Lucas Emanuel Resck Domingues

Inferring and Explaining Potential Citations to Binding Precedents in Brazilian Supreme Court Decisions

Rio de Janeiro, Brazil

2021

Lucas Emanuel Resck Domingues

Inferring and Explaining Potential Citations to Binding Precedents in Brazilian Supreme Court Decisions

Undergraduate Thesis presented to the School of Applied Mathematics as a partial requirement for obtaining a Bachelor's degree in Applied Mathematics.

Getulio Vargas Foundation – FGV

School of Applied Mathematics

Applied Mathematics Undergraduate Course

Supervisor: Jorge Luis Poco Medina

Rio de Janeiro, Brazil

2021

Lucas Emanuel Resck Domingues
Inferring and Explaining Potential Citations to Binding Precedents in Brazilian Supreme Court Decisions/ Lucas Emanuel Resck Domingues. – Rio de Janeiro, Brazil, 2021-
82p. : il. (algumas color.) ; 30 cm.

Supervisor: Jorge Luis Poco Medina

Thesis (Undergraduate) – Getulio Vargas Foundation – FGV
School of Applied Mathematics
Applied Mathematics Undergraduate Course, 2021.

1. Machine learning. 2. Natural language processing. 3. Machine learning explainability. 4. Legal documents. I. Jorge Luis Poco Medina. II. Getulio Vargas Foundation. III. School of Applied Mathematics. IV. Inferring and Explaining Potential Citations to Binding Precedents in Brazilian Supreme Court Decisions

LUCAS EMANUEL RESCK DOMINGUES

"INFERRING AND EXPLAINING POTENTIAL CITATIONS TO BINDING PRECEDENTS IN BRAZILIAN SUPREME COURT DECISIONS"

Trabalho de Conclusão de Curso - TCC apresentado ao Curso de Graduação em Matemática Aplicada da Escola de Matemática Aplicada para obtenção do grau de Bacharel (a) em Matemática Aplicada.

Data da Defesa: 08/12/2021

ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA

Jorge Luís Poco Medina
Orientador

Renato Rocha Souza
Membro

Elias Jacob de Menezes Neto
Membro

Nos termos da Lei nº 13.979 de 06/02/20 - DOU nº 27 de 07/02/20 e Portaria MEC nº 544 de 16/06/20 - DOU nº 114 de 17/06/20 que dispõem sobre a suspensão temporária das atividades acadêmicas presenciais e a utilização de recursos tecnológicos face ao COVID-19, as apresentações dos Trabalhos de Conclusão de Curso, de forma excepcional, serão realizadas de forma remota e síncrona, incluindo-se nessa modalidade membros da banca e discente.

Acknowledgements

I want to thank the Brazilian Mathematics Olympiad of Public and Private Schools (OBMEP) for the first opportunities it opened in my academic life; the Getulio Vargas Foundation (FGV), the Center for the Development of Mathematics and Sciences (CDMC) of FGV, and the School of Applied Mathematics (EMAp) of FGV for the opportunities they opened for my college and graduate academic life, including scholarship; the other authors of the original work on which this one was based — my supervisor Jorge Poco, Jean R. Ponciano, and Luis Gustavo Nonato; all the professors and teachers who have gone through my academic life; and my mother, who never measured efforts for my education.

Abstract

The Brazilian Supreme Court (STF) is the highest law court in Brazil and it is primarily responsible for guarding the Brazilian Constitution. To reduce judicial insecurity and the high Court's workload, a Constitutional Amendment from 2004 allowed STF to create binding precedents ("Súmulas Vinculantes," BPs). A BP is a statement that consolidates the understanding of STF about a legal matter and has mandatory application for lower branches of the Judiciary. Frequently, an STF Justice cites a BP in a decision, and it is trivial to search for these explicit citations using regular expressions. However, it is not trivial to assert whether a decision *potentially* cites the statement, in the sense of "it should have cited it, but it did not" or "it addresses a similar issue, so they are related." This work explores machine learning and natural language processing (NLP) algorithms to infer and explain these *potential* citations. The inference is performed using models from classical machine learning theory and recent NLP research, and the explanation is achieved using a machine learning explainability technique. The models learn what characterizes a citation through training on documents with explicit citations, in which we demonstrate they achieve high performance. We present two case studies that demonstrate the usefulness of the trained models to search for potential citations when accompanied by the explainability technique to inform the most relevant parts of the document for the potential citation assignment.

Keywords: machine learning. natural language processing. machine learning explainability. legal documents.

Resumo

O Supremo Tribunal Federal (STF) é a corte de maior instância no Brasil e é primariamente responsável por guardar a Constituição Brasileira. Para reduzir a insegurança jurídica e a alta carga de trabalho da Corte, uma Emenda Constitucional de 2004 permitiu que o STF criasse Súmulas Vinculantes (SV). Uma SV é um enunciado que consolida o entendimento do STF sobre algum assunto jurídico e possui aplicação obrigatória pelas instâncias inferiores do Poder Judiciário. Frequentemente, um Ministro do STF cita uma SV em uma decisão, e é trivial procurar por essas citações explícitas usando expressões regulares. Porém, não é trivial verificar se uma decisão cita *potencialmente* o enunciado, no sentido de que "ele deveria ter sido citado, mas não o foi" ou "ela aborda um assunto similar, então eles estão relacionados". Este trabalho explora algoritmos de aprendizado de máquina e de processamento de linguagem natural (PLN) para inferir e explicar essas citações *potenciais*. A inferência é realizada usando modelos da teoria clássica de aprendizado de máquina e da pesquisa recente em PLN, e a explicação é alcançada utilizando uma técnica de explicabilidade em aprendizado de máquina. Os modelos aprendem o que caracteriza uma citação através do treinamento com documentos que possuem citações explícitas, no qual nós demonstramos que eles atingem alta performance. Nós também apresentamos dois estudos de caso que demonstram a utilidade dos modelos treinados em buscar por citações potenciais quando acompanhados pela técnica de explicabilidade para informar as partes mais relevantes do documento para a atribuição da citação potencial.

Palavras-chave: aprendizado de máquina. processamento de linguagem natural. explicabilidade em aprendizado de máquina. documentos jurídicos.

Contents

1 Introduction

The Brazilian Supreme Court ("Supremo Tribunal Federal," in Portuguese, or just STF) is the highest law court in Brazil and is primarily responsible for guarding the Brazilian Constitution, according to the Constitution itself. It is composed of eleven Justices, Brazilians, appointed by the President and approved by the Brazilian Federal Senate.

The STF produces a large number of decisions during its functioning. For instance, it published more than 1 million documents between 2011 and 2020 (BRAZILIAN SUPREME COURT, 2011-2020). In fact, this situation is spread all over the Brazilian Judiciary, and the adoption of appropriated mechanisms is necessary. One is the "precedent," when new judicial decisions are taken based on similar, old decisions. When there are several precedents about the same subject, a court can formalize this understanding in what is called a "súmula": "a statement prepared by a court that represents its guidance," which "constitutes a means to standardize judicial decisions through a consolidated understanding" (TAVARES, 2007; TUCCI, 2004, free translation apud DIMOULIS; LUNARDI, 2016). However, a súmula application is not mandatory.

When similar cases have different decisions, it can lead to judicial insecurity. Associated with the fact that Brazil had more than 100 million pending cases in 2004 (CONSELHO NACIONAL DE JUSTIÇA, 2004), a Constitutional Amendment from the same year allowed STF to create binding precedents[1] ("Súmulas Vinculantes," BPs). Like ordinary súmulas, these BPs consolidate the understanding of a court, STF in this case. However, they have a mandatory application (they are binding) for lower branches of the Judiciary. From 2004 to 2020, the STF ruled 58 binding precedents, covering many judicial discussions. For example, BP 37, which addresses the issue of salaries of public servants, has the following statement: *"It is not up to the Judiciary, which has no legislative function, to increase the salaries of public servants on the grounds of isonomy."*

Frequently, STF Justices cite a BP in their decisions. Because the number of decisions is large, the analysis of these documents is hard for judicial experts. If one wants to find a BP citation inside an STF decision, it is trivial to search it using regular expressions. However, it is not trivial to assert whether a decision *potentially* cites a precedent, in the sense of "it should have cited it, but it did not," or "it addresses a similar issue, so they are related." With this situation in mind, this work explores machine learning and natural language processing (NLP) models to infer *potential* citations. Because

[1] We point that binding precedent is not a precedent. A usual translation for "Súmula Vinculante" would be "binding súmula," however, in the absence of a more suitable translation for "súmula," we chose the word "precedent."

trusting a model in a sensitive context (*e.g.*, legal documents) is not straightforward, we also explore a machine learning explainability technique to understand the inference of potential citations. More specifically, this work aims to:

- Collect, analyze, and process documents related to STF decisions;
- Utilize machine learning and NLP algorithms and techniques to infer potential citations to binding precedents through their training on legal documents;
- Apply a machine learning explainability technique to understand the models' decisions. We want to answer, for instance, "which parts of the document were most important for the inference of the potential citation?"

The context of this work is the development of a methodology of identifying and explaining potential citations in legal documents for integrating a visual analytics system designed to help judicial experts analyze legal data. The corresponding work is entitled *LegalVis: Exploring and Inferring Precedent Citations in Legal Documents*, with authors *Lucas E. R. Domingues, Jean R. Ponciano, Luis Gustavo Nonato, and Jorge Poco*, and is under peer review.

Part I

Theoretical references

2 Document embedding

Many machine learning algorithms need fixed-length vectors as input to do calculations (LE; MIKOLOV, 2014). Therefore, to apply machine learning models for textual data, we need to represent this data in a vector format. In this context, *text embedding* can be defined as "fixed-length vector representation for texts" (LE; MIKOLOV, 2014), which can be words (MIKOLOV et al., 2013), sentences, paragraphs, or entire documents (CER et al., 2018; LE; MIKOLOV, 2014; BELTAGY; PETERS; COHAN, 2020). The idea is to have a real vector representing a text, with its features "potentially describing syntactic or semantic properties" (LEBRET; COLLOBERT, 2014). We refer to *document embeddings* when dealing with more-than-a-word texts, *e.g.*, sentences, paragraphs, and documents.

2.1 TF-IDF

TF-IDF (Term Frequency, Inverse Document Frequency) is one of the most simple document embeddings techniques. Suppose there is a corpus (a collection of documents) of N documents and one wants to vectorize document j. We can vectorize document j assigning a score for each word in the corpus relatively to this document. If the frequency of a word i in the document j is f_{ij} and this word appears in n_i of the N documents of the corpus, the score between word i and document j is calculated as $\text{TF}_{ij} \times \text{IDF}_i$, where

$$\text{TF}_{ij} = \frac{f_{ij}}{\max_k f_{kj}}, \quad \text{IDF}_i = \log_2\left(\frac{N}{n_i}\right)$$

(LESKOVEC; RAJARAMAN; ULLMAN, 2020). In the end, all documents will have been assigned to a real nonnegative vector with a dimension equal to the vocabulary size (the number of unique words in the corpus). Intuitively, each document embedding carries information about the most common words in it (the TF part) but privileges words that appear only in a few documents (rare words, the IDF part). It is important to note that popular implementations of TF-IDF vectorization, *e.g.*, scikit-learn's[1], do not necessarily follow the same calculations, but the intuitions are the same.

To fix ideas, let's consider an example. Suppose we have a corpus of three documents only:

1. "My cats sleep all day."
2. "Cats are beautiful."

[1] Interested readers can refer to https://scikit-learn.org/stable/modules/feature_extraction.html#text-feature-extraction.

3. "Dogs are loyal."

The calculation (with scikit-learn) of the TF-IDF scores results in Table 1. The columns represent the vocabulary, and the rows represent the individual documents. Cell (i, j) represents the score between document i and word j. Therefore, each row is the embedding of the respective document. The greatest score is *beautiful*'s for document 2 because it is rare (it appears in one document only) and very frequent in this document (this document has few words[2]). On the other hand, the lowest score (different from zero) is *cats*' in document 1 because this word is not rare (it appears in two documents), but also not frequent in this document (the document has many words).

Table 1 – Example of TF-IDF table.

	all	**are**	**beautiful**	**cats**	**day**	**dogs**	**loyal**	**my**	**sleep**
1	0.47	0	0	0.32	0.47	0	0	0.47	0.47
2	0	0.49	0.73	0.49	0	0	0	0	0
3	0	0.43	0	0	0	0.64	0.64	0	0

Each row is the TF-IDF embedding of the respective document. Calculated with scikit-learn. Source: the authors.

In the following sections, we explore other, more complex document embedding techniques, which try to solve TF-IDF problems and incorporate more sophisticated calculations when building a document's representation vector.

2.2 Doc2Vec

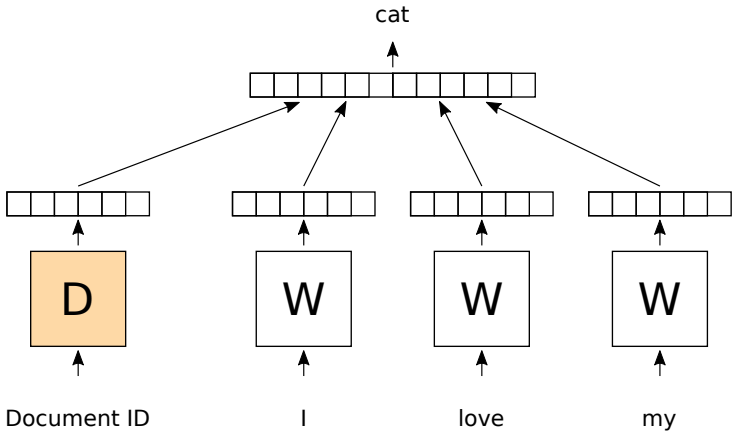
When using the TF-IDF technique, two different documents that share the exact words have the same document embeddings, even if the word order or the underlying semantics are different. For instance, changing the document "Dogs are loyal." to "Are dogs loyal?" results in a completely different meaning, but both documents have the same TF-IDF vector. As noticeable in Table 1, TF-IDF vectors also suffer from sparsity (few coordinates different from zero) and high dimensionality (one dimension per word in the vocabulary). For this reason, researchers proposed "Paragraph Vectors" (LE; MIKOLOV, 2014), usually called "Doc2Vec" embedding (in reference to the transformation of a document to a vector). Although it is named "paragraph," one can use this model to obtain encoding for any piece of text, *e.g.*, sentences, paragraphs, and documents. Intuitively, when vectorizing a document, it considers the present words and the context in which those words are inserted. This task is performed using a neural network, which allows a higher level of abstraction.

[2] It makes sense in scikit-learn's calculation.

This work does not aim to elaborate on explanations related to neural networks but on the algorithms that use them. The Doc2Vec algorithm and the subsequent techniques use neural networks, but their descriptions are self-sufficient.

The researchers proposed two Doc2Vec embedding algorithms: *Distributed Memory Model of Paragraph Vectors* (PV-DM) and *Distributed Bag of Words Model of Paragraph Vectors* (PV-DBOW). In this work, we consider the PV-DM, which we call Doc2Vec.

Figure 1 – Doc2Vec framework.



In this example, the objective is to predict the word "cat" from the context ("I love my") and the document ID. The document ID and each word are vectorized, concatenated, and passed through a shallow neural network, which predicts the word "cat" from the context and the document ID. Source: the authors.

Figure 1 represents the Doc2Vec framework. The context ("I love my") is used to predict the word "cat," but we also use the information of in which document these words are (document ID). More formally, consider a set of N documents $\{d_1, d_2, \cdots, d_N\}$, which are the indices of the respective documents in the corpus. Document i has words $\{w_{i1}, w_{i2}, \cdots, w_{iM_i}\}$, where M_i is the number of words of document i. w_{ij} is the index of the respective word in the vocabulary. Our objective is to predict w_{ij} using the context information,

$$w_{i,j-k}, w_{i,j-k+1}, \cdots, w_{i,j-1}, w_{i,j+1}, \cdots, w_{i,j+k-1}, w_{i,j+k},$$

(from now on we write $w_{i,j-k}, \cdots, w_{i,j+k}$ only), and the document, d_i.

Following the framework of Figure 1, consider matrix W, where each column is the vector of a specific word from the vocabulary. Similarly, each column of D is the vector representation of a specific document. h is a function that concatenates the vectors referent to the document (from D) and the words (from W) passed to it. The resulting vector $h(d_i, w_{i,j-k}, \cdots, w_{i,j+k})$ is passed through a shallow linear neural network with parameters b and U:

$$y = b + Uh(d_i, w_{i,j-k}, \cdots, w_{i,j+k}; W, D)$$

Vector y contains the information of the predicted words. To convert the coordinates of y to probabilities, we apply a softmax:

$$\mathbb{P}(w_{ij}|d_i, w_{i,j-k}, \cdots, w_{i,j+k}) := \frac{e^{y_{w_{ij}}}}{\sum_l e^{y_l}}.$$

Finally, our objective is to maximize the sum of the log-probability of the correctly predicted words, optimizing the choice of W, D, b, U:

$$\sum_{i=1}^{N} \sum_{j=k}^{M_i-k} \ln \mathbb{P}(w_{ij}|d_i, w_{i,j-k}, \cdots, w_{i,j+k}).$$

This task is performed using stochastic gradient descent, and the gradient is computed through backpropagation (RUMELHART; HINTON; WILLIAMS, 1986).

After the model's training, one can extract the document vectors (Doc2Vec embeddings) from the matrix D. If one wants to extract vectors of new documents, the process is similar, but the weights W, b, and U are frozen, only D is updated.

Because documents are optimized to consider their words and their words' contexts, Doc2Vec solves TF-IDF's word order problem. Besides, it also incorporates semantics because words with similar meanings share the same contexts more frequently.

2.3 Universal Sentence Encoder

Similar to Doc2Vec, researchers proposed ways of encoding text into a high-dimensional vector space, which they called *Universal Sentence Encoder* (USE) (CER et al., 2018). The authors proposed two models: the first is based on the *Transformer* (VASWANI et al., 2017); the second is based on *deep averaging networks* (DAN) (IYYER et al., 2015). In contrast to the original version, trained for the English language, there also exist multilingual versions (YANG et al., 2020) based on the Transformer and on *convolutional neural networks*[3] (CNN) (KIM, 2014).

The intuition of USE is similar to Doc2Vec's. Essentially, the model vectorizes the words and does some calculation with them, resulting in the document vector. The parameters of this model are then adjusted using neural network theory. For instance, the authors describe DAN USE's functioning more technically: for a given text, it works by tokenizing this text (splitting it into tokens, *i.e.*, words or part of words), taking the embeddings of tokens and bi-grams (two adjacent tokens), averaging them, and passing them through a neural network. It results in a vector with dimension 512, called document embedding. The model is trained in a multi-task learning approach, *i.e.*, it is trained in various *downstream* tasks, *e.g.*, a classification task. After training, the model is ready to be used: for each given text, it returns an embedding.

[3] USE's multilingual version based on CNN is available at https://tfhub.dev/google/universal-sentence-encoder-multilingual/3.

The USE's multilingual version based on CNN has a similar functioning, omitted in this work. Hereafter, we consider this model and call it USE.

2.4 Transformer

A recent breakthrough in natural language processing (NLP) research was the release of the Transformer (VASWANI et al., 2017), an attention-based neural network architecture initially specialized for texts. The NLP models based on this architecture were tested and achieved state-of-the-art results in this field (BELTAGY; PETERS; COHAN, 2020) for tasks like text classification, named entity recognition, and sentiment analysis. According to the authors, the Transformer is composed of two parts: an encoder and a decoder. Researchers discovered that using only the encoder is very effective for many NLP tasks. The consequence was the release of BERT (DEVLIN et al., 2019), a deep language model based on the Transformer encoder.

In this work, we consider the Transformer encoder when presenting the models BERT and Longformer.

2.4.1 BERT

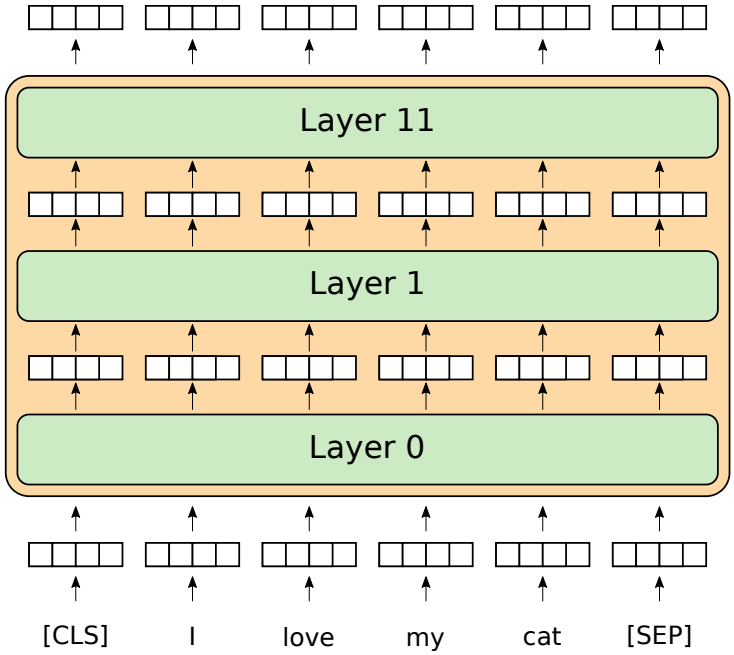
BERT stands for "Bidirectional Encoder Representations from Transformers," in reference to the fact that BERT creates representations for words conditioning on both left and right contexts of these words. The description of BERT, in particular the Transformer encoder, is very technical. Interested readers can refer to the original papers (VASWANI et al., 2017; DEVLIN et al., 2019) for the details. In this work, we plan to provide an overview of the model and an intuition of its functioning.

Figure 2 summarizes BERT's framework. Given a sentence, for instance, "I love my cat," this is pre-processed, tokenized (split into tokens, *i.e.*, words or parts of words), and special tokens are added at the beginning and end of the sentence. In this case, classifier ("[CLS]") and separation ("[SEP]") tokens. Each token is then converted into a text embedding, similarly to Doc2Vec's W matrix (section 2.2). All the vectors, together and simultaneously, pass through the BERT model and each of its layers, until the output at the top of the model. We can consider these output vectors as the text embedding of each token. All that remains to explain is which calculation is performed in each layer.

The functioning of a BERT's layer is pictured in Figure 3. Each layer is composed of two sublayers, a multi-head attention and a feed-forward sublayer. The feed-forward one is a simple neural network layer, similar to the one described in section 2.2, but, in this case, the function is different: if x is the input vector, the output is computed with

$$\mathrm{FFN}(x) = \max\left(0, xW_1 + b_1\right) W_2 + b_2,$$

Figure 2 – BERT's framework.



When a sentence is given, it is pre-processed, tokenized, and special tokens are included. The tokens are converted into vectors and passed through the model, generating their outputs at the top of BERT. Source: the authors.

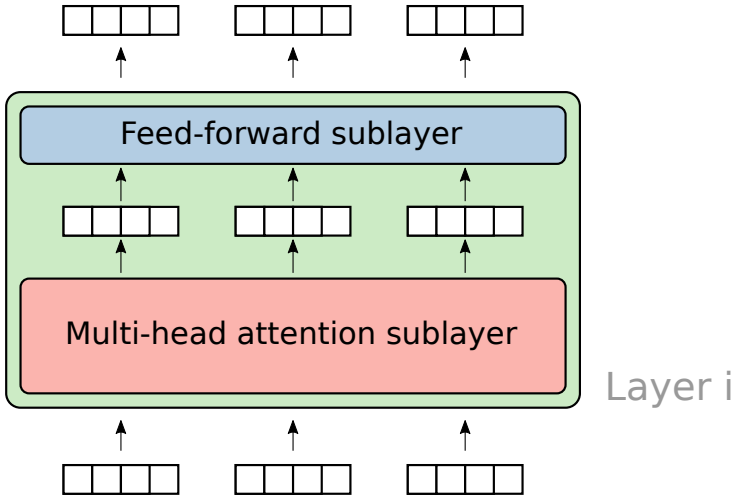
where $W., b.$ are parameters of the model (to be optimized). This sublayer is the same for all vectors.

The innovative part related to BERT (and Transformers, in general) belongs to the multi-head attention sublayer. Very didactically, the authors of the Transformer describe the meaning of an *attention* function: *"An attention function can be described as mapping a query and a set of key-value pairs to an output, where the query, keys, values, and output are all vectors. The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key."* (VASWANI et al., 2017) The authors consider the *scale dot-product attention* function in their work, where this "compatibility" is calculated through dot-products. As an example, consider the following vectors: a query q, keys $k_1, \cdots, k_n$, and values $v_1, \cdots, v_n$, where query and keys share the dimensionality. The compatibilities are calculated as $a_i = q \cdot k_i$ and these weights are used to compute a weighted sum of the values: $a_1 v_1 + \cdots + a_n v_n$.

Consider matrix Q, where each line is a query vector of dimension d_k; matrix K, where each line is a key of dimension d_k; and matrix V, where each line is a value. The scale dot-product attention can be computed through

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\intercal}{\sqrt{d_k}}\right)V.$$

Figure 3 – BERT's layer.



Each layer of BERT is composed of two sublayers: a multi-head attention and a feed-forward sublayer. Source: the authors.

The name "multi-head" comes from the fact that h attentions are concatenated and projected:

$$\begin{aligned} \text{multi\_head}(Q, K, V) &= \text{concat}(\text{head}_1, \cdots, \text{head}_h) W^O, \\ \text{head}_i &= \text{attention}\left(QW_i^Q, KW_i^K, VW_i^V\right). \end{aligned}$$

The projection matrices $W^O, W_i^{\cdot}$ are optimized during model training, as usual. In the case of BERT, the multi-head attention sublayer (Figure 3) is summarized by $\text{multi\_head}(Q, K, V)$, where $Q = K = V$ are matrices composed of the previous layer's vectors.

BERT's description is technical (and, in this work, superficial). Still, its intuition is that, inside the attention sublayer, the vectors of each token can attend to the vectors of all the other tokens, possible learning semantical and contextual information about natural language. It can be helpful in many different tasks, as this work aims to exemplify.

The training of BERT (usually referred to as *pre-training*) is performed using two procedures:

- Masked language modeling: the model is trained to predict randomly masked input tokens using the rest of the sentence. The tokens are masked by replacing them with the token "[MASK]." The predictions are made using the respective output vectors;

- Next sentence prediction: the model is trained to predict if two sentences are next to each other in the dataset. The predictions are done using the token "[CLS]."

The original BERT model (DEVLIN et al., 2019) was pre-trained using English textual data, which can be problematic if one wants to experiment with Portuguese text. This is our case (see chapter 5). Fortunately, BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) is a BERT model trained in Brazilian Portuguese textual data, which we consider in our work.

2.4.2 Longformer

As stated in subsection 2.4.1, the intuition of BERT is that the network can look into (it can "pay attention" to) the entire input sequence to extract contextual and semantical information. Although powerful, these models with self-attention suffer a $O(n^2)$ complexity of time and memory (BELTAGY; PETERS; COHAN, 2020), with n being the input sequence length (number of tokens). In practice, training these models considering a large input sequence length limit takes an enormous amount of time. Therefore, many pre-trained models have input sequence length limited to 512 tokens (DEVLIN et al., 2019; SOUZA; NOGUEIRA; LOTUFO, 2020; LIU et al., 2019). This is a difficulty when working with long documents, *e.g.*, those with thousands of tokens. Considering this problem, researchers proposed the Longformer model (BELTAGY; PETERS; COHAN, 2020), where the self-attention mechanism grows linearly (*i.e.*, it has complexity $O(n)$) with the input sequence length.

According to the authors, their self-attention mechanism "is a combination of a windowed local-context self-attention and an end task motivated global attention that encodes inductive bias about the task" (BELTAGY; PETERS; COHAN, 2020). In practice, the matrices used to calculate the self-attention are not "full" but consider only the local context, with the global context in only a few and specific tokens, *e.g.*, the "[CLS]" token, depending on the task. It grows linearly because the context length is fixed and the number of global inputs is small and independent of the input sequence length (BELTAGY; PETERS; COHAN, 2020). For this reason, the pre-trained and available Longformer model[4] has an input limit of 4096 tokens. Both model and tokenizer (*i.e.*, the tool that splits the text for the model) were trained with English data.

[4] The pre-trained Longformer model, in its simpler version, is available at `https://huggingface.co/allenai/longformer-base-4096`.

3 Classification

In the learning process of a machine learning algorithm, if the data has known labels, the process is called *supervised learning*, in contrast to the *unsupervised* one (JAIN; MURTY; FLYNN, 1999). For instance, the learning process could be identifying the weather from weather metrics, *e.g.*, air humidity. If the labels are classes (*i.e.*, they are discrete), we can call it a *classification* problem.

The goal of supervised learning is "to build a concise model of the distribution of class labels in terms of predictor features" (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2006). If we have new data where the predictor features are known but not the class labels, we can use the fitted classifier to predict, infer, or assign a class to this data (KOTSIANTIS; ZAHARAKIS; PINTELAS, 2006). In our study, we could fit a classifier to learn if a document cites binding precedent 37 using as predictor features the documents' raw texts. This way, if the model is suitable, we can use it to search for BP 37 citations in new documents.

We concentrate our attention on the classification of textual data, which is the goal of this work.

3.1 Support vector machines

Support vector machines (SVMs) are complex enough to be used for real-world classification problems and simple enough to be analyzed mathematically (HEARST et al., 1998). For this reason, they are suitable for text classification. A convenient way to build classifiers for texts with an SVM is using the document embeddings (chapter 2) as the predictor features to fit the algorithm, something we explore in our experiments (subsection 6.1.3).

The central intuition of SVMs is that the predictor features, in the form of vectors, are being separated using hyperplanes in a high-dimensional vector space, as in Figure 4. However, this separation happens in another *feature space* related to the original vector space in a non-linear form (HEARST et al., 1998).

Suppose we have data $(x_1, y_1), \cdots, (x_n, y_n) \in \mathbb{R}^m \times \{\pm 1\}$ and we want to separate the vectors x_i according to their classes y_i. We could search for a hyperplane $\{x \in \mathbb{R}^m : w \cdot x + b = 0,\ w \in \mathbb{R}^m,\ b \in \mathbb{R}\}$ such that $\text{sign}(w \cdot x_i + b) = y_i$ for all i (HEARST et al., 1998). It means that every vector of one class lies on one side of the hyperplane, while the vectors of the other class lie on the other side, as pictured in Figure 4. The

solution w, b can be computed through an optimization problem (JAMES et al., 2013):

$$\begin{aligned} \max_{b,w,\varepsilon_1,\cdots,\varepsilon_n,M} \quad & M \\ \text{s.t.} \quad & \|w\|^2 = 1, \\ & y_i(w \cdot x_i + b) \geq M(1 - \varepsilon_i), \quad i = 1, \cdots, n, \\ & \varepsilon_i \geq 0, \quad i = 1, \cdots, n, \\ & \sum_{i=1}^{n} \varepsilon_i \leq C, \end{aligned}$$

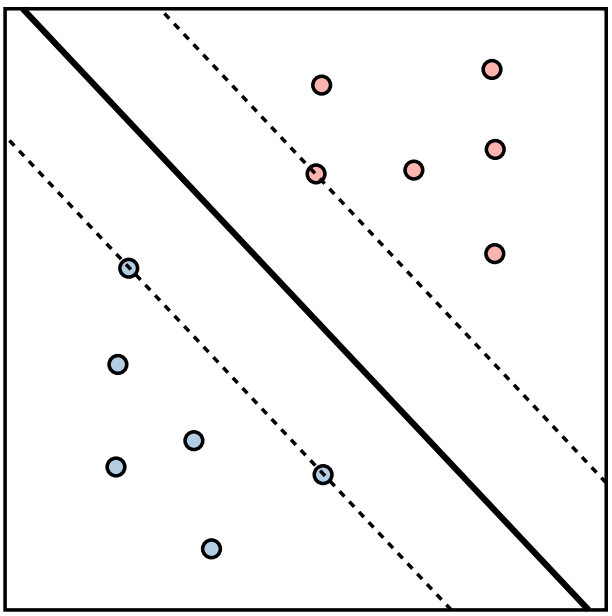
where C is a regularization parameter which allows misclassifications.

An optimal hyperplane, with the distance of all points to the hyperplane maximized, will have w as a linear combination of a subset of the closest vectors x_i, called the *support vectors*. They lie on the margin (HEARST et al., 1998), as pictured in Figure 4. Consequently, the decision function becomes

$$f(x) = \text{sign}(w \cdot x + b) = \text{sign}\left(\sum_{j=1}^{k} \alpha_j x_j \cdot x + b\right),$$

which depends only on dot products (HEARST et al., 1998).

Figure 4 – Optimal hyperplane example.



The blue dots are separated from the red dots using a hyperplane, represented by a line. The line is chosen to maximize the margin, *i.e.*, the distance between the line and each point. Source: the authors.

Things become more interesting when this hyperplane is asked to be computed in another *feature space* $\mathcal{F}$, related to the original space via a nonlinear function $\Phi : \mathbb{R}^m \to \mathcal{F}$.

In this case, the decision function becomes

$$f(x) = \text{sign}\left(\sum_{j=1}^{k} \alpha_j \Phi(x_j) \cdot \Phi(x) + b\right).$$

The explicit mapping $\Phi(x)$ can be expensive to compute. However, in a variety of special cases, we can substitute the dot product in the feature space by a kernel function $K(x, y) = \Phi(x) \cdot \Phi(y)$, chosen so that K is easy to calculate. The decision function becomes

$$f(x) = \text{sign}\left(\sum_{j=1}^{k} \alpha_j K(x_j, x) + b\right),$$

much less expensive to compute (HEARST et al., 1998). Examples of kernel functions, which we explore in this work, are:

- Linear: $k(x, y) = x \cdot y$;
- Radial basis function (RBF): $k(x, y) = \exp\left(-\frac{\|x - y\|^2}{2\sigma^2}\right)$ (HEARST et al., 1998).

3.2 BERT and Longformer

Given a text, the usual behavior of BERT-like models is to output text embeddings, one for each token (refer to section 2.4 and Figure 2). However, it is also possible to use these models for other tasks, *e.g.*, classification. Concerning the output vector of token "[CLS]," BERT authors say: *"The final hidden state corresponding to this token is used as the aggregate sequence representation for classification tasks."* (DEVLIN et al., 2019) So, in practice, we plug into BERT's "[CLS]" output a classification layer, similar to Doc2Vec's (section 2.2), but, instead of predicting words, the model predicts classes. Using less data than the pre-training (subsection 2.4.1), the classification layer, so as the BERT model, is optimized (*fine-tuned*) to learn to classify the data. The process is similar for Longformer.

3.3 Performance metrics

Suppose a model is trained for a classification task. If we continue training the model in the same dataset, we may reach a point in which the model knows the training data so well that it starts to repeat it. This is known as *overfitting*. Because of this, one may want to evaluate the performance of the classifier "in the wild," *i.e.*, in new data. Usually, after fitting the models on a training dataset, they are evaluated on a different one, called test dataset, that must follow the same distribution of the training one. By "follows the same distribution," we assume there exists a probability distribution that generated the training dataset, *i.e.*, each training example is a sample from this distribution, and that all test examples were also drawn from it.

There exist several performance metrics for binary classification tasks (like the one described in section 3.1). Suppose a model was trained to predict positive or negative. When it predicts a positive correctly, it is called a true positive (TP); when it incorrectly predicts a negative, it is a false negative (FN); and so on. The following summarizes some popular metrics (OLSON; DELEN, 2008):

- Accuracy: measures the hit rate of the model, *i.e.*, how many of the model predictions are correct. It is defined as
$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}.$$
- Precision: it measures how many of the predicted positives are correct. It is defined as
$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$
- Recall: how many of the true positives were predicted as a positive.
$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$
- F_1-score: the harmonic mean between precision and recall.
$$\text{F}_1 = \frac{2}{\frac{1}{\text{precision}} + \frac{1}{\text{recall}}}.$$

Another common approach for measuring the performance of a classifier is using a confusion matrix, like the one in Table 2. Each cell (i, j) displays the percentage of examples of the actual class i that were assigned to the predicted class j by the model. Cell (P, P) is the recall.

Table 2 – Confusion matrix.

		Predicted	
		N	**P**
True	**N**	$\frac{\text{TN}}{\text{TN} + \text{FP}}$	$\frac{\text{FP}}{\text{TN} + \text{FP}}$
	P	$\frac{\text{FN}}{\text{TP} + \text{FN}}$	$\frac{\text{TP}}{\text{TP} + \text{FN}}$

Source: the authors.

3.4 Cross-validation

In section 3.3, we detailed it is usual to divide the data between train and test datasets: the former for the fitting of the models, and the latter for the evaluation in new

data. However, it is common that the desired model must be chosen from a list of possible models, for instance, different hyperparameters for the same algorithm (as in section 3.1 for hyperparameter C). In this case, one could fit each model in training data and choose the best model in test. When doing this, the model is chosen based on information present in both train and test datasets, so the test data is no longer new for the model.

A solution to this problem is to use a third data split, usually called the validation dataset, where the models will be tweaked. In this case, test data remains unknown until evaluation. Another solution, called *k-fold cross-validation*, splits the training data into k folds and, for each fold, it is used to evaluate the model trained in all other folds. In the end, evaluation results are aggregated, the model is chosen, and, finally, it is evaluated in an independent test data.

There exist variations of cross-validation algorithms. In this work, we consider a version in which the fold split happens after a data shuffle, and the folds (approximately) preserve class proportionalities, *i.e.*, the splits are stratified.

4 Machine learning explainability

A well-fitted machine learning model is not enough: when the model is used for decision-making, explaining its decisions is also necessary (RIBEIRO; SINGH; GUESTRIN, 2016), especially when dealing with sensitive domains, *e.g.*, legal documents. Ideally, a model would be interpretable by design. For instance, it is trivial to interpret the weights of linear regression as the importance of the features; another example is a decision tree, which has its nodes as if-else clauses. However, this is not so trivial for neural networks, for example, something that motivates the development of machine learning explainability techniques. This work covers the standard, popular LIME technique (RIBEIRO; SINGH; GUESTRIN, 2016) for explaining classifiers.

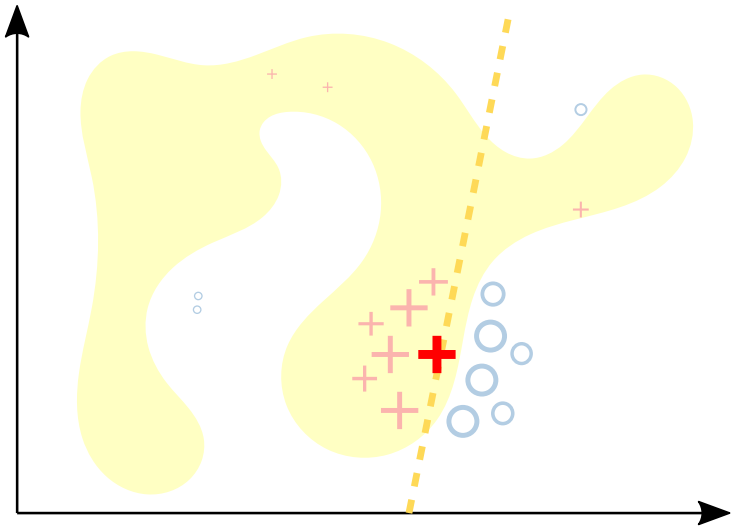
4.1 Local Interpretable Model-agnostic Explanations

Consider a classifier f which returns the probability $f(x)$ of a document embedding x of belonging to class C. One could require understanding why the model took this decision, *e.g.*, in the context of a text classifier, they could want to visualize the most important words or sentences for assigning the text to class C. A simple, possible solution would be to remove a few words from the text, extract the document embeddings, and verify how much the model changes its probability inference. If the prediction decreases too much, this indicates that these words are relevant for the model; if the probability increases, they have negative importance. This is an approach simple to understand but prone to instability in practice.

A next step in the above idea could be to repeat the process several times and train a simpler, interpretable model to predict the probabilities $f(x)$ using only the information of the presence or the absence of words in the text. If this surrogate model is well fitted, it knows to predict the original model's probabilities from the presence or absence of the words. Because it is interpretable, we can extract these explanations easily. This task can be understood as approximating the model f in the locality of the sample x using a simpler, interpretable model g, as represented in Figure 5. This is the foundation of the Local Interpretable Model-agnostic Explanations (LIME) technique (RIBEIRO; SINGH; GUESTRIN, 2016).

More formally, consider $x \in \mathbb{R}^d$ as our original representation, *e.g.*, USE document embedding, and $x' \in \mathbb{R}^{d'}$ as the *interpretable representation*, *e.g.*, a vector indicating the presence or absence of a word in the text. We want to explain the model $f\colon \mathbb{R}^d \to [0, 1]$ for the sample x, *i.e.*, we want to answer the question *"what are the most important words for saying the text (respective to the vector) x has probability $f(x)$?"* To solve this, we search

Figure 5 – Local Interpretable Model-agnostic Explanations representation.



Suppose $\mathbb{R}^2$ is the original vector space, *e.g.*, Doc2Vec with 2 dimensions. In this example, the classification boundary (yellow stain) is too complex to be explained. Given an example (red cross), LIME approximates the boundary perturbing this example and fitting a linear, interpretable model (dashed line). Note how the line approximates the classification boundary in the vicinity of the cross. Source: the authors.

for an interpretable model $g \in G$, with $G = \left\{g \colon \{0,1\}^{d'} \to \mathbb{R}\right\}$ being a pre-defined class of interpretable models, for instance, linear models. We also want a not too complex model, so g must have low complexity $\Omega(g)$. Our main goal is to find g *locally faithful* to f, in the sense of g being a good approximation of f (calculated by $\mathcal{L}$) in the neighborhood of x given by a function π_x. The explanation model g can be found solving the following optimization problem:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g).$$

The authors solve this optimization by sampling z' from x' randomly changing some coordinates of x' to zero, *i.e.*, removing some words of the original text. These new interpretable vectors $z' \in \mathcal{Z}$ are used to generate new document embeddings $z \in \mathbb{R}^d$ with new probabilities $f(z) \in [0, 1]$. The problem can be simplified choosing G to be the set of linear models $G = \{g(z') = w_g \cdot z'\}$ and $\mathcal{L}$ to be the weighted square loss

$$\mathcal{L}(f, g, \pi_x) = \sum_{z,z'} \pi_x(z)(f(z) - g(z'))^2, \tag{4.1}$$

where the locality π_x is given by

$$\pi_x(z) = \exp\left(-D(x', z')^2/\sigma^2\right), \quad D(x', z') = \frac{x' \cdot z'}{\|x'\|\|z'\|}.$$

The complexity is a limit K on the number of chosen words, which is represented by $\Omega(g) = \infty \cdot \mathbb{1}\left(\|w_g\|_0 > K\right)$. Finally, the task of explaining f for example x reduces to solving the weighted linear regression of Equation 4.1 with regularization.

To fix ideas, suppose we have a high-performance text classifier to predict if a text is talking about cats. We want to explain, for instance, why the following text is classified as "talking about cats": *"Cute little furballs that sleep all day long."* LIME randomly removes some words from this text and extracts the classifier probabilities. After, it fits a linear model to predict these probabilities from the presence or absence of the words. In the end, the weights of the linear regression can be mapped to a color scale for visualization purposes, where green means positive importance and red, negative importance:

Cute little furballs that sleep all day long .

Part II

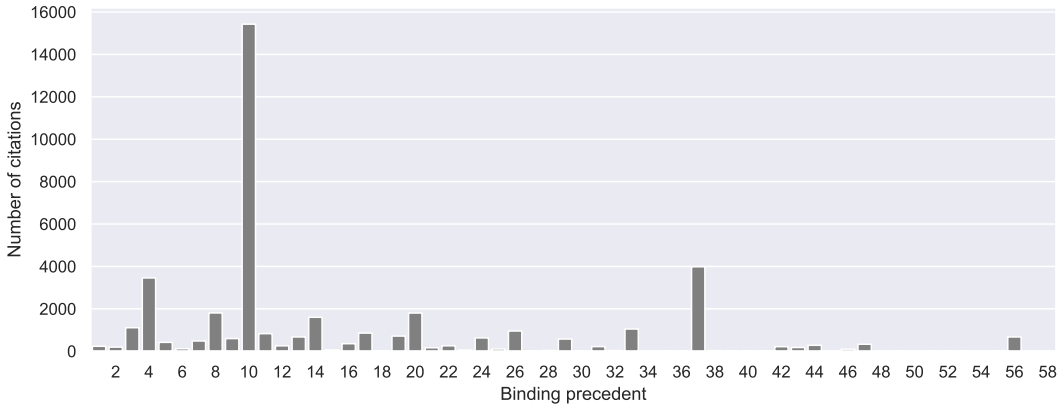
Methodology

5 Dataset

The Brazilian Supreme Court (STF) has to decide among many cases and processes during its functioning. The Court consolidates these decisions in documents that contain the decision text in Brazilian Portuguese, the publication date, the Justice Rapporteur (*i.e.*, the minister who took the decision), and the document type (*e.g.*, a Habeas Corpus or a Complaint). Frequently, STF documents cite one or more of the 58 existing binding precedents (BPs). For example, BP 10 is cited in 2018 in a decision of the process RCL 30135: *"(...) In the initial petition, the plaintiff alleges that the contested decision offended the authority of this Court, embodied in the binding precedent 10 (...)"* (free translation). In this work, we use the expressions "decision" and "document" interchangeably.

To be able to conduct our analyses and experiments, we collected data related to STF decisions in a partnership with the *Supremo em Números* project ("STF in Numbers," free translation) (FALCÃO et al., 2017). In this dataset, there are more than 2,500,000 documents since 1988, and they contain the decision text and metadata, *e.g.*, the document type and the BP being cited, if it is the case. Regarding binding precedents, there are 41,031 citations in a non-uniform distribution, as presented in Figure 6. Because one document can cite several BPs, the number of individual documents is lower: 38,364.

Figure 6 – Number of citations per binding precedent.

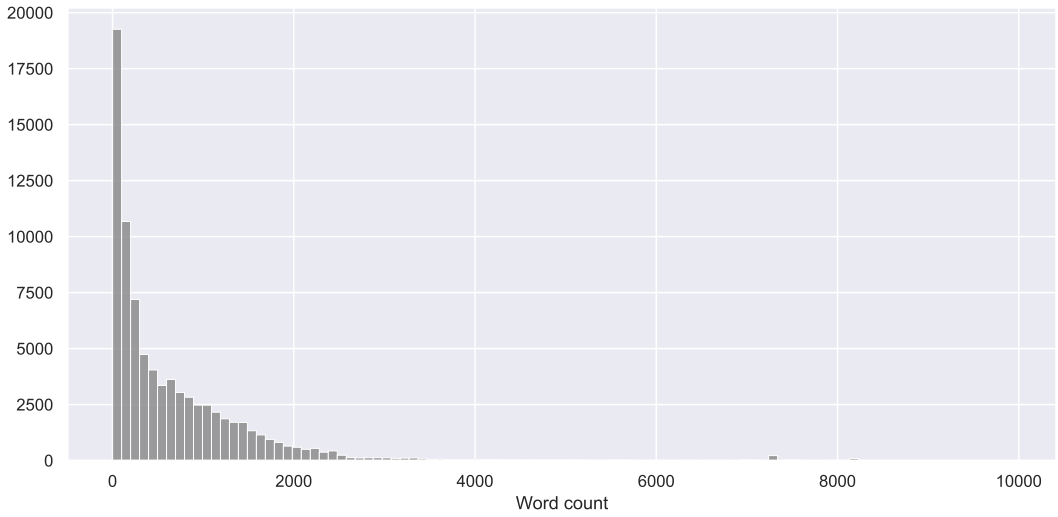


Source: the authors.

As it is clear from Figure 6, many BPs do not have enough citations for our experiments that involve machine learning models, so, in this work, we consider only the ten most-cited BPs. Most-cited BPs are calculated after filtering duplicated documents, *i.e.*, documents with different IDs in our dataset but that share the same raw texts. Following this criteria, the ten most cited BPs are $\{3, 4, 10, 11, 14, 17, 20, 26, 33, 37\}$, resulting in 31,070 BP citations (76% of the original value) and 29,743 documents (78%).

It is also essential to search for potential citations in documents that do not explicitly cite any BP, so we also gathered a set of 50,000 documents of this kind. Considering the entire dataset, *i.e.*, documents with citations to the ten most-cited BPs and documents without any BP citation, Figure 7 presents a word count histogram, showing that most documents have less than 4,000 words[1].

Figure 7 – Histogram of word count of the dataset.



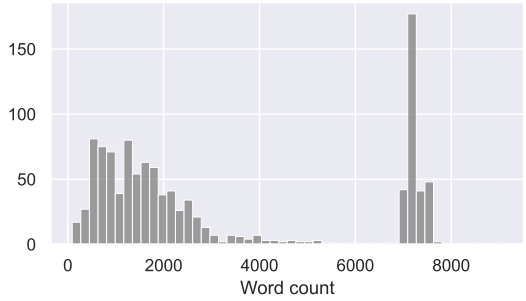
Source: the authors.

It is relevant to point that the word count distribution varies among documents that cite different BPs (or do not cite any of them). As an example, Figure 8 presents a word count histogram for different subsets of the dataset. The differences in word count distribution among documents that cite different BPs should not be considered a concern because it is natural to the data. We address this subject in chapter 9.

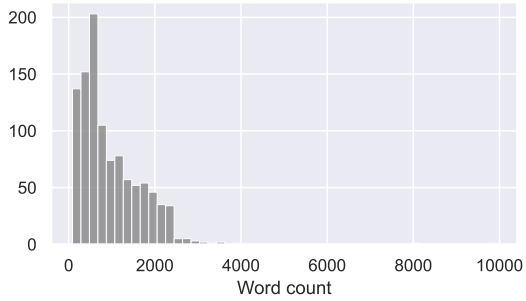
[1] The division of the texts into words was done using consecutive spaces as separators.

Figure 8 – Histogram of word count for subsets of the dataset.

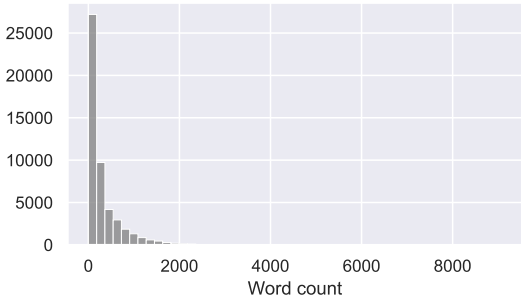
(a) Documents that cite BP 3.



(b) Documents that cite BP 33.



(c) Documents that do not cite any BP.



Source: the authors.

6 Identification of potential citations

Suppose a document does not cite a binding precedent explicitly, but the context is similar enough for a specialist to make an association with the BP or for a well-trained machine learning model to classify it as with an explicit citation. In this case, we define it as a *potential citation*. Finding potential citations has several applications:

- it allows searching for related cases that are not similar enough to cite a BP but that are similar enough to be used by a lawyer during their argument construction;
- it allows detecting lower courts' processes that should have cited the BP but they did not;
- it is possible to find decisions related to a BP before its creation, even decisions that motivated the creation of the statement;
- it allows finding documents with explicit citations, but that were not corrected labeled;
- it enables finding documents with explicit citations but with typing errors that made it impossible to have them correctly labeled.

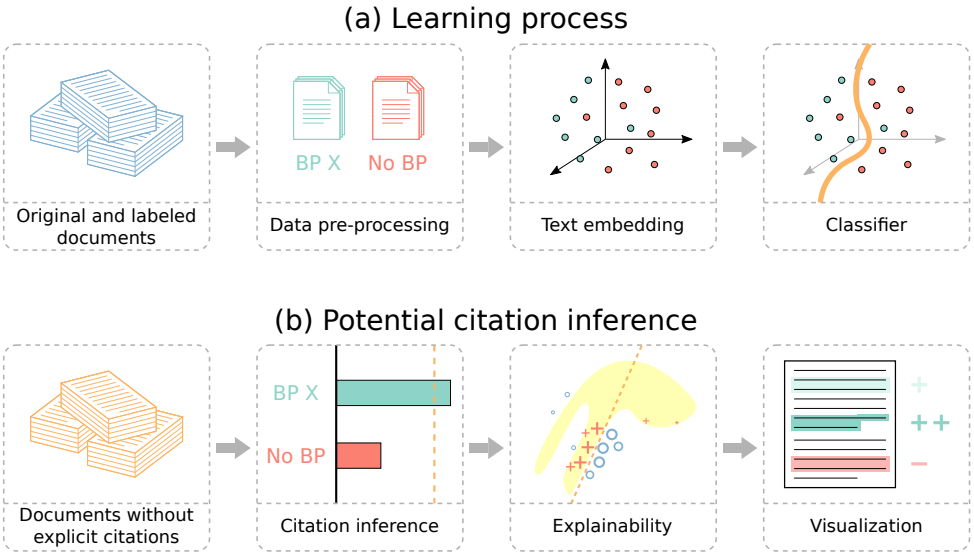
Also, the methodology of searching for potential citations to BPs can be extended to other types of relation among documents. For instance, one may desire to find documents in which an ordinary *súmula* could have been applied, but it was not because its application is not mandatory (chapter 1). Another interesting possibility is finding lower courts' decisions that should not have been taken because STF decided that all processes related to the subject should be "on hold" (free translation of "sobrestados") until an STF decision. Today, this process is manual and can be improved with our methodology.

Figure 9 presents our pipeline of identification of potential citations. This pipeline consists of a learning process, with models learning what characterizes a citation, and an inference process, when potential citations are inferred and explained. In the following, we present this pipeline in detail.

6.1 Learning process

To search for potential citations, we need to discover a function f that receives a document and returns the answer if there is a potential citation to a specific BP. A possible approach to reach this goal is applying an unsupervised algorithm that receives

Figure 9 – Pipeline of identification of potential citations.



(a) First, original, labeled documents are pre-processed and converted into high-dimensional vectors before being fed into classification machine learning models. (b) These models then receive documents without explicit citations and infer potential citations. An explainability technique acts to explain this inference and visually presents the most important sentences. Source: the authors, based on the original work (chapter 1).

a set of documents and returns them grouped in clusters of similar documents. In this context, non-labeled documents that are similar to several labeled documents should also be labeled, in this case, with a potential citation.

Another way to solve this problem is using a supervised technique: we can fit a classifier to labeled documents, teaching the model to predict an explicit citation from the documents' raw texts. When the model predicts a citation, but there are not any of them, we assign a potential citation. This method has some advantages when compared to the unsupervised. Because classifiers return probabilities (or can be adjusted to return them, subsection 6.1.3), it is easy, intuitive, and straightforward to define a probability threshold for the models' confidence in the potential citation assignment. Also, although classifiers can act like black-boxes, we can benefit from machine learning explainability literature (chapter 4).

Instead of applying a single classifier that predicts a citation to every of the ten most-cited binding precedents, *i.e.*, a multiclass supervised model with ten classes — as in the original work —, in this work, we fit a classifier for *each* BP, *i.e.*, ten binary classifiers. Although similar in practice, this methodology allows the probabilities not to add up to one, unlike the case for general multiclass classifiers. For instance, all BPs could be potentially cited with probability 95% (or 5%) simultaneously, impossible for the general multiclass approach. What we are proposing is a multilabel approach, although it is not

efficient in practice. This efficiency limitation is discussed in chapter 9.

6.1.1 Data pre-processing

We must pre-process our data (chapter 5) before feeding the embeddings and classifiers. For each of the ten most-cited binding precedents, we create a sample dataset to be used during training and evaluation. For a BP, this sample consists of all documents with citations to it[1] and several documents without any BP citation, ensuring balancing. The balancing guarantee is a simple, random undersampling approach to deal with our high imbalanced binary classification problem (KRAWCZYK, 2016) — this issue is discussed in chapter 9. For each sample dataset, we remove duplicated documents (decisions with different IDs but with the same raw text) to avoid overfitting the data during training. For instance, BP 37 has 3,984 citations (without duplicates), so its sample dataset contains these 3,984 documents plus 3,984 documents without any BP citation. Each resulting sample dataset is split into a train (64%), validation (16%), and test (20%) sets. The validation sets are used to compare the different models' results, while the test sets are used to evaluate the chosen model.

The texts that compose the sample datasets have explicit citations to BPs removed to avoid offering hints during the models' training. For TF-IDF and Doc2Vec document embeddings, the texts are pre-processed through a standard NLP data processing pipeline, including lowercase conversion, tokenization (text splitting into words), punctuation and stop words removal, and lemmatization (conversion of words into their lemmas, *e.g.*, "walking" is converted into "walk").

The sample dataset sizes are summarized in Table 3. It is relevant pointing that, although two BPs can have sample datasets with a high difference in the number of documents, the classifiers are fitted and evaluated for each BP individually. No comparisons are made among BPs.

Table 3 – Number of documents in the sample dataset of each BP.

3	**4**	**10**	**11**	**14**	**17**	**20**	**26**	**33**	**37**
2,002	5,404	8,000	1,446	2,950	1,652	3,484	1,880	2,094	7,968

Source: the authors.

[1] The exception is for BP 10, which has 14,269 citations after duplicate filtering, but only 4,000 of them are used. We do this to reduce the discrepancy of sample dataset sizes when compared to the other BPs (see Table 3).

6.1.2 Document embedding

To convert texts into vectors, we experiment with the following document embedding techniques: TF-IDF, Doc2Vec, USE, BERT, and Longformer (chapter 2). For TF-IDF, because of its high dimensionality (usually higher than tens of thousands), we apply a truncated singular value decomposition (SVD) dimensionality reduction (MANNING; RAGHAVAN; SCHÜTZE, 2008) to 50[2]. TF-IDF vectorizers and dimensionality reducers are fitted on training sets. For Doc2Vec, it is trained for 30 epochs on training sets individually. For USE, we consider its multi-lingual version based on CNN `universal-sentence-encoder-multilingual`[3]. For BERT, we consider BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), a BERT model pre-trained in Brazilian Portuguese data, specifically the `neuralmind/bert-base-portuguese-cased` version[4]. For Longformer, we consider the version `allenai/longformer-base-4096`[5], which is trained for the English language. The resulting embedding dimensionalities are summarized in Table 4.

Table 4 – Dimensionality of each document embedding.

TF-IDF	**Doc2Vec**	**USE**	**BERT**	**Longformer**
50	50	512	768	768

Source: the authors.

6.1.3 Classification

For TF-IDF, Doc2Vec, and USE, we experiment with a support vector machine (SVM) with kernels linear and RBF (section 3.1). In the case of kernel linear, we test the hyperparameters $C \in \{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$; in the case of kernel RBF, $C, \gamma \in \{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$[6]. Hyperparameters are chosen by 5-fold stratified and shuffled cross-validation on training data (section 3.4).

Before SVM, all vectors suffer a standardization, *i.e.*, each coordinate has its mean subtracted and is divided by its variance, resulting in vectors with mean zero and variance one in each dimension. This may improve SVM classification performance (ALI; SMITH-MILES, 2006). The standardization scalers are fitted on the training sets individually. SVM does not return probabilities by default. To solve this, we apply a classifier *calibration*, *i.e.*, the creation of a map from the classifier's output to a probability estimate. In practice,

[2] Details available at https://scikit-learn.org/stable/modules/decomposition.html#truncated-singular-value-decomposition-and-latent-semantic-analysis.
[3] Available at https://tfhub.dev/google/universal-sentence-encoder-multilingual/3.
[4] Available at https://huggingface.co/neuralmind/bert-base-portuguese-cased.
[5] Available at https://huggingface.co/allenai/longformer-base-4096.
[6] $\gamma = 1/(2\sigma^2)$, section 3.1.

this is done by fitting a logistic regression over the SVM scores (PLATT, 1999), using 5-fold cross-validation to split data for training and calibration.

For BERT and Longformer, we plug a classification layer into the "[CLS]" output vector (section 3.2). Because BERT has an input length limit of 512 tokens, the texts are split into 512-length texts (with an overlap of 64 tokens), trained individually. Both models are fine-tuned on training sets with batch sizes 128 (BERT) and 12 (Longformer), for 5 epochs[7], with AdamW optimizer (LOSHCHILOV; HUTTER, 2018), and a linear decrease of the learning rate from $2 \cdot 10^{-5}$ to zero.

All models are evaluated on validation sets (section 3.3). In the case of BERT, predictions are extracted for each text split and aggregated taking the maximum of the class logits (model outputs before softmax, as in section 2.2) over all the splits of the same text. This decision was taken based on preliminary tests. We reference it as *maximum logit* evaluation.

6.2 Potential citation inference

The association of a document embedding with a classifier (and potentially dimensionality reducers, scalers, and probability calibrators) is what we call a *model*. So a model receives a raw or processed text and returns the probability of a specific BP citation, *e.g.*, BP 37. With all models adjusted in our training sets and with proven performance, it is possible to use them to search for potential citations.

6.2.1 Citation inference

When a trained model is very efficient in pointing out an explicit citation, even when the explicit citation was removed (subsection 6.1.1), we can extrapolate its original role and start searching for citations where there are not any of them, *i.e.*, *potential citations*. Because a model gives us a probability of citation, we can interpret it as the model's confidence in assigning a potential citation. For instance, if BERT trained with BP 37 sample dataset points that a specific document has a BP 37 citation with a probability of 95%, but it has not, we say the model is 95% sure of a potential citation to this statement. We can also set a threshold, for instance, 95%, and say that every document without explicit citations but with a probability greater than this threshold has a potential citation.

[7] Models are saved each epoch and the one with best training accuracy (in an independent subset) is considered for evaluation.

6.2.2 Explainability

Understanding why a potential citation was assigned is not straightforward, especially for laypeople in the field of law. We rely on the Local Interpretable Model-agnostic Explanations (LIME) (RIBEIRO; SINGH; GUESTRIN, 2016) (section 4.1) explainability method to visually present the essential parts of the text for the potential citation assignment. By default, LIME computes the most important words. However, because the texts are long, as in Figure 7, we ask the method to compute the importance of entire sentences. This methodology is analogous to that explained in section 4.1.

6.3 Computational resources

We did all the computational work in Python. To facilitate reproducibility, Table 5 lists the main Python libraries and versions used in this work. In the near future, we plan to release data and code, which will be available at `https://github.com/lucasresck/potential-citations`. It is not possible to release them with the publication of this work because the original work is under peer review (chapter 1).

Table 5 – Main Python libraries used in this work.

Library	**Version**	**Function**
Datasets	1.15.1	Load datasets and metrics for BERT and Longformer
Gensim	4.0.1	Doc2Vec training
Lime	0.2.0.1	LIME calculations
NLTK	3.6.5	Split of texts into sentences
NumPy	1.21.2	Data manipulation
pandas	1.3.3	Data manipulation
PyTorch	1.10.0	BERT and Longformer training
scikit-learn	1.0.1	SVM, calibration, dimensionality reduction, TF-IDF, cross-validation, standardization, evaluation
spaCy	2.3.5	Text pre-processing
TensorFlow	2.2.0	USE extraction
Transformers	4.6.1	BERT and Longformer training

Source: the authors.

Part III

Results

7 Modeling results

We now present the results of the experiments described in chapter 6.

7.1 Quantitative analyses

Embeddings, dimensionality reducers, scalers, classifiers, and probability calibrators (subsection 6.1.3) were fitted on the training portion of each BP sample dataset individually. Hyperparameters were optimized using cross-validation. Table 6 presents the F_1-score evaluation of the models on validation data. We provide similar tables for accuracy, precision, and recall in Appendix A.

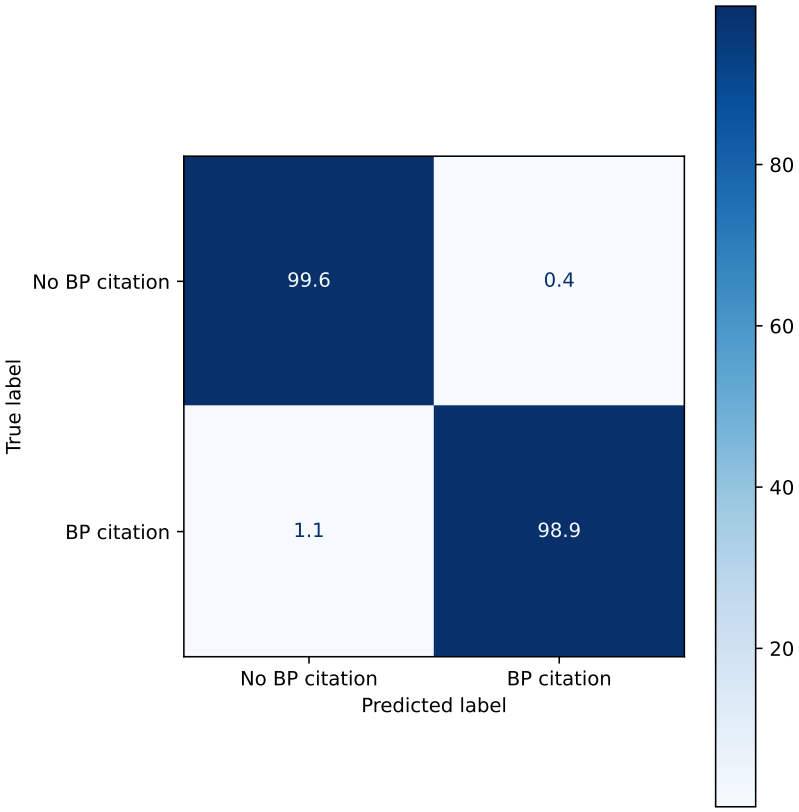
Table 6 – F_1-score (in percentage) of each model on validation data.

Model	Binding precedent									
	3	**4**	**10**	**11**	**14**	**17**	**20**	**26**	**33**	**37**
TF-IDF RBF	**98.5**	**98.7**	96.4	**97.0**	96.9	97.0	**99.3**	**99.0**	97.9	**99.1**
BERT	96.7	97.8	97.4	96.2	96.9	95.7	99.1	98.3	**98.5**	98.9
TF-IDF linear	98.1	97.3	95.2	95.7	96.9	**98.5**	99.1	98.7	97.0	98.9
Longformer	87.9	**98.7**	**98.8**	95.2	**98.7**	88.1	98.4	93.6	97.1	99.0
USE RBF	97.5	96.9	96.2	95.7	97.7	95.9	98.2	98.0	97.6	98.9
USE linear	97.5	96.2	94.6	94.9	97.3	95.9	97.8	98.0	97.3	98.6
Doc2Vec RBF	95.0	93.9	89.2	93.2	93.6	92.1	95.8	95.2	94.2	96.5
Doc2Vec linear	94.3	91.8	84.6	87.4	91.5	88.7	93.6	92.3	92.0	92.8

For each embedding, classifier, and BP sample dataset, the table presents the F_1-score (in percentage) on validation data. BERT is evaluated using the maximum logit evaluation (subsection 6.1.3). Bold values are the best values of the column. Source: the authors.

From Table 6, we notice that all models have performed well — under F_1-score on validation data — in the task of predicting the presence of a BP citation, especially those based on TF-IDF and Transformers. The model with the best performance is the SVM over TF-IDF embeddings with RBF kernel. It achieved the best F_1-score in the majority of the BPs; in those the model did not achieve, the performance was also excellent. In the other tables of Appendix A, this model's high performance is also evident. As an example, Figure 10 presents the confusion matrix of this model for the validation portion of the BP 20 sample dataset. The confusion matrices for the other BPs are available in Appendix B. We see this model really performs well in distinguishing the presence and absence of a BP citation because of the diagonal's dominance in the matrix.

Figure 10 – Validation confusion matrix (in percentage) of TF-IDF RBF for BP 20.



Source: the authors.

7.2 Qualitative analyses

The presented quantitative results support the idea of the feasibility of the potential citations methodology described in chapter 6. We saw the models can extrapolate training data in the task of identifying explicit citations.

Although we have little data, which can affect deep learning models' performance, and Longformer is specialized in English texts, it is not a surprise that the language models would end in the performance top 4. These models have been showing state-of-the-art results in NLP (BELTAGY; PETERS; COHAN, 2020). However, it is not straightforward to understand the results of TF-IDF because this technique is basically a count of words. We point again that the explicit citations were removed from the texts. These results suggest that the citation of a BP is predictable by the presence and absence of words, something that is reasonable after all, because it is common for Justices to include the entire BP statement when they cite it. Also, even when the statement is not fully present, it is expected the same words will be because of the related subject.

To continue our analyses, we chose to work with the TF-IDF RBF model. In our context, this model has some advantages when compared to the language models: (a) the

results — except for recall — were better in general; (b) it is a simple technique, and the simpler, the better; (c) it is an easier and lighter technique to implement, run, save, load, and test. Point (c) is particular: being faster is a desired characteristic when running an explainability technique like LIME because it relies on running the model thousands of times to generate the explanations.

7.3 Chosen model on test data

Table 7 presents the evaluation results in the test portions of the sample datasets for TF-IDF RBF. It is evident the model continued to generalize after being chosen using the validation portions. It supports the idea of the model being used to search for potential citations.

Table 7 – Evaluation (in percentage) of TF-IDF RBF in the test portions for each BP.

Metric	Binding precedent									
	3	**4**	**10**	**11**	**14**	**17**	**20**	**26**	**33**	**37**
Accuracy	97.3	97.9	96.9	96.6	97.8	98.8	98.9	98.1	99.0	99.1
F1-score	97.2	97.9	96.9	96.5	97.8	98.8	98.8	98.2	99.0	99.1
Precision	98.5	98.7	96.1	97.9	99.0	99.4	99.4	96.4	99.0	98.8
Recall	96.0	97.0	97.8	95.2	96.6	98.2	98.3	100.0	99.0	99.4

Source: the authors.

7.4 Identification analyses

We now use the chosen model, TF-IDF RBF, to search for potential citations in unlabeled, unknown data. For each binding precedent, we search for potential citations using the model trained on its sample dataset. The models are exposed to the unlabeled dataset, *i.e.*, without explicit citations, of 50 thousand documents (chapter 5) after removing the part of this data present in the corresponding sample dataset. We assign a potential citation to a document when the model's citation probability is higher than a threshold t_c.

Table 8 presents the number of potential citations found for each BP and probability threshold t_c. It is evident the lower the threshold, the higher the number of potential citations. The number of assigned documents is relevant, approximately 5.5% for $t_c = 0.9$. It is not possible in this work to validate all these potential citation assignments. However, in chapter 8, we present a few case studies that we believe help us demonstrate the methodology usefulness.

Table 8 – Number of potential citations assigned by TF-IDF RBF by BP and t_c.

t_c	**Binding precedent**									
	3	**4**	**10**	**11**	**14**	**17**	**20**	**26**	**33**	**37**
0.99	124	56	88	103	88	104	41	60	90	42
0.95	223	124	222	284	241	276	80	182	164	101
0.90	312	203	328	447	354	403	103	256	214	149

t_c is the probability threshold for the potential citation assignment. Source: the authors.

8 Case studies

In chapter 7, we presented quantitative and qualitative results we argued they demonstrate the feasibility of the methodology of potential citations search. However, validating all assigned potential citations is not trivial, so we restrict ourselves to presenting relevant examples, which we call case studies.

8.1 Binding precedent 37 and *súmula* 339

Binding precedent 37 was published on 10/24/2014 and has the following statement in Portuguese: *"Não cabe ao Poder Judiciário, que não tem função legislativa, aumentar vencimentos de servidores públicos sob o fundamento de isonomia."* An English free translation is present in chapter 1. Considering TF-IDF RBF trained to identify BP 37 citations, the model identifies a citation in a Justice Luiz Fux's decision from 06/29/2012 in the process ARE 694870 in our dataset with 50 thousand documents without BP citations[1]. We present part of this decision as it is, in Portuguese, in our dataset:

"RECURSO EXTRAORDINÁRIO COM AGRAVO 694.870 RIO GRANDE DO NORTE

RELATOR : MIN. LUIZ FUX RECTE.(S) :ANDRÉIA APARECIDA DOS SANTOS SALES E

OUTRO(A/S)

ADV.(A/S) :JOÃO PAULO DOS SANTOS MELO E OUTRO(A/S) RECDO.(A/S) :ESTADO DO RIO GRANDE DO NORTE PROC.(A/S)(ES) :PROCURADOR-GERAL DO ESTADO DO RIO

GRANDE DO NORTE

(...)

Ex positis, NEGO SEGUIMENTO ao agravo de instrumento, com

fundamento no artigo 21, § 1º, do RISTF.

Publique-se. Int.. Brasília, 29 de junho de 2012.

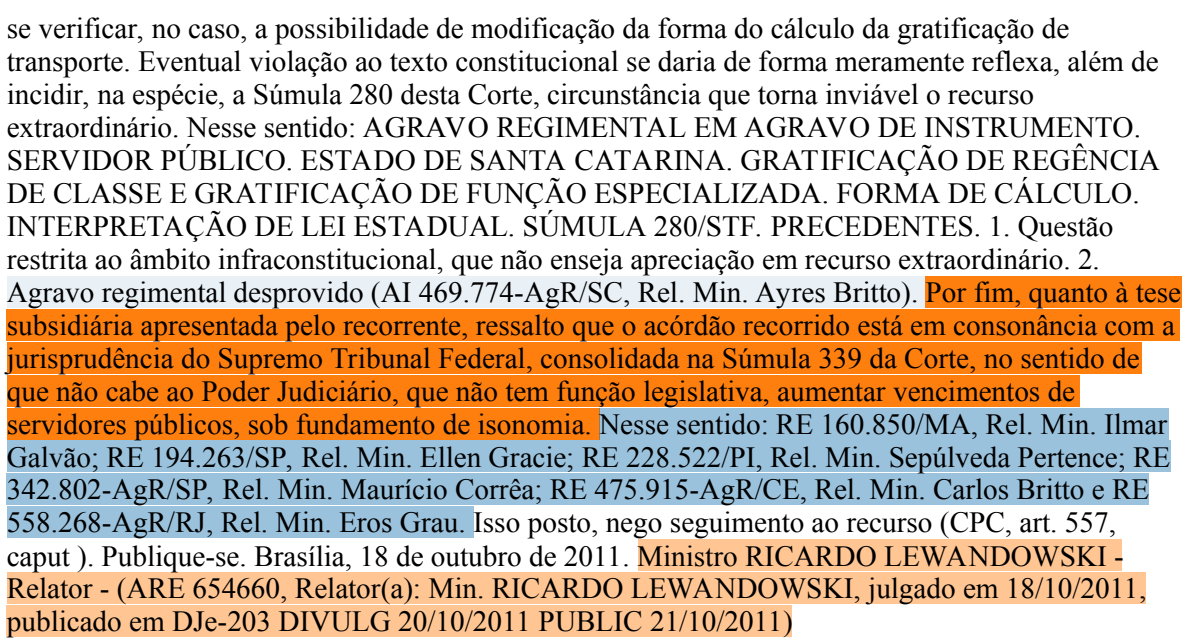
Ministro LUIZ FUX"

What is relevant about this document for our work is the fact that TF-IDF RBF predicts a BP 37 citation with a probability of 99.998%, *i.e.*, the model is sure this

[1] Before searching for the potential citations, we disregarded documents that compose the BP 37 sample dataset.

document cites BP 37. However, the decision (2012) was published before the creation of the statement (2014). We rely on LIME to present the most important sentences of this document for the model's decision. We ask the technique to select ten sentences only through the generation of five thousand disturbances of the original document. Part of the decision with highlighted sentences is in Figure 11. The entire decision is in section C.1. We carefully chose this part because it contains the most important sentence (in intense orange highlight).

Figure 11 – Part of a decision of the process ARE 694870 with highlighted sentences by LIME.



LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. This selected part of the document contains the most important sentence. Source: the authors.

Notice that the most highlighted sentence contains the statement of the BP 37 in Portuguese — the last two lines of the intense orange highlight. However, this could not be possible because the decision was published before the statement. If one looks carefully, one can see that this decision cites *súmula* 339, a non-binding súmula, but which was later converted into BP 37. This is an example of a decision that would have cited the BP 37 if it had existed. However, it did not, so the citation was not possible. Therefore, we characterize it as a potential citation.

A simple regular expression using variations of the BP 37 statement could have found all documents that cite súmula 339 before BP 37 creation without having to perform all of our experiments. Also, a specialist dealing with BP 37-related documents already

knows that BP 37 came from the conversion of súmula 339. However, the main point of this case study is to show that the methodology of inference and explanation is able to produce and explain coherent and valid potential citation assignments. In special, it shows that the model learned to use the BP statement, which is not always present in the documents, to search for citations.

8.2 Binding precedent 4

We use the TF-IDF RBF model trained to identify BP 4 citations to search for potential citations in our unlabeled dataset. The model identifies a citation in a decision of Justice Joaquim Barbosa from 03/06/2012 in the process ARE 637053. Although the document does not explicitly cite the BP, the model is 99.49% sure of this assignment. Figure 12 presents an excerpt of this decision with the highlighted sentences by LIME, similarly to section 8.1. The entire decision is in section C.2.

Figure 12 – Part of a decision of the process ARE 637053 with highlighted sentences by LIME.

I. - Indenização vinculada ao salário-mínimo: impossibilidade. C.F., art. 7º, IV. O que a Constituição veda - art. 7º, IV - é a fixação do quantum da indenização em múltiplo de salários-mínimos. STF, RE 225.488/PR, Moreira Alves; ADI 1.425. A indenização pode ser fixada, entretanto, em salários-mínimos, observado o valor deste na data do julgamento. A partir daí, esse quantum será corrigido por índice oficial. II. - Provimento parcial do agravo: RE conhecido e provido, em parte." (RE 409.427-AgR, rel. min. Carlos Velloso, Segunda Turma, DJ 02.04.2004) "EMENTA: Vinculação ao salário mínimo: a vedação do art. 7º, IV, da Constituição, restringe-se à hipótese em que se pretenda fazer das elevações futuras do salário mínimo índice de atualização da indenização fixada; não, qual se deu no acórdão recorrido, se o múltiplo do salário mínimo é utilizado apenas para expressar o valor inicial da condenação, a ser atualizado, se for o caso, conforme os índices oficiais da correção monetária". (RE 389.989, rel. min. Sepúlveda Pertence, Primeira Turma, DJ 05.11.2004). No mesmo sentido, o AI 493.494-AgR (rel. min. Gilmar Mendes, DJ 04.03.2005) e o AI 510.244-AgR (rel. min. Cezar Peluso, DJ 04.03.2005).

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. This selected part of the document contains the most important sentence. Source: the authors.

BP 4 has the following statement in Portuguese: *"Salvo nos casos previstos na Constituição, o salário mínimo não pode ser usado como indexador de base de cálculo de vantagem de servidor público ou de empregado, nem ser substituído por decisão judicial."* A free translation to English is: *"Except in the cases provided for in the Constitution, the minimum wage cannot be used as a basis for calculating the benefit of public servants or employees, nor can it be replaced by a court decision."* Comparing the most important

sentence[2] (Figure 12, intense orange highlight) with the original BP statement, both refer to the binding to the minimum wage. In the case of the BP, it deals with labor law; in the case of the document, it deals with civil law, specifically in the collection of advocative hours. Although the document could not cite the BP because it was not the case, both of them considered subjects similar enough to convince the model of a citation — bindings to the minimum wage. The document also brings a citation to art. 7th, IV, from the Brazilian Constitution, which considers the impossibility of binding to the minimum wage and is a Legislative Reference for the BP 4 (SUPREMO TRIBUNAL FEDERAL, 2021). Therefore, it is possible to consider this as an example of a potential citation.

[2] Note that the intense orange highlighted text contains more than one sentence. This is because the algorithm to split the text into sentences was unable to fully perform the split, probably, among other reasons, due to the particularities of the law field texts.

9 Discussion and limitations

Through an evaluation and a few case studies, we demonstrated our methodology of inferring and explaining potential citations is promising and can assist specialists in analyzing legal documents. However, this work has limitations.

The first limitation is regarded to the data. There is a discrepancy in the number of citations among the 58 existing binding precedents, and it can result in models performing differently for different BPs. If all BPs had enough citations, this should not be a concern. However, they do not. This, associated with the fact there are duplicated documents (documents that share the same raw text), happens to impact our analyses. Models training on a few data probably won't perform well, especially those language models, *e.g.*, BERT and Longformer. A possible solution is to apply data augmentation techniques (*e.g.*, EDA (WEI; ZOU, 2019)), an idea we leave for future work.

A minor limitation also related to the data is that we used documents without any BP citation when constructing the sample datasets. For instance, BP 10 sample dataset contains two types of documents: (a) decisions with BP 10 citations; (b) decisions without any BP citation. In this example, the ideal for item (b) would be decisions without BP 10 citations but allowed to cite the other BPs, respecting the natural conditional probability distribution of the documents that do not cite BP 10. In this work, (a) and (b) do not follow the original proportions of the classes, *i.e.*, few BP citations and many non-BP citations. This is a simple, random undersampling approach to the high imbalance binary classification problem we have (KRAWCZYK, 2016). We plan to explore more sophisticated imbalance classification solutions in future work.

In the analysis of the dataset (chapter 5), it was made clear that the document length distribution is not the same for documents that cite different BPs (or do not cite any of them) (Figure 8). Even though it is not a limitation, because it follows the natural distribution of the data, one may want to have all documents of the same length to avoid having the models utilizing this information. For example, the length of a document could be not related to a BP but to a past Justice who cited this BP several times. If this is the case, it should be straightforward to apply a pre-processing of the data, *e.g.*, splitting the documents, and rerun the experiments.

Because legal documents are indexed by publication date, it is possible to understand a legal corpus as a time series. Splitting the dataset into train, validation, and test without considering the publication date probably artificially improves models' performance during validation and test because the models are being assessed in data from the same date they were trained on. This situation does not happen in real life: models are trained in past data

and need to predict in actual and future data. This adaptation should be straightforward to be implemented in future work.

A major limitation of this work is the use of STF documents only. The methodology of inferring and explaining potential citations cannot be fully explored in STF documents because, in practice, it is not important to know if STF is following its own precedent, but if lower Judiciary branches are. In future work, it should be easy to perform the same experiments using documents from higher courts, *e.g.*, Federal Regional Courts ("Tribunais Regionais Federais," TRFs).

Regarding the models, there are some limitations and possible improvements, which we plan for future work. The metadata, *e.g.*, Justice Rapporteur and publication date, can be passed explicitly during models training. Currently, if the models want to use this information, they have to learn to extract them from the raw texts. In this work, we explored simple embeddings (TF-IDF, Doc2vec, and USE) and state-of-the-art language models (BERT and Longformer). However, we could also explore mid-term sophistication approaches, *e.g.*, Long Short-Term Memory (HOCHREITER; SCHMIDHUBER, 1997). In relation to the language models, we considered a Longformer version specialized in English data, which can prevent the model from performing better in our Brazilian Portuguese documents. It is not straightforward to overcome this limitation, but a possible solution is to effectively build a new Longformer model from BERTimbau's checkpoint. Another available improvement is using pre-trained models specialized for Brazilian Portuguese law documents, *e.g.*, recent LegalNLP (POLO et al., 2021). Finally, we could fine-tune USE similarly to what was done for BERT and Longformer (section 3.2).

The classification strategy adopted in our work is a simple multilabel approach where a classifier is adjusted for each output class (section 6.1). So, if we have n classes (*e.g.*, in our case, 10 BPs), we end with n classifiers. We argued that this is an improvement over the multiclass approach, in which there is only one model with n output probabilities that add up to one. However, our simple multilabel approach is not efficient because it does not scale well for many classes. For instance, in Brazil there are hundreds of binding understandings, like BPs, what makes it infeasible to fully apply our methodology. The application of more efficient multilabel approaches will be explored in future work.

We believe a critical limitation of this work is the absence of a methodological validation of the potential citation assignments (section 7.4). It is not straightforward to perform this validation, both because of the large number of assigned documents and because of the lack of knowledge by the authors in the legal area, so we leave this task for future work. However, the presented case studies (chapter 8), which provided interesting analyses and insights, demonstrate that one can readily benefit from our potential citation methodology.

10 Conclusion

In this work, we presented a methodology of inference and explanation of potential citations to binding precedents (BP) in decisions from the Brazilian Supreme Court (STF). More specifically:

- We collected, analyzed, and processed documents with citations to BPs in STF decisions. We decided to work with the ten most-cited BPs, but the results are naturally extensible to the rest of the BPs and other citation types among legal documents.
- We relied on document embedding techniques and classification algorithms from classical machine learning theory and recent natural language processing research to infer potential citations. The models learned to identify potential citations through their training over documents labeled with explicit citations, in which they achieved high performance. After learning, the best model was used to infer potential citations from unlabeled, unknown data.
- We applied a machine learning explainability technique to explain the best model's potential citation assignments, presenting the most relevant sentences. We demonstrated the model's capacity to find documents with potential citations and the explainability technique's reliability in explaining the model's decisions through two case studies.

These tasks were accomplished by an extensive, although not extraordinarily technical, description of the utilized models, methods, and techniques.

References

ALI, Shawkat; SMITH-MILES, Kate A. Improved Support Vector Machine Generalization Using Normalized Input Space. en. In: SATTAR, Abdul; KANG, Byeong-ho (Eds.). **AI 2006: Advances in Artificial Intelligence**. Berlin, Heidelberg: Springer, 2006. (Lecture Notes in Computer Science), p. 362–371. ISBN 978-3-540-49788-2. DOI: 10.1007/11941439_40. Cit. on p. 44.

BELTAGY, Iz; PETERS, Matthew E.; COHAN, Arman. Longformer: The Long-Document Transformer. **arXiv:2004.05150 [cs]**, Dec. 2020. arXiv: 2004.05150. Available from: <http://arxiv.org/abs/2004.05150>. Visited on: 11 Oct. 2021. Cit. on pp. 17, 21, 24, 50.

BRAZILIAN SUPREME COURT. **Number of Decisions**. [S.l.: s.n.], 2011-2020. Retrieved from the Brazilian Supreme Court statistics, http://portal.stf.jus.br/estatistica/. Accessed in: 2021-10-10. Cit. on p. 13.

CER, Daniel et al. Universal Sentence Encoder. **arXiv:1803.11175 [cs]**, Apr. 2018. arXiv: 1803.11175. Available from: <http://arxiv.org/abs/1803.11175>. Visited on: 11 Oct. 2021. Cit. on pp. 17, 20.

CONSELHO NACIONAL DE JUSTIÇA. **Total Number of Pending Cases**. [S.l.: s.n.], 2004. Data retrieved from the National Council of Justice database, https://www.cnj.jus.br/wp-content/uploads/2013/01/fc587d0cef86f585d6872a4f0ff43107.zip. Accessed in: 2021-10-10. Cit. on p. 13.

DEVLIN, Jacob et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: PROCEEDINGS of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, June 2019. P. 4171–4186. DOI: 10.18653/v1/N19-1423. Available from: <https://aclanthology.org/N19-1423>. Visited on: 24 Oct. 2021. Cit. on pp. 21, 24, 27.

DIMOULIS, Dimitri; LUNARDI, Soraya. **Curso de Processo Constitucional - Controle de constitucionalidade e remédios constitucionais**. 4. ed. São Paulo: Atlas, 2016. OCLC: 992571244. ISBN 978-85-970060-4-9. Cit. on p. 13.

FALCÃO, Joaquim et al. **V Relatório Supremo em números: o foro privilegiado e o Supremo**. [S.l.]: FGV Direito Rio, Mar. 2017. Accepted: 2017-03-24T14:39:00Z. ISBN 978-85-63265-83-8. Available from:

<http://bibliotecadigital.fgv.br/dspace/handle/10438/18097>. Visited on: 14 Nov. 2021. Cit. on p. 37.

HEARST, M.A. et al. Support vector machines. **IEEE Intelligent Systems and their Applications**, v. 13, n. 4, p. 18–28, July 1998. Conference Name: IEEE Intelligent Systems and their Applications. ISSN 2374-9423. DOI: 10.1109/5254.708428. Cit. on pp. 25–27.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long Short-Term Memory. **Neural Computation**, v. 9, n. 8, p. 1735–1780, Nov. 1997. ISSN 0899-7667. DOI: 10.1162/neco.1997.9.8.1735. Available from: <https://doi.org/10.1162/neco.1997.9.8.1735>. Visited on: 12 Dec. 2021. Cit. on p. 58.

IYYER, Mohit et al. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In: PROCEEDINGS of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China: Association for Computational Linguistics, July 2015. P. 1681–1691. DOI: 10.3115/v1/P15-1162. Available from: <https://aclanthology.org/P15-1162>. Visited on: 23 Oct. 2021. Cit. on p. 20.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. **ACM Computing Surveys**, v. 31, n. 3, p. 264–323, Sept. 1999. ISSN 0360-0300. DOI: 10.1145/331499.331504. Available from: <https://doi.org/10.1145/331499.331504>. Visited on: 3 Nov. 2021. Cit. on p. 25.

JAMES, Gareth et al. **An Introduction to Statistical Learning with Applications in R**. 1. ed. [S.l.]: Springer, New York, NY, 2013. v. 112. (Springer Texts in Statistics). ISBN 978-1-4614-7138-7. Available from: <https://doi.org/10.1007/978-1-4614-7138-7>. Cit. on p. 26.

KIM, Yoon. Convolutional Neural Networks for Sentence Classification. In: PROCEEDINGS of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, Oct. 2014. P. 1746–1751. DOI: 10.3115/v1/D14-1181. Available from: <https://aclanthology.org/D14-1181>. Visited on: 28 Nov. 2021. Cit. on p. 20.

KOTSIANTIS, S. B.; ZAHARAKIS, I. D.; PINTELAS, P. E. Machine learning: a review of classification and combining techniques. en. **Artificial Intelligence Review**, v. 26, n. 3, p. 159–190, Nov. 2006. ISSN 1573-7462. DOI: 10.1007/s10462-007-9052-3. Available from: <https://doi.org/10.1007/s10462-007-9052-3>. Visited on: 3 Nov. 2021. Cit. on p. 25.

KRAWCZYK, Bartosz. Learning from imbalanced data: open challenges and future directions. en. **Progress in Artificial Intelligence**, v. 5, n. 4, p. 221–232, Nov. 2016. ISSN 2192-6360. DOI: 10.1007/s13748-016-0094-0. Available from: <https://doi.org/10.1007/s13748-016-0094-0>. Visited on: 12 Dec. 2021. Cit. on pp. 43, 57.

LE, Quoc; MIKOLOV, Tomas. Distributed Representations of Sentences and Documents. en. In: PROCEEDINGS of the 31st International Conference on Machine Learning. [S.l.]: PMLR, June 2014. P. 1188–1196. ISSN: 1938-7228. Available from: <https://proceedings.mlr.press/v32/le14.html>. Visited on: 11 Oct. 2021. Cit. on pp. 17, 18.

LEBRET, Rémi; COLLOBERT, Ronan. Word Embeddings through Hellinger PCA. en. In: PROCEEDINGS of the 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden: Association for Computational Linguistics, 2014. P. 482–490. DOI: 10.3115/v1/E14-1051. Available from: <http://aclweb.org/anthology/E14-1051>. Visited on: 11 Oct. 2021. Cit. on p. 17.

LESKOVEC, Jure; RAJARAMAN, Anand; ULLMAN, Jeffrey D. **Mining of Massive Datasets**. 3. ed. [S.l.: s.n.], 2020. Available from: <http://infolab.stanford.edu/~ullman/mmds/book0n.pdf>. Visited on: 11 Oct. 2021. Cit. on p. 17.

LIU, Yinhan et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach. **arXiv:1907.11692 [cs]**, July 2019. arXiv: 1907.11692. Available from: <http://arxiv.org/abs/1907.11692>. Visited on: 31 Oct. 2021. Cit. on p. 24.

LOSHCHILOV, Ilya; HUTTER, Frank. Decoupled Weight Decay Regularization. en. In. Available from: <https://openreview.net/forum?id=Bkg6RiCqY7>. Visited on: 21 Nov. 2021. Cit. on p. 45.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to information retrieval**. New York: Cambridge University Press, 2008. OCLC: ocn190786122. ISBN 978-0-521-86571-5. Cit. on p. 44.

MIKOLOV, Tomas et al. Efficient Estimation of Word Representations in Vector Space. **arXiv:1301.3781 [cs]**, Sept. 2013. arXiv: 1301.3781. Available from: <http://arxiv.org/abs/1301.3781>. Visited on: 11 Oct. 2021. Cit. on p. 17.

OLSON, David; DELEN, Dursun. **Advanced Data Mining Techniques**. [S.l.: s.n.], Jan. 2008. Journal Abbreviation: Springer. USA Publication Title: Springer. USA. ISBN 978-3-540-76916-3. DOI: 10.1007/978-3-540-76917-0. Cit. on p. 28.

PLATT, John C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: ADVANCES in Large Margin Classifiers. [S.l.]: MIT Press, 1999. P. 61–74. Cit. on p. 45.

POLO, Felipe Maia et al. LegalNLP – Natural Language Processing methods for the Brazilian Legal Language. **arXiv:2110.15709 [cs]**, Oct. 2021. arXiv: 2110.15709. Available from: <http://arxiv.org/abs/2110.15709>. Visited on: 30 Nov. 2021. Cit. on p. 58.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: PROCEEDINGS of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: Association for Computing Machinery, Aug. 2016. (KDD '16), p. 1135–1144. ISBN 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. Available from: <https://doi.org/10.1145/2939672.2939778>. Visited on: 13 Nov. 2021. Cit. on pp. 31, 46.

RUMELHART, David E.; HINTON, Geoffrey E.; WILLIAMS, Ronald J. Learning representations by back-propagating errors. en. **Nature**, v. 323, n. 6088, p. 533–536, Oct. 1986. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 6088 Primary_atype: Research Publisher: Nature Publishing Group. ISSN 1476-4687. DOI: 10.1038/323533a0. Available from: <https://www.nature.com/articles/323533a0>. Visited on: 17 Oct. 2021. Cit. on p. 20.

SOUZA, Fábio; NOGUEIRA, Rodrigo; LOTUFO, Roberto. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. en. In_____. **Intelligent Systems**. Cham: Springer International Publishing, 2020. (Lecture Notes in Computer Science), p. 403–417. ISBN 978-3-030-61377-8. DOI: 10.1007/978-3-030-61377-8_28. Available from: <https://link.springer.com/chapter/10.1007%2F978-3-030-61377-8_28>. Cit. on pp. 24, 44.

SUPREMO TRIBUNAL FEDERAL. **Súmula vinculante 4**. Portuguese. [S.l.: s.n.]. Available from: <https://jurisprudencia.stf.jus.br/pages/search/seq-sumula740/false>. Visited on: 12 Dec. 2021. Cit. on p. 56.

TAVARES, André Ramos. Súmula vinculante. pt-br. **Dicionário brasileiro de direito constitucional**, p. 366–367, 2007. Available from: <https://repositorio.usp.br/item/001712199>. Visited on: 10 Oct. 2021. Cit. on p. 13.

TUCCI, José Rogério Cruz e. Precedente judicial como fonte do direito. pt-br. **Revista dos Tribunais**, 2004. Available from: <https://repositorio.usp.br/item/001397530>. Visited on: 10 Oct. 2021. Cit. on p. 13.

VASWANI, Ashish et al. Attention is All You Need. In: PROCEEDINGS of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY,

USA: Curran Associates Inc., 2017. (NIPS'17), p. 6000–6010. event-place: Long Beach, California, USA. ISBN 978-1-5108-6096-4. Cit. on pp. 20–22.

WEI, Jason; ZOU, Kai. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In: PROCEEDINGS of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, Nov. 2019. P. 6382–6388. DOI: 10.18653/v1/D19-1670. Available from: <https://aclanthology.org/D19-1670>. Visited on: 30 Nov. 2021. Cit. on p. 57.

YANG, Yinfei et al. Multilingual Universal Sentence Encoder for Semantic Retrieval. In: PROCEEDINGS of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Online: Association for Computational Linguistics, July 2020. P. 87–94. DOI: 10.18653/v1/2020.acl-demos.12. Available from: <https://aclanthology.org/2020.acl-demos.12>. Visited on: 28 Nov. 2021. Cit. on p. 20.

Appendix

APPENDIX A – Quantitative evaluation tables

Table 9 – Accuracy (in percentage) of each model on validation data.

Model	Binding precedent									
	3	**4**	**10**	**11**	**14**	**17**	**20**	**26**	**33**	**37**
TF-IDF RBF	**98.4**	**98.7**	96.4	**97.0**	96.8	97.0	**99.3**	**99.0**	97.9	**99.1**
BERT	96.6	97.8	97.3	96.1	96.8	95.5	99.1	98.3	**98.5**	98.9
TF-IDF linear	98.1	97.3	95.3	95.7	96.8	**98.5**	99.1	98.7	97.0	98.9
Longformer	88.2	**98.7**	**98.8**	95.3	**98.7**	87.9	98.4	93.7	97.0	99.0
USE RBF	97.5	96.9	96.2	95.7	97.7	95.8	98.2	98.0	97.6	98.9
USE linear	97.5	96.2	94.6	94.8	97.2	95.8	97.8	98.0	97.3	98.6
Doc2Vec RBF	95.0	94.1	89.5	93.5	93.6	92.5	95.9	95.3	94.3	96.5
Doc2Vec linear	94.4	92.3	85.7	88.4	91.7	89.4	93.7	92.7	92.2	93.2

For each embedding, classifier, and BP sample dataset, the table presents the accuracy (in percentage) on validation data. BERT is evaluated using maximum logit (subsection 6.1.3). Bold values are the best values of the column. Source: the authors.

Table 10 – Precision (in percentage) of each model on validation data.

Model	Binding precedent									
	3	**4**	**10**	**11**	**14**	**17**	**20**	**26**	**33**	**37**
TF-IDF linear	97.5	97.7	96.8	94.9	94.7	**100.0**	**99.6**	98.0	**98.2**	98.4
TF-IDF RBF	**98.1**	98.2	97.4	97.4	95.1	97.7	**99.6**	**98.7**	**98.2**	98.8
Longformer	90.2	**99.1**	**99.4**	95.7	**98.3**	86.9	98.6	95.2	95.4	98.4
Doc2Vec RBF	95.0	97.0	92.3	**99.0**	94.4	97.5	98.1	97.2	96.9	98.1
USE RBF	**98.1**	95.3	97.0	96.5	97.1	95.5	98.2	96.2	97.1	**98.9**
BERT	94.6	96.0	95.1	94.2	94.4	92.3	98.6	**98.7**	97.7	97.8
USE linear	**98.1**	94.8	95.1	94.1	96.7	95.5	98.2	96.2	97.0	98.6
Doc2Vec linear	96.7	97.4	91.9	94.9	94.2	95.7	95.9	97.1	95.5	98.6

For each embedding, classifier, and BP sample dataset, the table presents the precision (in percentage) on validation data. BERT is evaluated using maximum logit (subsection 6.1.3). Bold values are the best values of the column. Source: the authors.

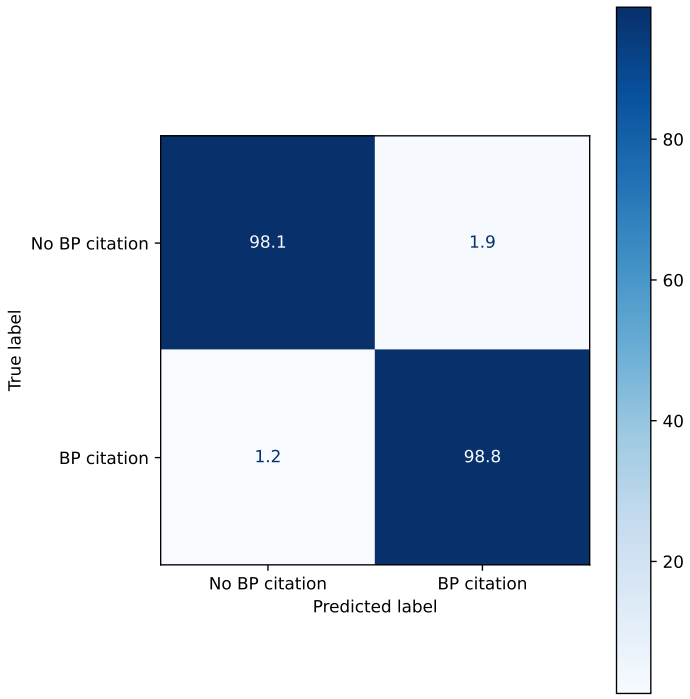
Table 11 – Recall (in percentage) of each model on validation data.

Model	Binding precedent									
	3	**4**	**10**	**11**	**14**	**17**	**20**	**26**	**33**	**37**
BERT	**98.8**	**99.8**	**99.8**	**98.3**	**99.6**	**99.2**	**99.6**	98.0	**99.4**	**100.0**
USE linear	96.9	97.7	94.1	95.7	97.9	96.2	97.5	**100.0**	97.6	98.6
USE RBF	96.9	98.6	95.5	94.8	98.3	96.2	98.2	**100.0**	98.2	98.9
Longformer	85.7	98.4	98.3	94.8	99.2	89.5	98.2	92.0	98.8	99.5
TF-IDF linear	**98.8**	97.0	93.8	96.6	99.2	97.0	98.6	99.3	95.8	99.4
TF-IDF RBF	**98.8**	99.3	95.3	96.6	98.7	96.2	98.9	99.3	97.6	99.4
Doc2Vec RBF	95.0	91.0	86.2	87.9	92.8	87.2	93.5	93.3	91.7	95.0
Doc2Vec linear	91.9	86.8	78.3	81.0	89.0	82.7	91.4	88.0	88.7	87.6

For each embedding, classifier, and BP sample dataset, the table presents the recall (in percentage) on validation data. BERT is evaluated using maximum logit (subsection 6.1.3). Bold values are the best values of the column. Source: the authors.

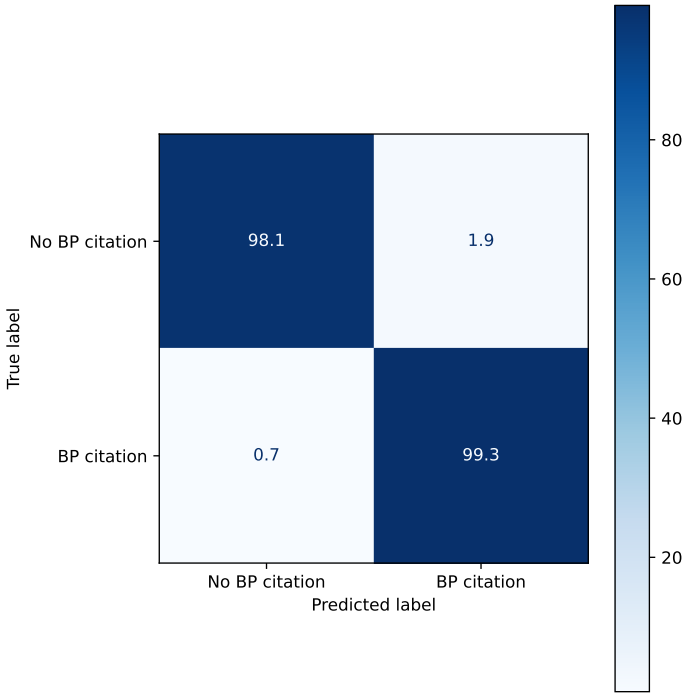
APPENDIX B – TF-IDF RBF validation confusion matrices

Figure 13 – TF-IDF RBF for BP 3 validation confusion matrix (in percentage).



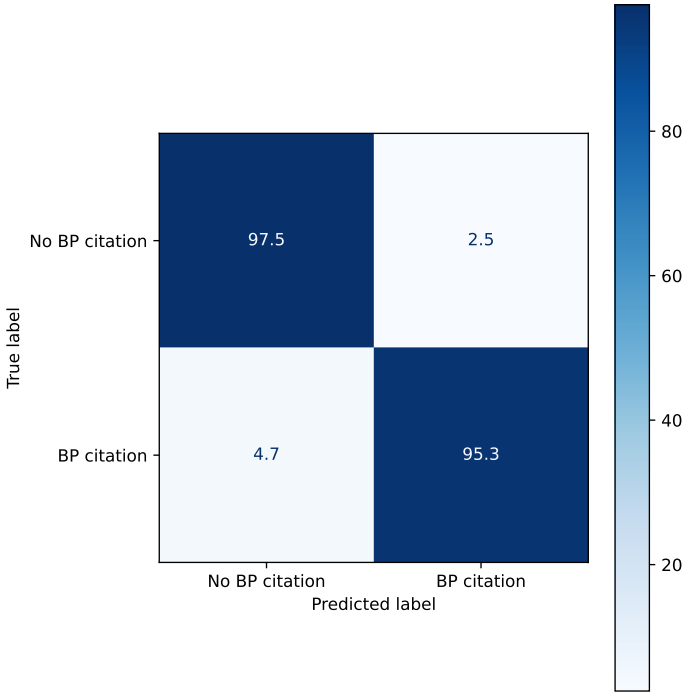
Source: the authors.

Figure 14 – TF-IDF RBF for BP 4 validation confusion matrix (in percentage).



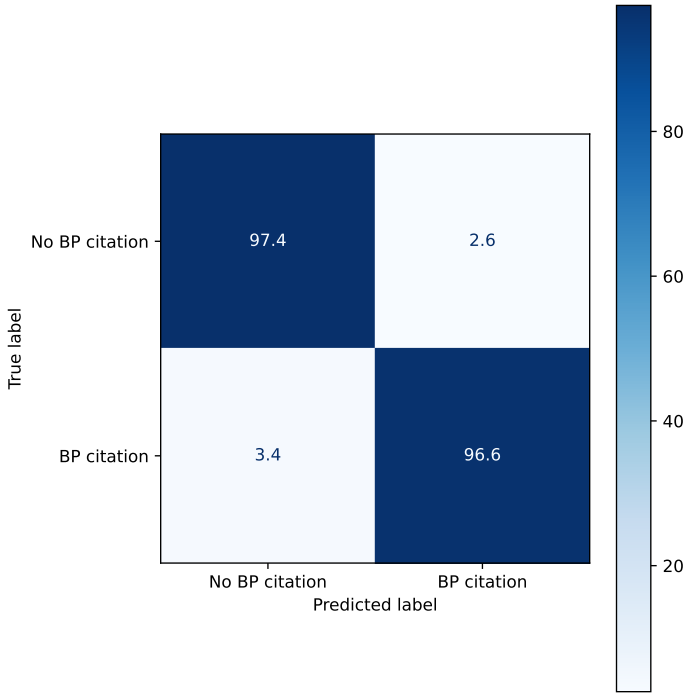
Source: the authors.

Figure 15 – TF-IDF RBF for BP 10 validation confusion matrix (in percentage).



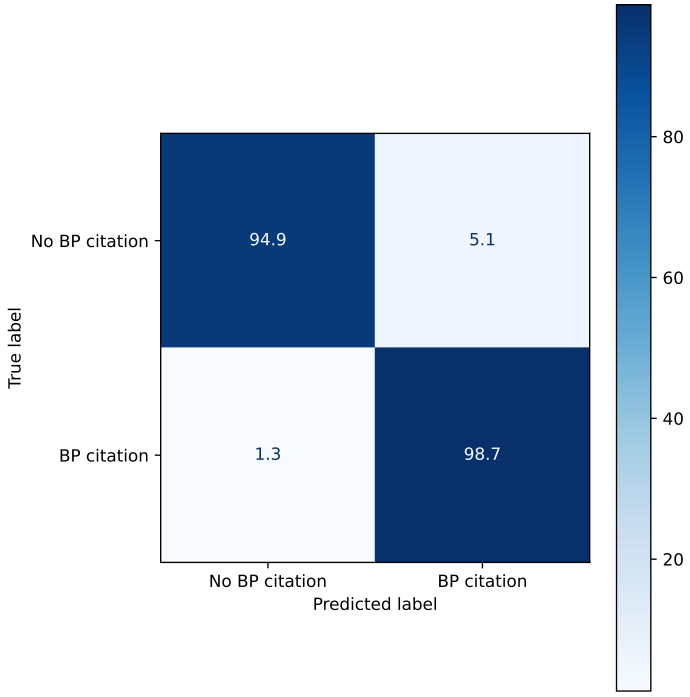
Source: the authors.

Figure 16 – TF-IDF RBF for BP 11 validation confusion matrix (in percentage).



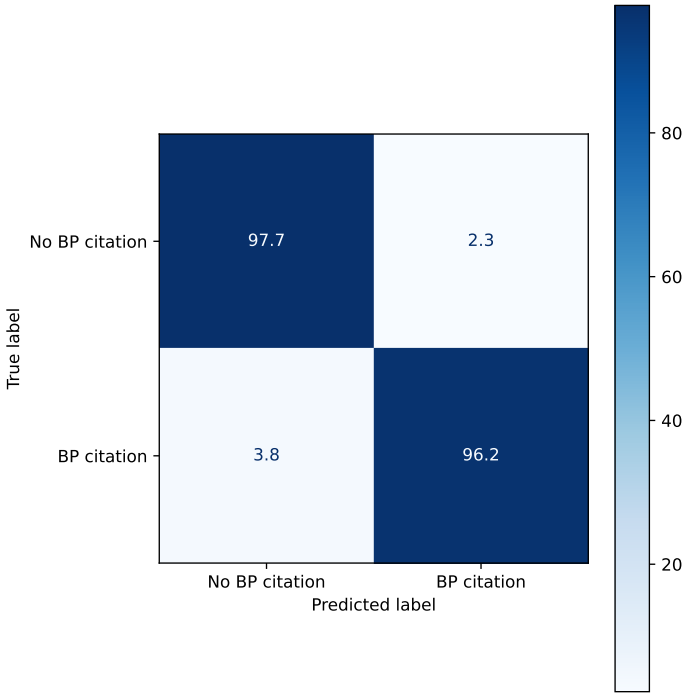
Source: the authors.

Figure 17 – TF-IDF RBF for BP 14 validation confusion matrix (in percentage).



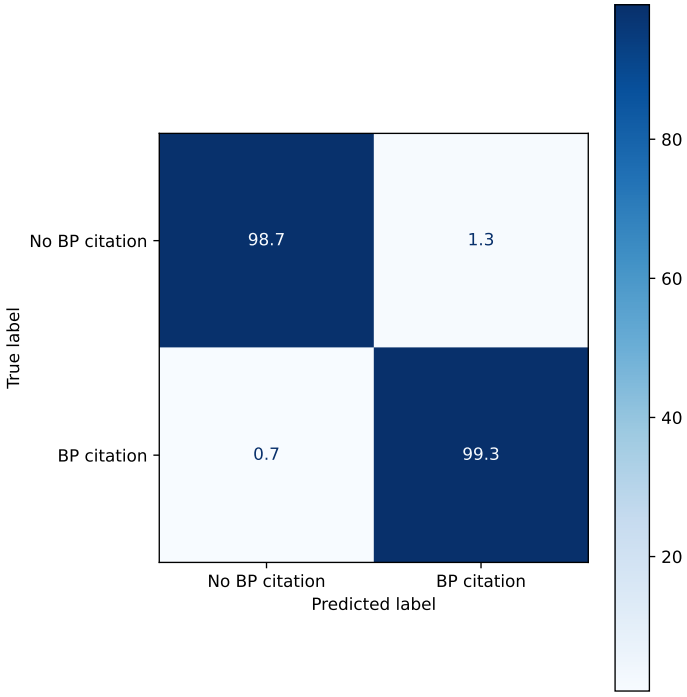
Source: the authors.

Figure 18 – TF-IDF RBF for BP 17 validation confusion matrix (in percentage).



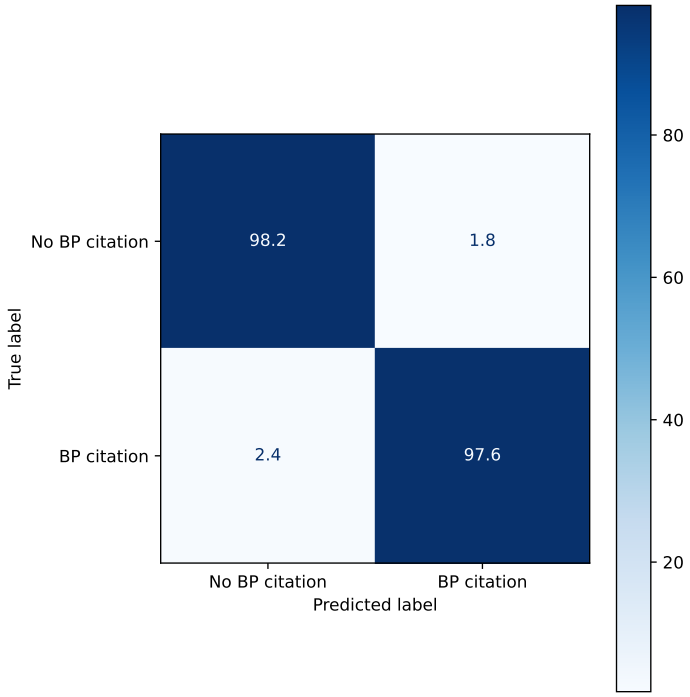
Source: the authors.

Figure 19 – TF-IDF RBF for BP 26 validation confusion matrix (in percentage).



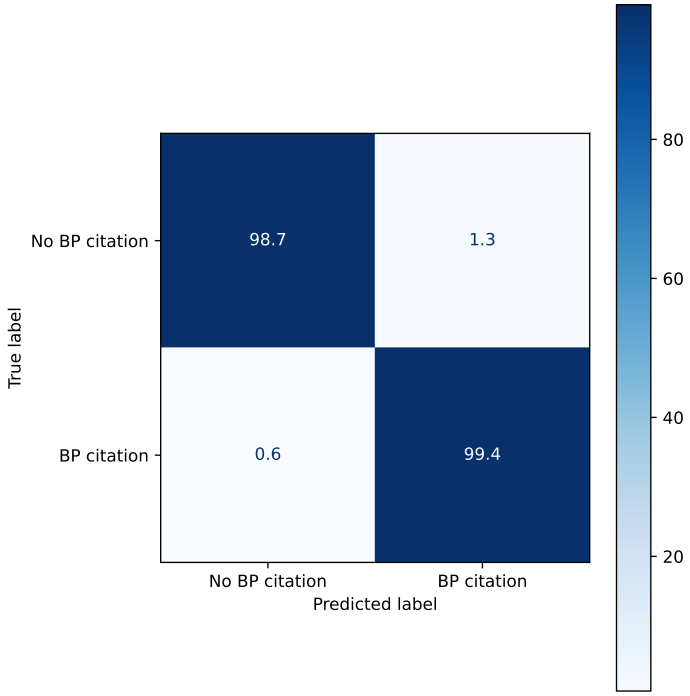
Source: the authors.

Figure 20 – TF-IDF RBF for BP 33 validation confusion matrix (in percentage).



Source: the authors.

Figure 21 – TF-IDF RBF for BP 37 validation confusion matrix (in percentage).

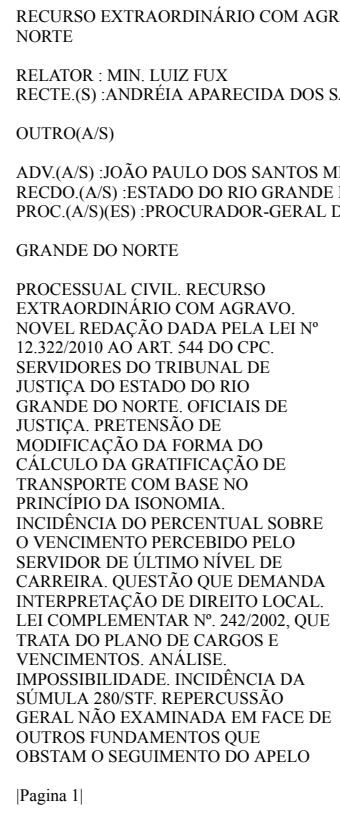


Source: the authors.

APPENDIX C – Documents of the case studies

C.1 ARE 694870

Figure 22 – A decision of the process ARE 694870, part 1.



RECURSO EXTRAORDINÁRIO COM AGRAVO 694.870 RIO GRANDE DO NORTE

RELATOR : MIN. LUIZ FUX
RECTE.(S) :ANDRÉIA APARECIDA DOS SANTOS SALES E

OUTRO(A/S)

ADV.(A/S) :JOÃO PAULO DOS SANTOS MELO E OUTRO(A/S)
RECDO.(A/S) :ESTADO DO RIO GRANDE DO NORTE
PROC.(A/S)(ES) :PROCURADOR-GERAL DO ESTADO DO RIO

GRANDE DO NORTE

PROCESSUAL CIVIL. RECURSO EXTRAORDINÁRIO COM AGRAVO. NOVEL REDAÇÃO DADA PELA LEI Nº 12.322/2010 AO ART. 544 DO CPC. SERVIDORES DO TRIBUNAL DE JUSTIÇA DO ESTADO DO RIO GRANDE DO NORTE. OFICIAIS DE JUSTIÇA. PRETENSÃO DE MODIFICAÇÃO DA FORMA DO CÁLCULO DA GRATIFICAÇÃO DE TRANSPORTE COM BASE NO PRINCÍPIO DA ISONOMIA. INCIDÊNCIA DO PERCENTUAL SOBRE O VENCIMENTO PERCEBIDO PELO SERVIDOR DE ÚLTIMO NÍVEL DE CARREIRA. QUESTÃO QUE DEMANDA INTERPRETAÇÃO DE DIREITO LOCAL. LEI COMPLEMENTAR Nº. 242/2002, QUE TRATA DO PLANO DE CARGOS E VENCIMENTOS. ANÁLISE. IMPOSSIBILIDADE. INCIDÊNCIA DA SÚMULA 280/STF. REPERCUSSÃO GERAL NÃO EXAMINADA EM FACE DE OUTROS FUNDAMENTOS QUE OBSTAM O SEGUIMENTO DO APELO

|Pagina 1|

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. Source: the authors.

Figure 23 – A decision of the process ARE 694870, part 2.

ARE 694.870 / RN

EXTREMO.
1. A repercussão geral pressupõe recurso admissível sob o crivo dos demais requisitos constitucionais e processuais de admissibilidade (art. 323 do RISTF). Consectariamente, se o recurso é inadmissível por outro motivo, não há como se pretender seja reconhecida "a repercussão geral das questões constitucionais discutidas no caso" (art. 102, III, § 3º, da CF).
2. A ofensa ao direito local não viabiliza o apelo extremo.
3. Na hipótese sub judice, a questão relativa ao aumento da gratificação de transporte, percebida pelos servidores do Poder Judiciário do Estado do Rio Grande do Norte, encontra-se disciplinada na Lei Complementar Estadual n º 242/2002, o que obsta a abertura da via excepcional ante o teor da súmula nº. 280/STF, que assim dispõe, verbis: Por ofensa a direito local não cabe recurso extraordinário.) Precedentes.
4. In casu, o acórdão recorrido assentou: CONSTITUCIONAL E ADMINISTRATIVO. SERVIDORES DO PODER JUDICIÁRIO DO ESTADO DO RIO GRANDE DO NORTE. OFICIAIS DE JUSTIÇA. PRETENSÃO DE MODIFICAÇÃO DA FORMA DO CÁLCULO DA GRATIFICAÇÃO DE TRANSPORTE, COM BASE NO PRINCÍPIO DA ISONOMIA. LEI COMPLEMENTAR ESTADUAL Nº

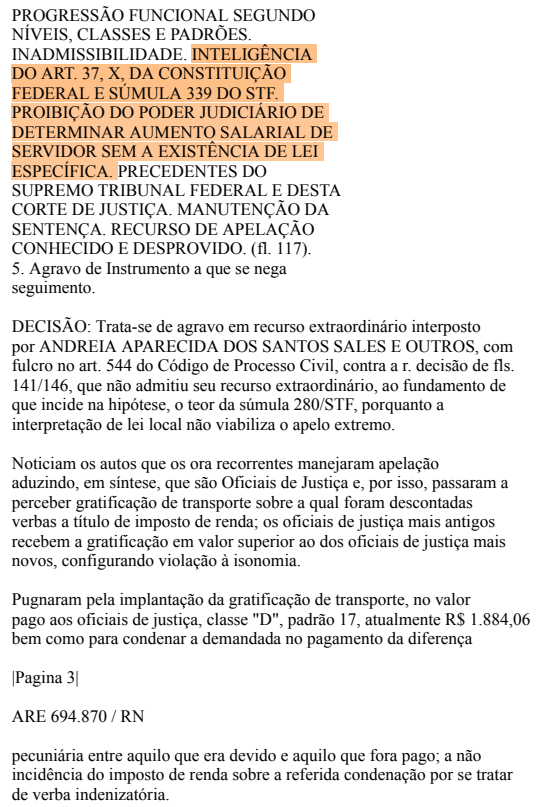
|Pagina 2|

ARE 694.870 / RN

242/2002 QUE TRATA DO PLANO DE CARGOS E VENCIMENTOS QUE PREVÊ

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. Source: the authors.

Figure 24 – A decision of the process ARE 694870, part 3.



PROGRESSÃO FUNCIONAL SEGUNDO NÍVEIS, CLASSES E PADRÕES. INADMISSIBILIDADE. INTELIGÊNCIA DO ART. 37, X, DA CONSTITUIÇÃO FEDERAL E SÚMULA 339 DO STF. PROIBIÇÃO DO PODER JUDICIÁRIO DE DETERMINAR AUMENTO SALARIAL DE SERVIDOR SEM A EXISTÊNCIA DE LEI ESPECÍFICA. PRECEDENTES DO SUPREMO TRIBUNAL FEDERAL E DESTA CORTE DE JUSTIÇA. MANUTENÇÃO DA SENTENÇA. RECURSO DE APELAÇÃO CONHECIDO E DESPROVIDO. (fl. 117).
5. Agravo de Instrumento a que se nega seguimento.

DECISÃO: Trata-se de agravo em recurso extraordinário interposto por ANDREIA APARECIDA DOS SANTOS SALES E OUTROS, com fulcro no art. 544 do Código de Processo Civil, contra a r. decisão de fls. 141/146, que não admitiu seu recurso extraordinário, ao fundamento de que incide na hipótese, o teor da súmula 280/STF, porquanto a interpretação de lei local não viabiliza o apelo extremo.

Noticiam os autos que os ora recorrentes manejaram apelação aduzindo, em síntese, que são Oficiais de Justiça e, por isso, passaram a perceber gratificação de transporte sobre a qual foram descontadas verbas a título de imposto de renda; os oficiais de justiça mais antigos recebem a gratificação em valor superior ao dos oficiais de justiça mais novos, configurando violação à isonomia.

Pugnaram pela implantação da gratificação de transporte, no valor pago aos oficiais de justiça, classe "D", padrão 17, atualmente R$ 1.884,06 bem como para condenar a demandada no pagamento da diferença

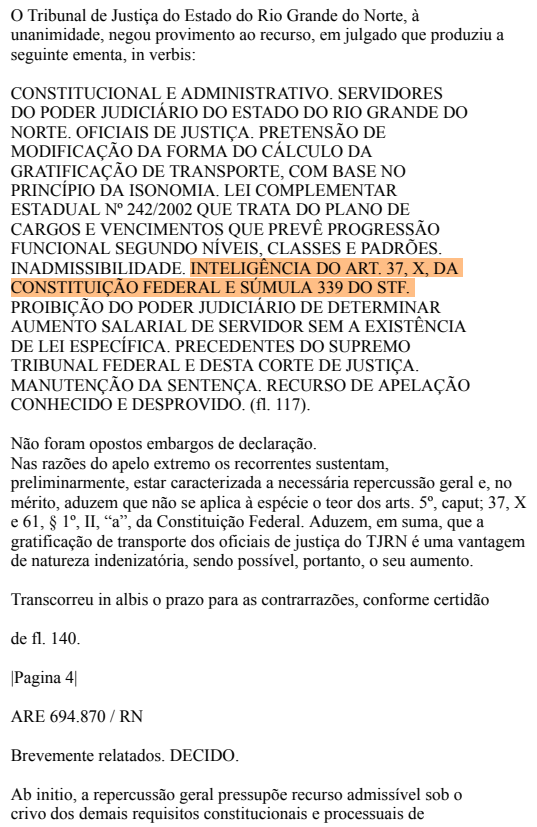
|Pagina 3|

ARE 694.870 / RN

pecuniária entre aquilo que era devido e aquilo que fora pago; a não incidência do imposto de renda sobre a referida condenação por se tratar de verba indenizatória.

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. Source: the authors.

Figure 25 – A decision of the process ARE 694870, part 4.



O Tribunal de Justiça do Estado do Rio Grande do Norte, à unanimidade, negou provimento ao recurso, em julgado que produziu a seguinte ementa, in verbis:

CONSTITUCIONAL E ADMINISTRATIVO. SERVIDORES DO PODER JUDICIÁRIO DO ESTADO DO RIO GRANDE DO NORTE. OFICIAIS DE JUSTIÇA. PRETENSÃO DE MODIFICAÇÃO DA FORMA DO CÁLCULO DA GRATIFICAÇÃO DE TRANSPORTE, COM BASE NO PRINCÍPIO DA ISONOMIA. LEI COMPLEMENTAR ESTADUAL Nº 242/2002 QUE TRATA DO PLANO DE CARGOS E VENCIMENTOS QUE PREVÊ PROGRESSÃO FUNCIONAL SEGUNDO NÍVEIS, CLASSES E PADRÕES. INADMISSIBILIDADE. INTELIGÊNCIA DO ART. 37, X, DA CONSTITUIÇÃO FEDERAL E SÚMULA 339 DO STF. PROIBIÇÃO DO PODER JUDICIÁRIO DE DETERMINAR AUMENTO SALARIAL DE SERVIDOR SEM A EXISTÊNCIA DE LEI ESPECÍFICA. PRECEDENTES DO SUPREMO TRIBUNAL FEDERAL E DESTA CORTE DE JUSTIÇA. MANUTENÇÃO DA SENTENÇA. RECURSO DE APELAÇÃO CONHECIDO E DESPROVIDO. (fl. 117).

Não foram opostos embargos de declaração.
Nas razões do apelo extremo os recorrentes sustentam, preliminarmente, estar caracterizada a necessária repercussão geral e, no mérito, aduzem que não se aplica à espécie o teor dos arts. 5º, caput; 37, X e 61, § 1º, II, "a", da Constituição Federal. Aduzem, em suma, que a gratificação de transporte dos oficiais de justiça do TJRN é uma vantagem de natureza indenizatória, sendo possível, portanto, o seu aumento.

Transcorreu in albis o prazo para as contrarrazões, conforme certidão

de fl. 140.

|Pagina 4|

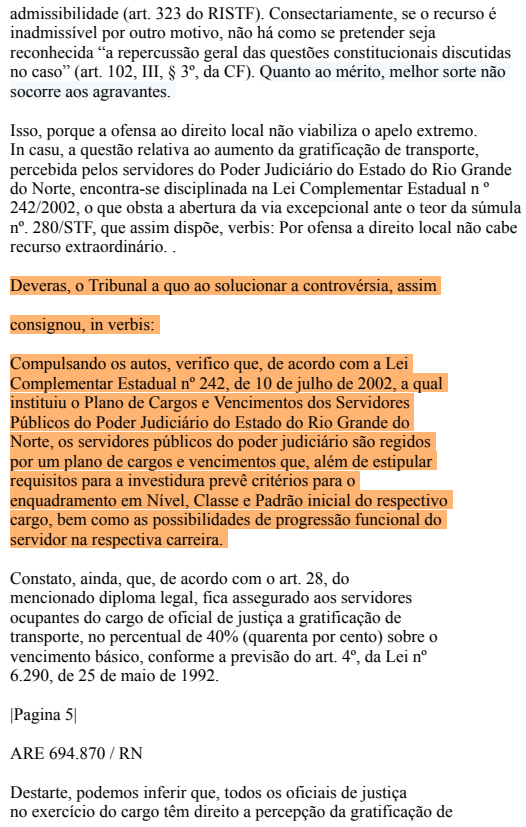
ARE 694.870 / RN

Brevemente relatados. DECIDO.

Ab initio, a repercussão geral pressupõe recurso admissível sob o crivo dos demais requisitos constitucionais e processuais de

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. Source: the authors.

Figure 26 – A decision of the process ARE 694870, part 5.



admissibilidade (art. 323 do RISTF). Consectariamente, se o recurso é inadmissível por outro motivo, não há como se pretender seja reconhecida "a repercussão geral das questões constitucionais discutidas no caso" (art. 102, III, § 3º, da CF). Quanto ao mérito, melhor sorte não socorre aos agravantes.

Isso, porque a ofensa ao direito local não viabiliza o apelo extremo. In casu, a questão relativa ao aumento da gratificação de transporte, percebida pelos servidores do Poder Judiciário do Estado do Rio Grande do Norte, encontra-se disciplinada na Lei Complementar Estadual n º 242/2002, o que obsta a abertura da via excepcional ante o teor da súmula nº. 280/STF, que assim dispõe, verbis: Por ofensa a direito local não cabe recurso extraordinário. .

Deveras, o Tribunal a quo ao solucionar a controvérsia, assim

consignou, in verbis:

Compulsando os autos, verifico que, de acordo com a Lei Complementar Estadual nº 242, de 10 de julho de 2002, a qual instituiu o Plano de Cargos e Vencimentos dos Servidores Públicos do Poder Judiciário do Estado do Rio Grande do Norte, os servidores públicos do poder judiciário são regidos por um plano de cargos e vencimentos que, além de estipular requisitos para a investidura prevê critérios para o enquadramento em Nível, Classe e Padrão inicial do respectivo cargo, bem como as possibilidades de progressão funcional do servidor na respectiva carreira.

Constato, ainda, que, de acordo com o art. 28, do mencionado diploma legal, fica assegurado aos servidores ocupantes do cargo de oficial de justiça a gratificação de transporte, no percentual de 40% (quarenta por cento) sobre o vencimento básico, conforme a previsão do art. 4º, da Lei nº 6.290, de 25 de maio de 1992.

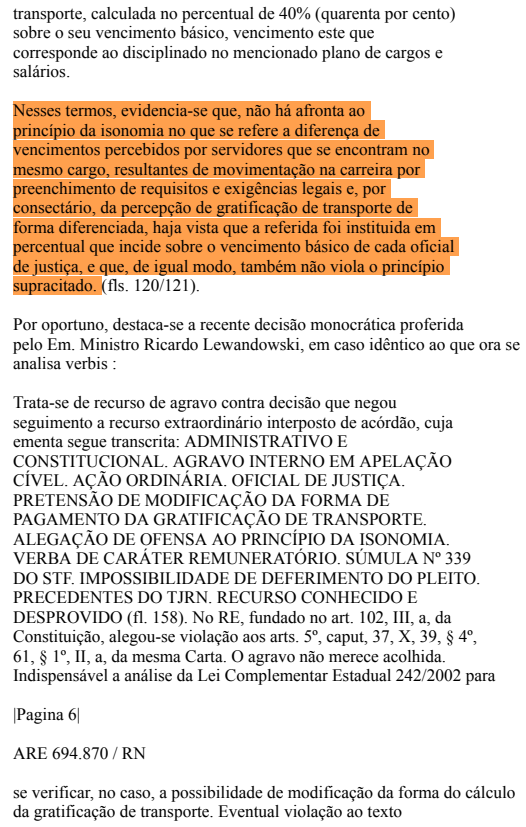
|Pagina 5|

ARE 694.870 / RN

Destarte, podemos inferir que, todos os oficiais de justiça no exercício do cargo têm direito a percepção da gratificação de

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. Source: the authors.

Figure 27 – A decision of the process ARE 694870, part 6.



transporte, calculada no percentual de 40% (quarenta por cento) sobre o seu vencimento básico, vencimento este que corresponde ao disciplinado no mencionado plano de cargos e salários.

Nesses termos, evidencia-se que, não há afronta ao princípio da isonomia no que se refere a diferença de vencimentos percebidos por servidores que se encontram no mesmo cargo, resultantes de movimentação na carreira por preenchimento de requisitos e exigências legais e, por consectário, da percepção de gratificação de transporte de forma diferenciada, haja vista que a referida foi instituida em percentual que incide sobre o vencimento básico de cada oficial de justiça, e que, de igual modo, também não viola o princípio supracitado. (fls. 120/121).

Por oportuno, destaca-se a recente decisão monocrática proferida pelo Em. Ministro Ricardo Lewandowski, em caso idêntico ao que ora se analisa verbis :

Trata-se de recurso de agravo contra decisão que negou seguimento a recurso extraordinário interposto de acórdão, cuja ementa segue transcrita: ADMINISTRATIVO E CONSTITUCIONAL. AGRAVO INTERNO EM APELAÇÃO CÍVEL. AÇÃO ORDINÁRIA. OFICIAL DE JUSTIÇA. PRETENSÃO DE MODIFICAÇÃO DA FORMA DE PAGAMENTO DA GRATIFICAÇÃO DE TRANSPORTE. ALEGAÇÃO DE OFENSA AO PRINCÍPIO DA ISONOMIA. VERBA DE CARÁTER REMUNERATÓRIO. SÚMULA Nº 339 DO STF. IMPOSSIBILIDADE DE DEFERIMENTO DO PLEITO. PRECEDENTES DO TJRN. RECURSO CONHECIDO E DESPROVIDO (fl. 158). No RE, fundado no art. 102, III, a, da Constituição, alegou-se violação aos arts. 5º, caput, 37, X, 39, § 4º, 61, § 1º, II, a, da mesma Carta. O agravo não merece acolhida. Indispensável a análise da Lei Complementar Estadual 242/2002 para

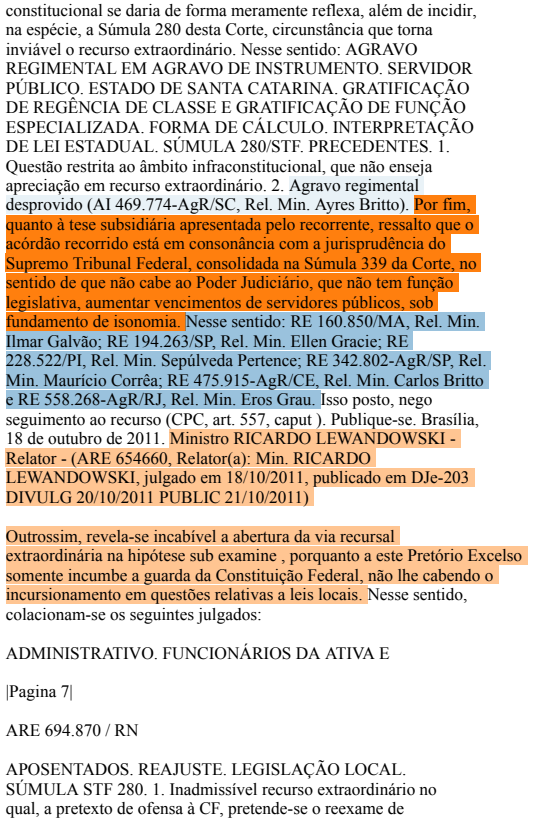
|Pagina 6|

ARE 694.870 / RN

se verificar, no caso, a possibilidade de modificação da forma do cálculo da gratificação de transporte. Eventual violação ao texto

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. Source: the authors.

Figure 28 – A decision of the process ARE 694870, part 7.



constitucional se daria de forma meramente reflexa, além de incidir, na espécie, a Súmula 280 desta Corte, circunstância que torna inviável o recurso extraordinário. Nesse sentido: AGRAVO REGIMENTAL EM AGRAVO DE INSTRUMENTO. SERVIDOR PÚBLICO. ESTADO DE SANTA CATARINA. GRATIFICAÇÃO DE REGÊNCIA DE CLASSE E GRATIFICAÇÃO DE FUNÇÃO ESPECIALIZADA. FORMA DE CÁLCULO. INTERPRETAÇÃO DE LEI ESTADUAL. SÚMULA 280/STF. PRECEDENTES. 1. Questão restrita ao âmbito infraconstitucional, que não enseja apreciação em recurso extraordinário. 2. Agravo regimental desprovido (AI 469.774-AgR/SC, Rel. Min. Ayres Britto). Por fim, quanto à tese subsidiária apresentada pelo recorrente, ressalto que o acórdão recorrido está em consonância com a jurisprudência do Supremo Tribunal Federal, consolidada na Súmula 339 da Corte, no sentido de que não cabe ao Poder Judiciário, que não tem função legislativa, aumentar vencimentos de servidores públicos, sob fundamento de isonomia. Nesse sentido: RE 160.850/MA, Rel. Min. Ilmar Galvão; RE 194.263/SP, Rel. Min. Ellen Gracie; RE 228.522/PI, Rel. Min. Sepúlveda Pertence; RE 342.802-AgR/SP, Rel. Min. Maurício Corrêa; RE 475.915-AgR/CE, Rel. Min. Carlos Britto e RE 558.268-AgR/RJ, Rel. Min. Eros Grau. Isso posto, nego seguimento ao recurso (CPC, art. 557, caput). Publique-se. Brasília, 18 de outubro de 2011. Ministro RICARDO LEWANDOWSKI - Relator - (ARE 654660, Relator(a): Min. RICARDO LEWANDOWSKI, julgado em 18/10/2011, publicado em DJe-203 DIVULG 20/10/2011 PUBLIC 21/10/2011)

Outrossim, revela-se incabível a abertura da via recursal extraordinária na hipótese sub examine , porquanto a este Pretório Excelso somente incumbe a guarda da Constituição Federal, não lhe cabendo o incursionamento em questões relativas a leis locais. Nesse sentido, colacionam-se os seguintes julgados:

ADMINISTRATIVO. FUNCIONÁRIOS DA ATIVA E

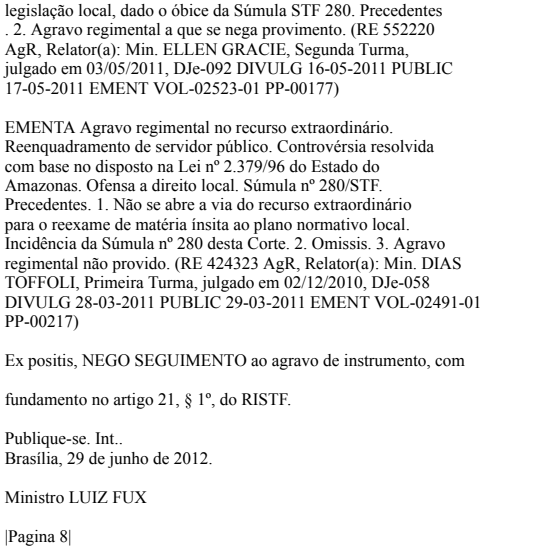
|Pagina 7|

ARE 694.870 / RN

APOSENTADOS. REAJUSTE. LEGISLAÇÃO LOCAL. SÚMULA STF 280. 1. Inadmissível recurso extraordinário no qual, a pretexto de ofensa à CF, pretende-se o reexame de

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. Source: the authors.

Figure 29 – A decision of the process ARE 694870, part 8.



legislação local, dado o óbice da Súmula STF 280. Precedentes . 2. Agravo regimental a que se nega provimento. (RE 552220 AgR, Relator(a): Min. ELLEN GRACIE, Segunda Turma, julgado em 03/05/2011, DJe-092 DIVULG 16-05-2011 PUBLIC 17-05-2011 EMENT VOL-02523-01 PP-00177)

EMENTA Agravo regimental no recurso extraordinário. Reenquadramento de servidor público. Controvérsia resolvida com base no disposto na Lei nº 2.379/96 do Estado do Amazonas. Ofensa a direito local. Súmula nº 280/STF. Precedentes. 1. Não se abre a via do recurso extraordinário para o reexame de matéria ínsita ao plano normativo local. Incidência da Súmula nº 280 desta Corte. 2. Omissis. 3. Agravo regimental não provido. (RE 424323 AgR, Relator(a): Min. DIAS TOFFOLI, Primeira Turma, julgado em 02/12/2010, DJe-058 DIVULG 28-03-2011 PUBLIC 29-03-2011 EMENT VOL-02491-01 PP-00217)

Ex positis, NEGO SEGUIMENTO ao agravo de instrumento, com

fundamento no artigo 21, § 1º, do RISTF.

Publique-se. Int..
Brasília, 29 de junho de 2012.

Ministro LUIZ FUX

|Pagina 8|

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. Source: the authors.

C.2 ARE 637053

Figure 30 – A decision of the process ARE 637053, part 1.

DECISÃO: Trata-se de agravo de decisão que inadmitiu recurso extraordinário (art. 102, III, a , da Constituição) interposto contra acordão cuja a ementa transcrevo (fls. 131):

"EMENTA: Apelação Cível. Ação de cobrança de honorários devidos a Defensor Dativo que atuou em defesa de parte hipossuficiente. Valor que deve ser suportado pelo Estado. Art. 272 da Constituição Estadual e art. 10 da Lei Estadual nº 1.166/99. Preliminar. Ausência de interesse processual. Exaurimento da via administrativa. Desnecessidade. Afastamento. Direito reconhecido. Procedência do pedido. Honorários advocatícios. Inobservância ao art. 20, § 4º, CPC. Provimento do primeiro recurso e desprovimento do segundo recurso."

Alega a parte ora recorrente vulneração do artigo 7º, IV, da Constituição federal.

O recurso não merece provimento.

Sobre o arbitramento do valor da compensação em múltiplo de salários mínimos, o Supremo Tribunal Federal já se pronunciou definitivamente, no julgamento da ADI 1.425 (DJ 26.03.1999), entendendo que o art. 7º, IV, da Carta Magna quis "evitar que interesses estranhos aos versados na norma constitucional venham a ter influência na fixação do valor do mínimo a ser observado ".

Nesse sentido, vários julgados desta Corte têm proibido a utilização do salário mínimo como fator de atualização da indenização (RE 205.455, DJ 06.04.2001; RE 225.488, DJ 16.06.2000; RE 140.940, DJ 15.09.1995, v.g.).

Contudo, não há problema quando a condenação, apesar de fixada em múltiplo de salários mínimos, tem apenas a intenção de expressar o valor inicial da indenização, o qual, se necessário, será atualizado pelos índices oficiais de correção monetária. Confiram-se as seguintes decisões que manifestam esse entendimento:

"EMENTA: CONSTITUCIONAL. INDENIZAÇÃO: SALÁRIO-MÍNIMO. C.F., art. 7º, IV.

I. - Indenização vinculada ao salário-mínimo: impossibilidade. C.F., art. 7º, IV. O que a Constituição veda - art. 7º, IV - é a fixação do quantum da indenização em múltiplo de salários-mínimos. STF, RE 225.488/PR, Moreira Alves; ADI 1.425. A indenização pode ser fixada, entretanto, em salários-mínimos, observado o valor deste na data do julgamento. A partir daí, esse quantum será corrigido por índice oficial.

II. - Provimento parcial do agravo: RE conhecido e provido, em parte." (RE 409.427-AgR, rel. min. Carlos Velloso, Segunda Turma, DJ 02.04.2004)

"EMENTA: Vinculação ao salário mínimo: a vedação do art. 7º, IV, da Constituição, restringe-se à hipótese em que se pretenda fazer das elevações futuras do salário mínimo índice de atualização da indenização fixada; não, qual se deu no acórdão recorrido, se o múltiplo do salário mínimo é utilizado apenas para expressar o valor inicial da condenação, a ser atualizado, se for o caso, conforme os índices oficiais da correção monetária". (RE 389.989, rel. min. Sepúlveda Pertence, Primeira Turma, DJ 05.11.2004).

No mesmo sentido, o AI 493.494-AgR (rel. min. Gilmar Mendes, DJ 04.03.2005) e o AI 510.244-AgR (rel. min. Cezar Peluso, DJ 04.03.2005).

Esse entendimento aplica-se, mutatis mutandis, ao caso ora em exame.

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. Source: the authors.

Figure 31 – A decision of the process ARE 637053, part 2.

Anoto que a sentença julgou parcialmente procedente o pedido nos seguintes termos:

"Condena-se o Réu no pagamento dos valores constantes das certidões, cujo SM será aquele vigente na data do trânsito em julgado da sentença em cujo processo houve a nomeação do Defensor Dativo. Todavia, incidirá CM a partir da data da certidão e, também, juros moratórios de 1% (um por cento) ao mês, estes a contar da data da citação." (fls. 94)

O acórdão recorrido reformou parcialmente a sentença apenas para majorar os honorários advocatícios nesta ação de cobrança, mantendo-a integralmente no restante.

Portanto, a decisão recorrida não divergiu da orientação deste Tribunal.

Do exposto, conheço do agravo e nego seguimento ao recurso.

Publique-se.

Brasília, 06 de março de 2012.

Ministro JOAQUIM BARBOSA

Relator

LIME marked orange sentences as with positive importance for the model's decision, *i.e.*, the potential citation assignment. Blue sentences have negative importance. The intensity of the colors reflects the degree of importance of the sentences. Source: the authors.