightly higher accuracy when compared to the Fine-tuning approach, showing the effectiveness of the BERT embeddings used, especially considering the particularity of the terms used in the legal domain.

Despite the high accuracies achieved, it is important to notice the large imbalance present in the dataset when categorized by the provisions: the restoration of supply and the removal from credit downgrade score list. When analyzing the individual accuracies related to the minority classes in these cases, the results are far below the mean results, reaching only around 50% and 70%, turning these models highly biased towards the majority classes. This indicates the need for larger amounts of training instances belonging to the minority classes for effective extraction of these provisions.

An early version of this study has been presented at the 24^{th} International Conference on Enterprise Information Systems (ICES) and was awarded as the best student paper in the field of Artificial Intelligence and Decision Support Systems. In addition, it received an invitation for applying an extended version of the study to the Springer Nature Computer Science Journal.

The most direct contribution from this work is the development of a highly accurate tool for extracting provisions from legal opinions in the given context. In addition, the proposed framework can be adapted to other tasks, by simply providing an annotated dataset for the optimization and evaluation of the models, enabling the extraction of other categorical and numerical information from documents. Furthermore, the conclusion related to the effectiveness of the C-LSTM is also a contribution, especially when compared to the previous results presented by (COELHO. et al., 2022), where the MuDEC framework was shown to outperform the remaining models. The main differences from the previous work are the increase in the dataset size (from 193 to 964 instances) and the training process related to word2vec embeddings, which are performed exclusively in the given dataset, in oppose to the previous work, where pre-trained embeddings are used. This shows the challenge of reusing word2vec models in different contexts.

In the next steps, the low performance associated with the minority classes of imbalanced categories should be addressed. A possible solution could be based on the use of semi-supervised methods to decrease the costs associated with manual annotations. Once most provisions are related to a small variation of specific terms, synthetic annotations based on similarity might be suitable in this context. In addition, text classification can benefit from BERT word and sentence embeddings when applied to the feature extraction step. As shown by the NER results, BERT has the potential to achieve high-performance results with pre-trained embeddings, even when applied to a specific context, such as the legal domain.

Finally, the methods explored can be applied for the extraction of other categorical and numerical provisions, expanding to other legal contexts, and possibly to other relevant legal features, such as the facts and plaintiffs' claims. Hopefully, the extraction of these features from legal documents can provide valuable structured information to AI-based applications, adding value to the justice system.

7
Bibliography

ARAUJO, P. H. L. D. et al. Victor: a dataset for brazilian legal documents classification. In: **Proceedings of the 12th Language Resources and Evaluation Conference**. [S.l.: s.n.], 2020. p. 1449–1458.

ARAUJO, P. H. Luz de et al. Lener-br: a dataset for named entity recognition in brazilian legal text. In: SPRINGER. **International Conference on Computational Processing of the Portuguese Language**. [S.l.], 2018. p. 313–323.

ARORA, S.; LIANG, Y.; MA, T. A simple but tough-to-beat baseline for sentence embeddings. In: **International conference on learning representations**. [S.l.: s.n.], 2017.

CHALKIDIS, I.; KAMPAS, D. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. **Artificial Intelligence and Law**, Springer, v. 27, n. 2, p. 171–198, 2019.

CHAWLA, N. V. et al. Smote: synthetic minority over-sampling technique. **Journal of artificial intelligence research**, v. 16, p. 321–357, 2002.

COELHO., G. et al. Text classification in the brazilian legal domain. In: INSTICC. **Proceedings of the 24th International Conference on Enterprise Information Systems - Volume 1: ICEIS,**. [S.l.]: SciTePress, 2022. p. 355–363. ISBN 978-989-758-569-2. ISSN 2184-4992.

CORREIA, F. A. et al. Fine-grained legal entity annotation: A case study on the brazilian supreme court. **Information Processing & Management**, Elsevier, v. 59, n. 1, p. 102794, 2022.

DEVLIN, J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

FERNANDES, W. P. D. et al. Extracting value from brazilian court decisions. **Information Systems**, Elsevier, v. 106, p. 101965, 2022.

FERNANDES, W. P. D. et al. Appellate court modifications extraction for portuguese. **Artificial Intelligence and Law**, Springer, v. 28, n. 3, p. 327–360, 2020.

HARRIS, Z. S. Distributional structure. **Word**, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954.

JIANG, J. Information extraction from text. In: **Mining text data**. [S.l.]: Springer, 2012. p. 11–41.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, MCB UP Ltd, 1972.

KOWSARI, K. et al. Text classification algorithms: A survey. **Information**, Multidisciplinary Digital Publishing Institute, v. 10, n. 4, p. 150, 2019.

LE, Q.; MIKOLOV, T. Distributed representations of sentences and documents. In: PMLR. **International conference on machine learning**. [S.l.], 2014. p. 1188–1196.

LEITNER, E.; REHM, G.; MORENO-SCHNEIDER, J. Fine-grained named entity recognition in legal documents. In: SPRINGER. **International Conference on Semantic Systems**. [S.l.], 2019. p. 272–287.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. **arXiv preprint arXiv:1301.3781**, 2013.

MINAEE, S. et al. Deep learning–based text classification: a comprehensive review. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 54, n. 3, p. 1–40, 2021.

PENNINGTON, J.; SOCHER, R.; MANNING, C. D. Glove: Global vectors for word representation. In: **Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)**. [S.l.: s.n.], 2014. p. 1532–1543.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. Practical bayesian optimization of machine learning algorithms. **Advances in neural information processing systems**, v. 25, 2012.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Portuguese named entity recognition using bert-crf. **arXiv preprint arXiv:1909.10649**, 2019.

SULEA, O.-M. et al. Exploring the use of text classification in the legal domain. **arXiv preprint arXiv:1710.09306**, 2017.

SURDEN, H. Artificial intelligence and law: An overview. **Ga. St. UL Rev.**, HeinOnline, v. 35, p. 1305, 2018.

WEI, F. et al. Empirical study of deep learning for text classification in legal document review. In: IEEE. **2018 IEEE International Conference on Big Data (Big Data)**. [S.l.], 2018. p. 3317–3320.

XIAO, C. et al. Cail2019-scm: A dataset of similar case matching in legal domain. **arXiv preprint arXiv:1911.08962**, 2019.

YADAV, V.; BETHARD, S. A survey on recent advances in named entity recognition from deep learning models. **arXiv preprint arXiv:1910.11470**, 2019.

ZHONG, H. et al. Legal judgment prediction via topological learning. In: **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing**. [S.l.: s.n.], 2018. p. 3540–3549.

ZHONG, H. et al. How does nlp benefit legal system: A summary of legal artificial intelligence. **arXiv preprint arXiv:2004.12158**, 2020.

ZHOU, C. et al. A c-lstm neural network for text classification. **arXiv preprint arXiv:1511.08630**, 2015.