UNIVERSIDADE FEDERAL DE GOIÁS (UFG)
INSTITUTO DE INFORMÁTICA (INF)
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EURIPEDES BALSANULFO EVANGELISTA

# Mineração de argumentos em documentos jurídicos em Português

Goiânia
2024

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

# TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES

# E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

**1. Identificação do material bibliográfico**

[ X ] Dissertação [ ] Tese [ ] Outro*:_____________

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

**Exemplos**: Estudo de caso ou Revisão sistemática ou outros formatos.

**2. Nome completo do autor**

Euripedes Balsanulfo Evangelista

**3. Título do trabalho**

Mineração de argumentos em documentos jurídicos em Português

**4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)**

Concorda com a liberação total do documento [ X ] SIM [ ] NÃO[1]

**[1]** Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
**a)** consulta ao(à) autor(a) e ao(à) orientador(a);
**b)** novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

**Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.**

Documento assinado eletronicamente por **Euripedes Balsanulfo Evangelista**, **Discente**, em 24/02/2025, às 17:19, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020 .

Documento assinado eletronicamente por **Nadia Felix Felipe Da Silva**, **Professor do Magistério Superior**, em 24/02/2025, às 20:17, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5133680** e o código CRC **613B2E6D**.

**Referência:** Processo nº 23070.051937/2024-09 SEI nº 5133680

EURIPEDES BALSANULFO EVANGELISTA

# Mineração de argumentos em documentos jurídicos em Português

Dissertação apresentada ao Programa de Pós–Graduação em Ciência da Computação do INSTITUTO DE INFORMÁTICA (INF) da UNIVERSIDADE FEDERAL DE GOIÁS (UFG), como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

**Área de concentração:** Ciência da Computação.

**Linha de pesquisa:** Sistemas Inteligentes e Aplicações.

**Orientadora:** Profa. Dra. Nádia Félix Felipe da Silva

Goiânia
2024

Ficha de identificação da obra elaborada pelo autor, através do
Programa de Geração Automática do Sistema de Bibliotecas da UFG.

EVANGELISTA, EURÍPEDES BALSANULFO
Mineração de argumentos em documentos jurídicos em Português [manuscrito] / EURÍPEDES BALSANULFO EVANGELISTA. - 2024.
XLVII, 47 f.: il.

Orientador: Profa. Dra. Nádia Félix Felipe da Silva; co-orientador Pedro Vitor Quinta de Castro .
Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2024.
Bibliografia. Apêndice.

1. Mineração de argumentos. 2. Documentos da Justiça Trabalhista Brasileira. 3. Modelo de linguagem baseado em Transformers. I. Felipe da Silva, Nádia Félix , orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL

DE GOIÁS INSTITUTO DE

INFORMÁTICA

**ATA DE DEFESA DE DISSERTAÇÃO**

Ata nº **44** da sessão de Defesa de Dissertação de **Euripedes Balsanulfo Evangelista**, que confere o título de Mestre em **Ciência da Computação**, na área de concentração em **Ciência da Computação**.

Aos dois dias do mês de dezembro de dois mil e vinte e quatro, a partir das catorze horas, via webconferência, realizou-se a sessão pública de Defesa de Dissertação intitulada "**Mineração de argumentos em documentos jurídicos em Português**". Os trabalhos foram instalados pela Orientadora, Professora Doutora Nádia Félix Felipe da Silva (INF/UFG) com a participação dos demais membros da Banca Examinadora: Professor Doutor Douglas Farias Cordeiro (FIC/UFG), membro titular externo; Professora Doutora Fabíola Souza Fernandes Pereira (FACOM/UFU), membra titular externa. A realização da banca ocorreu por meio de videoconferência. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pela Professora Doutora Nádia Félix Felipe da Silva, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos dois dias do mês de dezembro de dois mil e vinte e quatro.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Fabíola Souza Fernandes Pereira**, **Usuário Externo**, em 31/01/2025, às 12:00, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Euripedes Balsanulfo Evangelista**, **Discente**, em 21/02/2025, às 17:46, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Nadia Felix Felipe Da Silva**, **Professor do Magistério Superior**, em 24/02/2025, às 20:17, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **5133665** e o código CRC **996534F7**.

**Euripedes Balsanulfo Evangelista**

Graduou em Engenharia da Computação pela Pontifícia Universidade Católica de Goiás, adquirindo sólida formação em tecnologia e sistemas computacionais. Possui mais de 15 anos de experiência dedicados ao desenvolvimento e engenharia de software, trabalhando em projetos complexos e inovadores para diversas empresas. Durante o mestrado participou ativamente de projetos de Inteligência Artificial no Centro de Excelência em Inteligência Artificial da Universidade Federal de Goiás (CEIA/UFG), contribuindo para avanços significativos na aplicação e pesquisa em IA.

À minha querida orientadora, Nádia Félix, cuja orientação firme e sabedoria foram fundamentais para o desenvolvimento deste trabalho. Sua dedicação e apoio foram imprescindíveis para alcançar este marco em minha jornada acadêmica.

Ao meu estimado coorientador, Pedro Vitor, pelo seu suporte constante e valiosas contribuições que enriqueceram significativamente esta pesquisa. Sua experiência e orientação foram essenciais para o meu avanço na presente pesquisa.

À minha amada família, Jaqueline e Enzo, pelo amor incondicional, compreensão e apoio ao longo de todos esses anos. Vocês foram minha força e inspiração, tornando possível cada conquista.

# Agradecimentos

Agradeço à empresa Datalawyer pelo generoso apoio concedido durante o desenvolvimento deste trabalho de pesquisa. Seus recursos foram fundamentais para enriquecer e viabilizar este estudo. A colaboração de vocês foi essencial para o sucesso deste projeto acadêmico.

# Resumo

Evangelista, Euripedes. **Mineração de argumentos em documentos jurídicos em Português**. Goiânia, 2024. 53p. Dissertação de Mestrado. PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, INSTITUTO DE INFORMÁTICA (INF), UNIVERSIDADE FEDERAL DE GOIÁS (UFG).

Este trabalho apresenta uma abordagem de mineração de argumentos aplicada a documentos da Justiça do Trabalho brasileira. Embora a mineração de argumentos em documentos jurídicos seja um tema de estudo há mais de uma década, apenas um trabalho foi encontrado que aplica especificamente esse estudo ao português brasileiro no domínio jurídico. Neste artigo, exploramos detalhadamente todas as etapas necessárias para alcançar o objetivo da tarefa de mineração de argumentos. Assim, nossa abordagem consiste no uso de um Modelo de Linguagem baseado em Transformers treinado em um corpus específico da Justiça do Trabalho brasileira, e relatamos um F1-score de 88,86% na tarefa de classificação. A proposta superou o BERTimbau em 1,88% e o Deberta em 3,39%.

**Palavras–chave**

Mineração de argumentos, Documentos da Justiça Trabalhista Brasileira, Modelo de linguagem baseado em Transformers.

# Abstract

Evangelista, Euripedes. **Argument Mining in Brazilian Labor Justice**. Goiânia, 2024. 53p. MSc. Dissertation. PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, INSTITUTO DE INFORMÁTICA (INF), UNIVERSIDADE FEDERAL DE GOIÁS (UFG).

This work presents an argument mining approach applied to Brazilian labor court documents. Although the mining of arguments in legal documents has been a subject of study for over a decade, only one work has been found that specifically applies this study to Brazilian Portuguese in the legal domain. In this dissertation, we thoroughly explore all the necessary steps to achieve the objective of the argument mining task. Thus, our approach consists of use a Transformers-based Language Model trained on a specific domain corpus of Brazilian labor justice and we report an F1-score of 88.86% on the classification task. The proposal outperformed BERTimbau by 1.88% and Deberta by 3.39%.

**Keywords**

Argument Mining, Brazilian Labor Court Documents, Transformers-based Language Model.

# Contents

# List of Figures

# List of Tables

CHAPTER 1

# Introduction

The Brazilian Labor Court reported a volume of 2,943,886 new cases, according to the analytical report published in 2,022 by the National Council of Justice (CNJ)[1]. In addition, this same sphere of justice had 5,186,775 pending cases in the same year. Attached to these lawsuits are tens of millions of textual documents that make up the nature of a legal document, such as initial petitions, sentences, and hearing minutes.

In the Labor Court, the lawyer supports the requests made using arguments that strengthen and justify the fulfillment of the requests made by him. Therefore, it is very common for these professionals to visit similar cases that were previously judged to understand the scenarios for granting and rejecting requests [Nascimento e Fernandes 2023]. From another point of view, when a judge wants to sentence a decision, he may want to understand how the judges have judged a certain type of case, and they also need to visit similar cases. Textual documents derived from these scenarios are rich in arguments as they need to justify decisions between legal disputes.

Argument is an element of linguistics characterized by supporting an idea or information that you want to convey. One of the pioneers in building a representative model of arguments was [Hardin 1959]. His model is a tool for analysis and construction of arguments and is composed of six parts: claim, grounds, warrant, backing, rebuttal, and degree. Based on Toulmin's model [Freeman 2011] in his work, proposes the representation of arguments as a product using diagrams that include elements that are represented by premises and conclusions.

Argument Mining (AM) is the task of automatically extracting arguments from texts, and according to Walton [Walton 2009], it consists of three standard tasks: *(1)* information detection: Identify the argumentative sentences in the text. This step is an approach to separating texts with arguments from those that are not argumentative; *(2)* Classification of the argument components: Once the arguments have been identified, this task aims to classify the elements of the sentence. Walton, for example, classifies his works into a set of premises and conclusions; And *(3)* Relationship prediction between

[1] https://www.cnj.jus.br/justica-em-numeros/

Arguments: After classifying the elements of argumentative sentences, described in the previous topic, this task aims to make inferences about the relationships that the premises have on the conclusions, whether they support or attack, for example. This work will focus only on task 1, and we modeled the argument task (information detection) as multiclass classification problem, which one of the classes is the argumentation.

The research reported in this paper has the following contributions: **(i)** Argument mining is not a recent area of knowledge of Natural Language Processing, although it has been studied in his first work in mid-2008 [Mochales e Moens 2008], and until the date of this paper there is only one work focused on this topic addressing the context of legal documents for the Portuguese language (3. Therefore, this document aims to propose a study on this subject and contribute to the evolution of this area. We adapted the Transformer-based Model [Vaswani et al. 2017] for identify arguments in Brazilian Labor Justice texts modeled this task as classification of sentences of multiple classes where one of them is of the argumentative type. (ii) We present a comparative review of related works for argument mining across multiple corpus types context; ***(iii)*** We describe a protocol of annotation of a legal argumentative corpus.

This work is organized as follows. The second section provides an overview of related work. The section three introduces the methodology and the experiments. Section four presents the experimental evaluation. The following section comprises the discussions and concludes the work, summarizing the findings, advantages, limitations, contributions and research opportunities.

## 1.1 Justification

### 1.1.1 Relevance to the Portuguese Language

Argument mining, although not a particularly recent field of study, with its first appearances in works published in 2008 by [Mochales e Moens 2008], has no works addressing documents in Brazilian Portuguese. Portugal, through the University of Porto, initiated some projects on argument mining, such as [Rocha 2016], where the authors propose a framework called ArgMine [2], which provides a comprehensive ecosystem for conducting argument mining tasks. The framework aims to automatically select text zones containing argumentative content. The corpus used to train this model was obtained from Portuguese news articles. Furthermore, [Rocha et al. 2018] proposes a cross-lingual approach from English to Portuguese capable of identifying relationships between argumentative text segments. Although there have been works found

[2] https://web.fe.up.pt/~argmine/

for European Portuguese, Brazilian Portuguese still lacks efforts in this area. In 2021, [Aragy, Fernandes e Caceres 2021] et al, presented the paper "Rhetorical Role Identification for Portuguese Legal Documents" that focuses on automating the identification of rhetorical roles in legal documents written in Portuguese.

### 1.1.2 Application in the Domain of Brazilian Labor Justice

According to a report issued by the CNJ in 2022, 2023 and 2024, the Brazilian Labor Justice system received more than 3 million new cases each year. Additionally, this branch of the judiciary had the same proportion of pending cases, as shown in Figure 1.1:


Figure 1.1: New cases by branch of justice. Data obtained from the ***Justice in Numbers*** report.

The presence of such a large volume of active cases feeds the courts' databases with hundreds of millions of legal documents containing texts to be explored. Unfortunately, so far, no research work has been found that addresses argument mining in this vast ecosystem of documents, leaving the Brazilian labor justice system without the benefits that could be gained.

In labor law, attorneys substantiate their requests using arguments that strengthen and justify the fulfillment of their demands. Consequently, it is common for these professionals to review similar cases previously adjudicated to understand scenarios of approval and denial of requests. Alternatively, when a judge seeks to issue a ruling, they may wish to grasp how fellow magistrates have adjudicated a particular type of case. Nonetheless, such analysis complicates the task due to the extensive nature of cases, which encompass substantial textual content, including arguments that necessitate thorough scrutiny for comprehensive comprehension.

This motivates this line of study, as it not only proposes pertinent research aimed at extracting argumentative segments from legal texts but also envisages a technological solution to enhance the work of legal professionals, thereby elevating the quality of service provided.

## 1.2 Objectives

The aim of this work is to propose an argument extraction model based on the *Transformers* architecture [Vaswani et al. 2017] within the scope of the Portuguese language in documents pertaining to Brazilian labor law.

### 1.2.1 General Objective

The general objective of this work is to understand how argument mining can be explored within the context of Brazilian labor law.

### 1.2.2 Specific Objectives

- Propose an annotation protocol for a corpus based on sentences/phrases that can be classified as argumentative components. This will enable the construction of a specific corpus in Portuguese aimed at the domain of Brazilian labor law;
- Use the corpus obtained in the previous phase to train models based on the *Transformers* architecture to law sentences classification;
- Evaluate model effectiveness with metrics such as F1-score, precision, and recall;
- Contribute to advancing the literature on Natural Language Processing for Portuguese by reporting the results in scientific articles.

## 1.3 Research Hypotheses

Over the past decade, various techniques have been explored with the aim of performing argument mining tasks, as illustrated in previous works and revisited in the Related Works chapter of this dissertation.

**Hypothesis 1:** Is it possible to identify arguments in Portuguese legal documents through a sentence classification model using the *Transformers* architecture?

## 1.4 Document Organization

The structure of this document is designed to systematically address the core elements of our research, from foundational theory to evaluation and interpretation of results.

The Section 2 presents the theoretical underpinnings of the research, detailing the primary concepts of labor justice and technical aspects relevant to argument mining in the Brazilian legal domain. Here, we explore key concepts and previous methodologies that inform our approach, providing a robust context for the study. The Section 4 describes the dataset produced and utilized, encompassing the data source origin, selection of documents and annotation processes involved in constructing a representative corpus. The Section 3 we review pertinent literature, focusing on the most recent advancements in natural language processing (NLP) and argument mining. This review synthesizes prior research, identifying gaps and highlighting where our work contributes to or extends existing findings. The Section 5 details the methodological approach employed in the pipeline of the experiment. It includes obtaining documents, selection of documents to annotate, preprocessing data, model's training and analysis of results. The Section 6 presents the results findings derived from our experiments, followed by an in-depth discussion of their implications. We analyze the results in relation to our research questions, assessing both the strengths and limitations of our model training. Where relevant, we also explore potential directions for improvement and future research. The Section 7 synthesizes the study's key findings, emphasizing the contributions made to the field of argument mining using NLP. We summarize the limitations identified and offer insights into advancements in applying large language models to the legal domain.

CHAPTER 2

# Theoretical foundation

This chapter aims to provide a basic understanding of the business area being explored in this work. Therefore, the field of Law, as well as concepts of labor justice will be explained in order to provide a better understanding of what is being covered in this dissertation.

## 2.1 Argument

Argument is an element of linguistics characterized by supporting an idea or information that you want to convey. One of the pioneers in building a representative model for arguments was [Hardin 1959]. His model is a tool for analysis and construction of arguments and is composed of six parts: claim, grounds, warrant, backing, rebuttal, and degree. Based on Toulmin's model [Freeman 2011] in his work, proposes the representation of arguments as a product using diagrams that include elements that are represented by premises and conclusions.

## 2.2 Brazilian Labor Law and Justice

Reale [Reale 2001] defines *Law* as the set of obligatory rules to which a society is subjected, imposing limits on its members and ensuring social interaction among them. The first way in which the science of Law was divided was between *Public* and *Private* [Reale 2001]. Within Public Law, there are two divisions that are the focus of this work: *Procedural Law* and *Labor Law*.

Procedural Law portrays the State as a service provider to society, as it must act as a mediator in conflicts that occur among its members [Reale 2001]. Therefore, the objective of Procedural Law is to clearly define how the State should carry out its role through a system of rules and principles to be observed and complied with. This system is governed by a set of procedures known as the ***process***. The remainder of this chapter will describe the elements and procedures of a process within the context of a Labor Law action.

According to Reale [Reale 2001], Labor Law is yet another manifestation of Public Law, whose focus is to regulate relations between employers and employees. In Brazil, the State manifests itself in Labor Law in the form of ***Labour Justice***, which was organized in 1943 through *Consolidation of Labor Laws* (CLT) [Basile 2012]. Basile [Basile 2012] thus describes the division of jurisdiction[1] in the Labor Court into *instances* (or textitdegrees):

**First Instance:** Labour *Judges* who act in ***Labor Courts*** (LC), with each court being composed of a titular judge and a substitute judge (the latter, if the court's budget allows).

**Second Instance:** ***Regional Labor Courts*** (RLC), distributed in 24 *regions* across the national territory, each composed of at least 7 *judges*.

**Extraordinary Instance:** ***Superior Labor Court***, composed of 27 *ministers*.

Basile [Basile 2012] defines the elements of a labor action as follows:

- *Parties* are the subjects involved in the conflict: the one who felt harmed and initiated the action, against whom the harm is claimed, who is the target of the action. These two subjects are also referred to as the *active party* and the *passive party* of the process, respectively. In the context of labor claims (another term for ordinary labor action), these roles are also commonly referred to as *claimant* and *defendant*.
- *Claim*, also called the *object* of the action. The claim in an action represents a right of the active party of the process, granted to them by labor legislation, which in some way would have been denied by the passive party. It is possible to make several claims in a single process, as established in article 292 of the *Civil Procedure Code* (CPC).
- *Cause of action*, which must be described in the form of facts and ***legal grounds*** that justify and underpin the claim. This grounding can be made argumentatively, pointing out and describing wrongdoings caused by the passive party of the process, as well as through citation of ***legal provisions*** (such as legislation and jurisprudence) that constitute the relevant Labor Law rules in the action.

There are still other individuals involved in a labor lawsuit, performing different roles besides the ***judge*** and the ***parties***. The parties are represented by their ***lawyers***, who possess knowledge about the process proceedings to guide their clients in the course of

[1]Representation of the State through a judge in a process [Basile 2012].

the action. ***Witnesses*** contribute to substantiating the parties' claims, whether to affirm or refute any allegation made by them. ***Experts*** carry out the task of providing an assertive and well-founded validation of evidence or allegations, providing support for the judge to make their decision.

When initiating a labor lawsuit, the claiming party must present, through the lawyer representing them, a document called the ***initial petition***. The petition must enumerate the claims and the grounds for each one, also indicating the respondent party that allegedly caused them harm. According to article 319 of the Civil Procedure Code (CPC) [Código de Processo Civil 2015], the initial petition must define the value of each claim sought, whether this value is precise or estimated, according to the scenario. The sum of all the requested values is called the ***amount in dispute***. When making a decision, the judge either grants or denies each claim made by the claimant. Similarly, even for claims that have been granted, the amount awarded for each may not correspond to the requested value. The sum of the amounts granted by the judge for each claim is called the ***award amount***. At any point in the process, the parties can settle and reach an agreement, which must be approved by the judge. The amount agreed upon by the parties, to be paid by the claimant to the respondent, is called the ***settlement amount***. There is also the ***court costs***, which correspond to the sum of expenses resulting from the process proceedings, payable to the Judiciary, for the provision of public service. Labor Justice defines the calculation of the cost value in article 789 of the Labor Code (CLT) [Basile 2012]

## 2.3 Technical aspects

### 2.3.1 Transformer Architecture

Transformer [Vaswani et al. 2017] is a promising deep neural network model that has been widely used in several areas of machine learning, such as: Natural language processing (NLP), Computer Vision and speech processing. This model, previously proposed as a *sequence-to-sequence* [Sutskever, Vinyals e Le 2014] model with the aim of solving text translation problems, soon proved to be very efficient in solving other tasks with the pre-trained module proposal. [Qiu et al. 2020], reaching state-of-the-art performance in your application.

The Vanilla Transformer [Vaswani et al. 2017] consists of a *sequence-to-sequence* model that is divided into two blocks of L identical layers. As illustrated by figure 2.1, the *encoder* block is basically formed by a *Multi-head attention* and *Position-wise-feed forward (FFN)* module. As it is a deep learning model, a normalization layer is added for each [Ba, Kiros e Hinton 2016] module. Compared to the *encoder* block, *decoder* has an additional layer of attention that is adapted so that attention is not activated for subsequent positions, since the main function of this block is, given one word, predict what the next one will be.

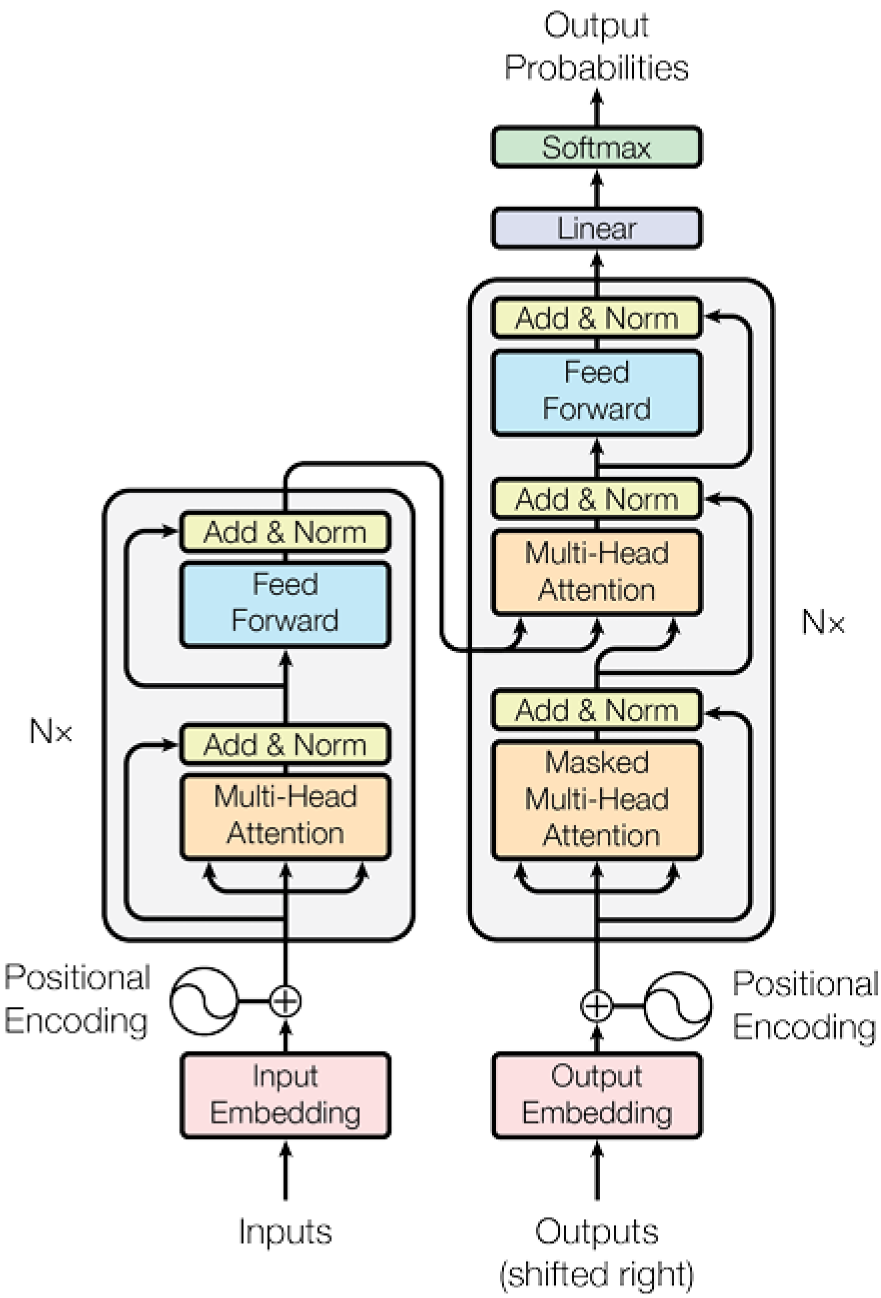


Figure 2.1: Vanilla Transformer's Architecture [Vaswani et al. 2017]

### 2.3.2 Attention modules

The Transformer architecture adopts a Query-Key-Value (QKV) attention mechanism strategy where each component is a representative matrix.

$$\mathbf{Q} \in \mathbb{R}^{NxD_k} \tag{2-1}$$

$$\mathbf{K} \in \mathbb{R}^{MxD_k} \tag{2-2}$$

$$\mathbf{V} \in \mathbb{R}^{NxD_v} \tag{2-3}$$

The dot product used in the attention mechanism used by Transformer is given by:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V = AV \tag{2-4}$$

where N and M represent the size of the queries and keys (or values); $D_k$ and $D_v$ indicate the dimensions of the keys (or queries) and values.

The block $A = softmax(\frac{QK^T}{\sqrt{d_k}})$ is commonly called the attention matrix; The softmax algorithm is applied sequentially. The scalar product of queries and keys is divided by $\sqrt{d_k}$ in order to alleviate the gradient vanishing problem that exists in the softmax function.

Instead of applying just a simple attention function, Transformer uses so-called multi-head attention. This method projects the original queries, keys and values with ($D_m$) dimensions into $D_k$, $D_k$ and $D_v$ dimensions respectively with H different combinations of learning projections. For each projection of queries, keys and values, the output is computed with the attention function 2-4. After this step, the model concatenates the results of the outputs and then projects them back to the initial representation ($D_m$)

$$MultiHead(Q, K, V) = Concat(head_1, ..., head_h)W^O \tag{2-5}$$

where

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{2-6}$$

In Transformers, there are three basic forms of attention taking into account the input source of queries and key-value pairs:

- Self-attention. Used in the encoder block, the sets Q = K = V = X in Eq (2-5), where X is the output of the previous layer.
- Masked Self-attention: Used in the Transformer decoder block it is essentially a way to prevent the model from looking at information we don't want it to look at.
- Cross-attention. Queries are projected from the output of the last decoder layer, while keys and values are projected using the encoder outputs.

### 2.3.3 Residual connection and normalization

Transformer implements a deep neural network and therefore employs a resitual connection [He et al. 2020] between the blocks of its architecture, followed by a Normalization layer [Ba, Kiros e Hinton 2016]. This way, each Transformer encoder block could be written as:

$$H' = LayerNorm(SelfAttention(X) + X) \tag{2-7}$$

$$H = LayerNorm(FFN(H') + H') \tag{2-8}$$

Where SelfAttention indicates the attention module and LayerNorm indicates a layer with normalization operations.

### 2.3.4 Transformer usage

The Transformer architecture allows the use of its components in a decoupled way:

- *Encoder-Decoder*. The Transformer architecture in its entirety is used. This is a typical example of a *sequence-to-sequence* architecture being used. Example: automatic text translation.
- Only *Encoder*. Only the encoder is employed and the output with the encoded tokens is used as the input representation. It is generally used in Natural Language Understanding (NLU) tasks. Example: text classification.
- Only *Decoder*. Only the decoder is used. This way, the *encoder-decoder cross-attention* module is also removed. It is generally a use for generalist models. Example: language model.

### 2.3.5 RoBERTa: A Robustly Optimized BERT Pretraining Approach

The BERT model [Devlin et al. 2019] presented in 2019 brought a turnaround in the results of existing benchmarks to date. However, this recently launched model had limitations highlighted by the authors themselves. Although the authors of the original BERT work presented a good work in exploring system ablation and different forms of optimization, a wide range of possible ideas were not explored in the original paper. Such factors motivated a meticulous study of this model in order to adjust hyperparameters and size of the training set to propose a version with higher performance. It was exactly what RoBERTa proposed [Liu et al. 2019].

Below is a brief summary of the main differences between the BERT and RoBERTa models.

The first difference concerns the issue of static masking versus dynamic masking. In the original BERT model, four copies of its original dataset were created, each with their respective masks. These four copies were used repeatedly during training seasons. The RoBERTa team had the intuition that it would be useful to introduce some type of diversification into this training process. Therefore, they went to another extreme, dynamic masks. Each example was presented to the model in a different form of masking, based on some randomness function. Another important difference is the way in which the examples were presented to the model. BERT presented two segments of concatenated documents. This was crucial for the next sentence prediction task. While for Roberta, we only have sequences of sentences, that is, pairs, which can even include document links. While Bert had Next Sentence Prediction as one of his main tasks, Roberta removed that part of the objective. This made the presentation of examples simpler and also facilitated the modeling of the objective. Therefore, Roberta proposed to use only modeling based on masked language modeling. There was also a change in the size of training batches. For BERT, batches of 256 examples were used while in Roberta the training was done using batches of 2000 examples. Another important difference was in the form of tokenization used in the models. Bert used word-piece tokenization while Roberta used Character-level byte-pair encoding, which made it possible to form many more word fragments. Another difference is in the way the models were trained. Bert was pre-trained based on the BooksCorpus and English Wikipedia corpus. Roberta used the BooksCorpus, CC-news, OpenWebText and Stories corpus as pre-training datasets. Substantial growth in the volume of training data. The number of training steps also showed differences. Bert with 1M training steps and Roberta with 500k steps, which sounds like a lower volume but actually presents an absurdly higher training volume, due to the fact that the training batch sizes are larger for Roberta. Finally, the authors of the original Bert had the intuition that they would optimize the training process by presenting only smaller input sequences first. Roberta's team removed this idea. They trained the model with the idea of always presenting the sequence completely throughout the training cycle.

Below is some evidence that supported the decision-making listed above.

The Table 2.1 displays published results in [Devlin et al. 2019]'s $BERT_{BASE}$ with static masking compared to dynamic masking work proposed by the Roberta team. The results were very similar and, if analyzed in general, the results with dynamic masking were better.

The table 2.2 shows that numerically the results using the DOC-SENTENCES strategy performed better. However, this approach is limited to the document limit and therefore does not make it possible to work with a fixed size of sequences. For this

| Masking | SQuAD 2.0 | MNLI-m | SST-2 |
|---|---|---|---|
| reference | 76.3 | 84.3 | 92.8 |
| Our reimplementation: | | | |
| static | 78.3 | 84.3 | 92.5 |
| dynamic | 78.7 | 84.0 | 92.9 |

Table 2.1: Comparison between static masking in $BERT_{BASE}$ and the dynamic masking adopted by Roberta in [Liu et al. 2019]. Results were reported in F1 for SQuAD and accuracy for MNLI-m and SST-2. The reported results are the averages of 5 random seedings. Benchmark results are at [You et al. 2020]

| Model | SQuAD 1.1/2.0 | MNLI-m | SST-2 | RACE |
|---|---|---|---|---|
| *Our reimplementation (with NSP loss):* | | | | |
| SEGMENT-PAIR | 90.4/78.7 | 84.0 | 92.9 | 64.2 |
| SENTENCE-PAIR | 88.7/76.2 | 82.9 | 92.1 | 63.0 |
| *Our reimplementation (without NSP loss):* | | | | |
| FULL-SENTENCES | 90.4/79.1 | 84.7 | 92.5 | 64.8 |
| DOC-SENTENCES | 90.6/79.7 | 84.7 | 92.7 | 65.6 |
| $BERT_{BASE}$ | 88.5/76.3 | 84.3 | 92.8 | 64.3 |
| $XLNet_{BASE}$ (K=7) | -/81.3 | 85.8 | 92.7 | 66.1 |
| $XLNet_{BASE}$ (K=6) | -/81.0 | 85.6 | 93.4 | 66.7 |

Table 2.2: Result set results for the base models trained with the BOOKCORPUS and WIKIPEDIA datasets. Both were trained in 1M steps with a batch size of 256 sequences. Results were reported in F1 for SQuAD and accuracy for MNLI-m, SST-2 and RACE. The reported results are the averages of 5 random seedings. Reference results are at [You et al. 2020]

reason, even with slightly lower results, the Roberta team preferred to follow the FULL-SENTENCES approach due to the ease of creating lots of batches with exactly the same size, since here the limits of the document would not need to be respected, which leads to all kinds of gains when you want to optimize a model of this size.

| bsz | steps | lr | ppl | MNLI-m | SST-2 |
|---|---|---|---|---|---|
| 256 | 1M | 1e-4 | 3.99 | 84.7 | 92.7 |
| 2K | 125K | 7e-4 | **3.68** | **85.2** | **92.9** |
| 8K | 31K | 1e-3 | 3.77 | 84.6 | 92.8 |

Table 2.3: Perplexity on extended training data (ppl) and accuracy results for basic models (Roberta) trained on the BOOKCORPUS and WIKIPEDIA corpus with varying batch sizes (bsz). The Roberta team adjusted the learning rate (lr) for each training set. The models go through the data the same number of times (epochs) and have the same computational cost.

In the table 2.3 we can observe the accuracy of the BERT model, trained with batch size 256, on the MNLI-m and SST-2 benchmarks. Next, we have a sampling of Roberta's results being run with batch sizes 2k and 8k. In the first option, Roberta managed to surpass the BERT results, which did not happen as the batch size increased. Clear evidence of why batch size 2k was selected.

| Model | data | bsz | steps | SQuAD (v1.1/2.0) | MNLI-m | SST-2 |
|---|---|---|---|---|---|---|
| RoBERTa | | | | | | |
| with BOOKS + WIKI | 16GB | 8K | 100K | 93.6/87.3 | 89.0 | 95.3 |
| additional data (§3.2) | 160GB | 8K | 100K | 94.0/87.7 | 89.3 | 95.6 |
| pretrain longer | 160GB | 8K | 300K | 94.4/88.7 | 90.0 | 96.1 |
| pretrain even longer | 160GB | 8K | 500K | **94.6/89.4** | **90.2** | **96.4** |
| $\text{BERT}_{\text{LARGE}}$ | | | | | | |
| with BOOKS + WIKI | 13GB | 256 | 1M | 90.9/81.8 | 86.6 | 93.7 |
| $\text{XLNet}_{\text{LARGE}}$ | | | | | | |
| with BOOKS + WIKI | 13GB | 256 | 1M | 94.0/87.8 | 88.8 | 94.4 |
| + additional data | 126GB | 2K | 500K | 94.5/88.8 | 89.8 | 95.6 |

Table 2.4: Results of the development of Roberta's pre-training for a volume of data that went from 16GB to 160GB of text, in addition to undergoing training that went from 100k steps to reaching 500k steps. Each line accumulates improvements in relation to the line above. RoBERTa and $BERT_{LARGE}$. Roberta's results can be found at [Liu et al. 2019] and the results of $BERT_{LARGE}$ and $XLNet_{LARGE}$ are in [Devlin et al. 2019] and [You et al. 2020], respectively.

In the table 2.4 the lesson is "more data is better". Although BERT's training was done in 1M steps compared to Roberta's 500k, this second one presents a much larger amount of training due to the 12x larger data volume and combined with the size of the batch used, which is also much larger. This shows better results compared to the original model proposed by [Devlin et al. 2019].

### 2.3.6 RoBERTaLexPT: A Legal RoBERTa Model pretrained with deduplication for Portuguese

[Garcia et al. 2024] developed a RoBERTa-based language model fine-tuned for legal documents in Portuguese. The model is pre-trained using a deduplicated corpus to improve the performance and relevance of the results for legal tasks. It particularly targets the Brazilian legal system, making it a specialized tool for processing and analyzing legal documents within this context.

This work contributes to Natural Language Processing (NLP) in the legal domain, aiming to improve the quality and precision of tasks like text classification, named entity recognition (NER), and other legal document-related processes. The study also highlights the comparative performance of this model against existing models like BERTimbau, demonstrating enhanced results in legal document processing tasks. Considering the context of applying Transformer models to legal document analysis in Portuguese, this is the reason why this model was chosen to be explored in the present work.

CHAPTER 3

# Related works

The study of argumentation has a long tradition, inside and outside the legal field. Many fields have shown an interest in argumentation, e.g. philosophy, logic, psychology or, more recently, approaches with artificial intelligence. We focus on argumentation inside the legal field.

Since [Mochales e Moens 2011] presented their work in 2011, the topic of argument mining already had an approach divided into detecting all arguments in a text and their relationships with each other. They focused on highlighting the challenges and applications of argument mining, offering an initial study on how machine learning and other state-of-the-art techniques at the time could assist in execution. In this study, the authors bring important reflections to the study of argumentation: (i) What is the "correct" abstract structure of argumentation? (ii) Should we represent argumentation as a tree-structure or is it better to use a graph-structure? (iii) What are the constraints that characterize this structure? (iii) What are the elementary units of argumentation? And of an individual argument? (iv) Can the units of argumentation and/or arguments be determined automatically? (v) Can argumentation structures be determined automatically? If so, how?

[Mochales e Moens 2011]'s work used two Corpus as a base. The first, called Araucaria, was constructed from two distinct data sources: one structured in English, collected and analyzed according to a specific methodology as part of a research project at the University of Dundee (UK) and a second part of Araucaria was extracted from structured data sources from magazines, newspapers, parliamentary records, among others. The second Corpus, called ECHR, consists of documents extracted from legal texts from the European Court of Human Rights (ECHR).

[Mochales e Moens 2011] used a set of features combined with classic machine learning techniques such as Support Vector Machines (SVM), Naive Bayes, and Maximum Entropy classifiers to identify arguments from the corpus used in the experiments. They achieved an accuracy of 73% for the Araucaria corpus [Mochales e Moens 2008]. The accuracy increases to 80% when applying the heuristic method on the ECHR corpus. In their first work [Mochales e Moens 2008] they achieved 90% accuracy for this last corpus in a previous version of it. It is worth remembering that the results mentioned above refer to the binary classification of the text between argument/non-argument, not the classification based on the structure of the argument.

Next, [Poudyal et al. 2020] present in their work an annotated corpus of 42 decisions from the European Court of Human Rights (ECHR). The corpus was annotated in terms of three types of clauses heavily used in argument mining: premises, conclusion, and non-argumentative text. With an approach based on bidirectional encoders from Transformers (BERT), they obtained results that the authors themselves classified as promising. Poudyal et al also understood the problem as a binary classification and based on a classification algorithm employing the pre-trained language model RoBERTa, they achieved an F1 score of 0.76 for the positive classes. Observing the corpus, they noticed that the arguments always contained the clauses “As to the law”/“The law”, and they used this to improve the model. As the information of the corpus is structured, they considered that any text outside the limits where the clauses mentioned above were observed would be non-argumentative. The argument detection task obtained in the method proposed by Poudyal et al presented the Precision, Recall and F1 score metrics of 0.697, 0.848 and 0.765, respectively.

[Stab e Gurevych 2014] in his work presents an approach to identifying arguments in persuasive essays. They propose the structure of an argument identified in claims and premises and also explore the identification of relationships between arguments. The identification stage, which is the one we are interested in developing this work, is done using multi-class classification. For this task, an F1-score of 0.726 was obtained. Stab et al. used a set of persuasive essays compiled by themselves for their experiment. This corpus has annotations of clause-level arguments as well as the relationships between them. As for the clause-level annotations, they particularly added the major claims, claims and premises labels. Three annotators were used to produce corpus annotations with a cohesion factor of $\alpha = 0.72$ [Antoine, Villaneau e Lefeuvre 2014] for arguments. In total, the corpus was composed of 90 essays including 1,673 sentences. In their experiment they used several classifiers like Support Vector Machine, Naive Bayes, C4.5 Decision Tree and Random Forest. To preprocess the corpus, they used the Stanford POS-Tagger [Toutanova et al. 2003] and Parsers.

The features used by them were:

a) Lexical features: Definition of n-grams, verbs, adverbs and modals. Verbs and adverbs play an important role in identifying arguments. For example, the presence of certain verbs such as ’believe’, ’think’ or ’agree’ usually signal the presence of arguments in the sentence. At the same time as ’also’, ’often’ or ’really’ emphasize the importance of a premise. Modal verbs such as ’should’ and ’could’ are often used in argumentative speech to signal the degree of certainty when expressing a statement. b) Indicators:

Discourse markers generally indicate the components of an argument. For example, claims are often started with 'therefore', 'thus', or 'consequently', while premises have markers like 'because', 'reason', or 'furthermore'. The authors collected a list of discourse markers from Penn Discourse Treebank 2.0 Annotation Manual [Prasad et al. 2008] and removed markers that do not indicate argumentative discourse.

To run the experiments they randomly divided the data into 80% training set and 20% test set containing the same class distributions and determined the best performance for 10-fold cross-validation on the training set only. Of the classifiers used, the one that performed best was SVM (Support Vector Machine), which presented an F1-score of 0.726 compared to an F1-score of 0.871 when using humans to classify arguments.

[Aragy, Fernandes e Caceres 2021] et al, presented in 2021 the paper "Rhetorical Role Identification for Portuguese Legal Documents" that focuses on automating the identification of rhetorical roles in legal documents written in Portuguese. It builds on existing work in legal Natural Language Processing (NLP), particularly around the structuring and summarization of legal texts. The core aim is to classify different parts of legal documents into rhetorical roles (such as facts, arguments, legal reasoning, etc.) to help with document understanding, summarization, and information retrieval. The approach likely uses machine learning techniques, potentially relying on language models such as BERT or other transformers, to segment legal texts and assign a rhetorical label to each section. This is crucial for simplifying the processing of legal documents, making it easier for legal professionals and AI systems to extract relevant information for tasks like legal research, case summarization, or judgment prediction. By addressing the challenge of processing Portuguese legal texts, this work contributes to the growing body of research that applies NLP to non-English languages in specialized domains like law.

In our approach, we adapted a Transformer based Model which will have the task of classifying sentences in a multi-classes approach to identify arguments in Brazilian Labor Justice texts. One of these classes is specifically **Argumentation** and will make it possible to identify argumentative sentences in legal texts. As observed in the works listed above, the use of a specific corpus from the domain area contributes to more performed results. Unlike the approaches proposed by [Mochales e Moens 2011] and [Stab e Gurevych 2014], we will follow the line of experiments proposed by [Poudyal et al. 2020] using Transformers that have shown themselves to be very promising, reaching the state of the art in several tasks previously performed by traditional classification methods with dependent hand-craft on domain knowledge features.

[Silveira et al. 2023] introduced LegalBert-pt, a specialized pretrained language model tailored to the Brazilian Portuguese legal domain. The authors addressed the challenges posed by the unique linguistic characteristics of legal texts, such as their formal tone, extensive use of technical vocabulary, and frequent inclusion of legal references, which are not adequately captured by general-purpose models.

The LegalBert-pt model was pretrained on a comprehensive and diverse corpus of Brazilian legal documents, ensuring that it captures the nuances of the legal domain. The model demonstrated superior performance in tasks such as Named Entity Recognition (NER) and text classification, significantly outperforming generic language models. These results highlight the importance of domain-specific pretraining for achieving robust performance in specialized contexts.

Furthermore, the authors made the model publicly available through the Hugging Face platform, enabling its use and customization for various applications in the legal domain. The study underscores the potential of domain-adapted models in enhancing the efficacy of Natural Language Processing (NLP) tasks within specialized fields.

CHAPTER 4

# Corpus

A well-constructed corpus alone does not guarantee good training performance for models based on supervised learning, but it can contribute positively to this. Such an assignment requires annotators to have specific knowledge of the textual domain to deal with legal texts, as well as a well-aligned annotation protocol. This subsection describes the proposed protocol for annotating the corpus of arguments, as well as the characteristics of the annotators, the tools used, and a descriptive analysis of resulting phase.

Annotating documents in a specific domain requires knowledge of the vocabulary and characteristics of the text being annotated. Thus, we were able to note the importance of the annotators involved in the projects having legal training. Therefore, the annotators selected for this task are lawyers, so that they can understand the specific legal terms covered in these documents. Furthermore, for tasks such as argument mining, they must be able to recognize the argument itself in legal texts.

Three annotators collaborated in this annotation process. This team was formed by lawyers who work as data analysts. The knowledge gained in annotation was based on past projects to address other NLP tasks. With this background, the more experienced guided and shared knowledge with the more inexperienced. The most experienced annotator has been working with document annotations for four years. The rest have 6 months and 24 months working as note-takers.

The most experienced annotator, responsible for the annotation process, was responsible for collecting the documents that would be part of the corpus from the repository where it is stored and guiding the other annotators on the guidelines that would be followed in order to achieve maximum coherence between the labeling. This also served as a focal point for clarifying doubts when an annotation scenario arose that had not yet been foreseen. As soon as the first documents were being annotated, the person responsible was evaluating the result in order to point out supposed corrections that should be applied to realign the agreement between the annotations. This process continued until the person responsible could trust the cohesion of the notes between the three work fronts.

Defining the limits of an argument is not a trivial task. This is because an argument can start in one sentence and connect with others sentences during the text, thus forming a complex argumentative section. Given this, authors adopt different approaches to defining the limits of arguments in argument mining researches. In this work it was defined that the limits for annotations would be at sentence level as can be observed in other works that deal with legal corpus such as [Savelka et al. 2017] and [Mochales e Moens 2011].

With the aim of producing a more flexible Corpus, which can also be used for other tasks besides specific argument mining, the sentences were annotated with different types of classes where one of them is of the **Argument** type, instead of a binary approach containing only argument/non-argument.

The resulting corpus consists of 49,848 annotated sentences.The labels considered are related in the Table 4.1:

Table 4.1: Number of sentences noted per class

| ***Label*** | ***Quant. annot.*** | ***Tokens/sent*** |
|---|---|---|
| NO_CLASS | 14,230 | 15 |
| **ARGUMENTATION** | **10,338** | 56 |
| STATEMENT | 7,253 | 50 |
| DEFINITION | 6,538 | 40 |
| CLAIM | 4,132 | 42 |
| VERBATIM | 3,311 | 10 |
| DECISION | 2,596 | 39 |
| REFERENCE | 1,103 | 45 |
| QUALIFICATION | 347 | 97 |

The corpus consists of 514 documents. At the beginning of the annotation process, a sample of 300 documents were selected. It was divided into three equal parts, 100 documents for each annotator, who worked in parallel. As mentioned above, filtering the types of documents before selecting the annotation sample is of utmost importance as it ensures greater use of the target type of class we are seeking to identify. In this way, we selected documents such as sentence, initial petition and agreement.

The histogram shown in Figure 4.1 illustrates the distribution of word counts, with the x-axis representing the number of words per text segment and the y-axis representing the frequency of these word counts. The majority of texts contain a relatively low number of words, particularly within the range of 0 to 50 words, as indicated by the high frequency in this area of the histogram. This suggests that the dataset consists predominantly of short texts, which could include concise responses, brief statements, or short paragraphs. Few instances have over 100 words, and texts exceeding 150 to 200 words are rare. This suggests that long texts are uncommon in the dataset, and most entries fall into shorter length categories. This finding may be relevant for Natural Language Processing (NLP) tasks, as different preprocessing techniques or model configurations might be required for datasets with such distributions.

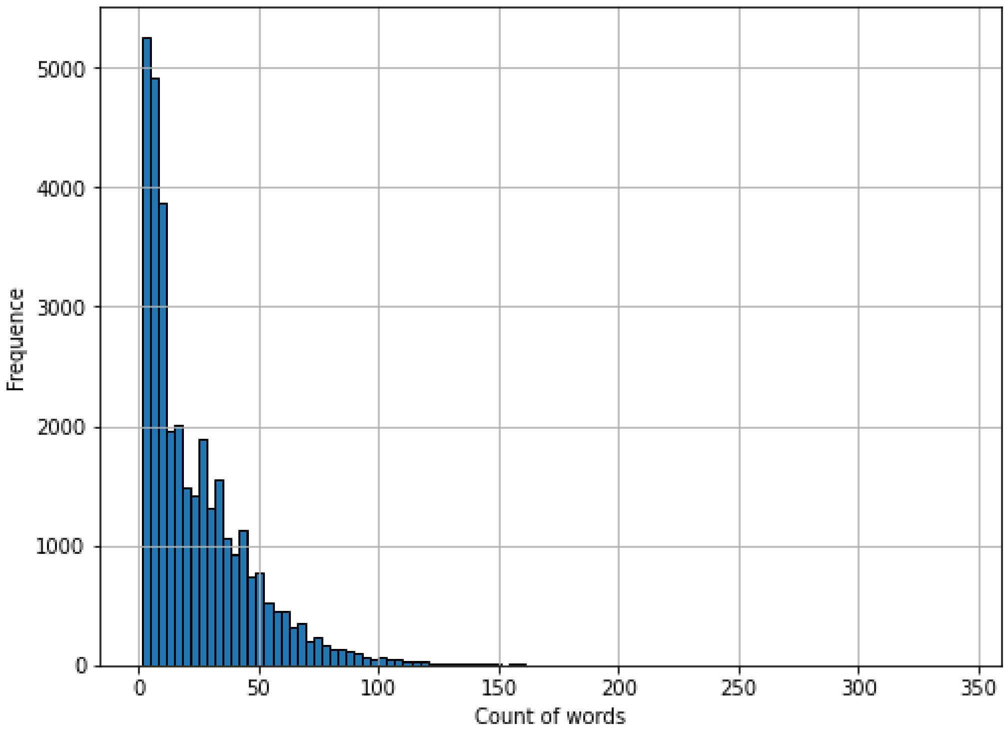


Figure 4.1: Size of text sensences distribuition

The *Inception* [Klie et al. 2018] tool was used for annotation, which allows user identification and role management. Therefore, it is possible to define who are the annotators and curators responsible for resolving conflicts between different annotations on the same piece of text.

The Figure 4.2 represents a real example of an excerpt from a corpus document with its respective annotations. As the corpus deals with documents in Brazilian Portuguese, the original text was kept to highlight the native characteristics of the annotated sentences. People's names were duly anonymized. For comparison purposes and to simplify understanding, the Figure A.1 containing the translation was made available in the appendix. No contextual or semantic modifications were made during translation.

**THE ARGUMENT TEXT**

*<Nome de pessoa anonimizado> conceitua o Mandado de Segurança como um remédio constitucional, com natureza de ação civil, posto à disposição de titulares de direito líquido e certo, lesado ou ameaçado de lesão, por ato ou omissão de autoridade pública ou agente de pessoa jurídica no exercício de atribuição do Poder Público.*
*(Curso de Direito Constitucional Positivo. São Paulo: Malheiros, 2003, pág. 446)*
*Em conjunto com as garantias constitucionais do hábeas corpus e do hábeas data, da inafastabilidade da jurisdição e do pleno acesso à Justiça, o mandado de segurança representa uma das maiores ferramentas de proteção dos direitos individuais e coletivos do cidadão contra os arbítrios de autoridades detentoras de poder.*

**ANNOTATIONS**

**DEFINITION**
*<Nome de pessoa anonimizado> conceitua o Mandado de Segurança como um remédio constitucional, com natureza de ação civil, posto à disposição de titulares de direito líquido e certo, lesado ou ameaçado de lesão, por ato ou omissão de autoridade pública ou agente de pessoa jurídica no exercício de atribuição do Poder Público.*

**REFERENCE**
*(Curso de Direito Constitucional Positivo. São Paulo: Malheiros, 2003, pág. 446)*

**ARGUMENTATION**
*Em conjunto com as garantias constitucionais do hábeas corpus e do hábeas data, da inafastabilidade da jurisdição e do pleno acesso à Justiça, o mandado de segurança representa uma das maiores ferramentas de proteção dos direitos individuais e coletivos do cidadão contra os arbítrios de autoridades detentoras de poder.*

Figure 4.2: Example of annotation of the resulting corpus. This example translated is attached on Appendice A.

# CHAPTER 5

# Experimental Evaluation

The Figure 5.1 shows and overview of proposed pipeline for Argument Mining in Brazilian Labor Justice.

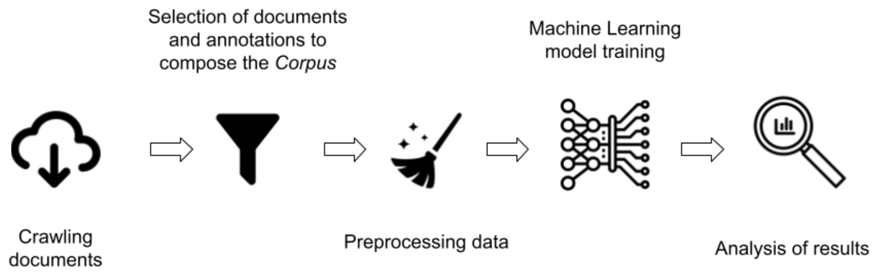


Figure 5.1: End-to-end cycle of learning model training for mining arguments.

The first step is marked by the scraping of documents from the Labor Courts portals [1]. We are interested to get metadata and textual documents that make up a labor lawsuit. With the exception of those under judicial secrecy, the lawsuits belonging to the labor courts are public.

Once you have the data, it is necessary to define the volume of information that will make up the experiment. As this work proposes to mine arguments, it would therefore be coherent to select documents where the presence of argumentative texts is more present, as is the case with initial petitions, where lawyers justify the requests made to their clients, as well as the sentencing documents, where judges base the decisions made on the requests. Furthermore, these documents were selected from different courts, to prevent the *Corpus* from being biased based on a specific Court. This stage, as well as the construction of the corpus, is marked by step 2 in Figure 5.1.

Before the data is presented to the model, it needs to go through a preprocessing process. This step was based on the same approach used by [Matthew et al. 2018, Chelba et al. 2013], which consists of the following steps:

[1] `https://www.pje.jus.br/navegador/`

1. Removal of all XML content (*tags*) that may be in the *corpus*;
2. Ordering of all sentences in the *corpus* so that repeated ones are eliminated;
3. Removal of sentences formed only by numbers, punctuation or characters outside the Portuguese alphabet. This step was added in this work with the aim of cleaning the *corpus*, eliminating lines such as those in Table 5.1.
4. Randomization of the remaining sentences, so that the language is learned without sticking to a specific claim of sentences;
5. Punctuation normalization and *tokenization* of the text;
6. Creation of a vocabulary keeping only words that were repeated at least 3 times in the *corpus*;

Table 5.1: Example of lines removed from Portuguese Wikipedia, formed only by numbers or characters outside the Portuguese alphabet.

| **Exemple removed lines** |
|---|
| ´ |
| ‘ |
| ∼ |
| < |
| <!– |
| <!– |
| <!– == |
| +————+—-+——————+ |
| – |
| __ |
| __ _*(#(#( |
| 1004 |

After the previous step, the data is subjected to Machine Learning model training (Step 4). The models used will be a pre-trained model on legal data [Garcia et al. 2024], a pre-trained model on portuguese language without using legal data [Souza, Nogueira e Lotufo 2020] and a pre-trained model base [He et al. 2020].

Finally, with the trained models, we were able to apply the test set to validate the training by analyzing the metrics obtained. This is the last step showed in Figure 5.1.

## 5.1 Parametrization, Training, and Experimental Setup

To train the learning model, a tool was used to automate the process of adjusting the parameters of an algorithm, called *GridSearch* [Adnan et al. 2022]. This tool systematically makes different combinations of parameters and after evaluating them, stores them in a single object. We used 4-fold-cross-validation, and results reported here are with this configuration.

The problem is being treated as a sentence classification task. Therefore, the resulting model will be given by a sentence classifier where, among several class possibilities, as discussed in the Section 4, one of them is ARGUMENTATION.

CHAPTER 6

# Results and Discussion

In this stage of the experiment, we employed the Grid Search technique to optimize the performance of a set of artificial intelligence models for a classification task. Three pre-trained models focused on natural language processing were utilized: RoBERTaLexPT-base [Garcia et al. 2024], bert-base-portuguese-cased [Souza, Nogueira e Lotufo 2020], and deberta-v3-base [He et al. 2020]. The training process was carried out using k-fold cross-validation with 4 folds to ensure robust model evaluation. Multiple **learning rates** were tested, including 1e-5, 9e-6, 2e-5, 3e-5, and 4e-5, in order to identify the most suitable configuration for fine-tuning each model. Additionally, three different **seeds** (10, 20, and 42) were used to control for randomness in the training process, aiming to enhance the reproducibility of the results.

Table 6.1: Comparison of F1-score grouped by learning rate for each model used

| Model | LR | Mean ± Std |
|---|---|---|
| **RoBERTaLexPT-base** | **1e-5** | **0.885736 ± 0.001854** |
| RoBERTaLexPT-base | 9e-6 | 0.885715 ± 0.001708 |
| RoBERTaLexPT-base | 2e-5 | 0.884025 ± 0.001891 |
| RoBERTaLexPT-base | 3e-5 | 0.881664 ± 0.002468 |
| RoBERTaLexPT-base | 4e-5 | 0.880146 ± 0.002131 |
| bert-base-portuguese-cased | 9e-6 | 0.869411 ± 0.003259 |
| bert-base-portuguese-cased | 1e-5 | 0.869275 ± 0.003755 |
| bert-base-portuguese-cased | 2e-5 | 0.868505 ± 0.002519 |
| bert-base-portuguese-cased | 3e-5 | 0.866900 ± 0.002395 |
| bert-base-portuguese-cased | 4e-5 | 0.865534 ± 0.002242 |
| deberta-v3-base | 4e-5 | 0.856723 ± 0.002873 |
| deberta-v3-base | 2e-5 | 0.856678 ± 0.002179 |
| deberta-v3-base | 3e-5 | 0.855770 ± 0.002715 |
| deberta-v3-base | 9e-6 | 0.851928 ± 0.002727 |
| deberta-v3-base | 1e-5 | 0.845968 ± 0.024663 |

Based on Table 6.1 the analysis of the performance results for the models RoBERTaLexPT-base, bert-base-portuguese-cased, and deberta-v3-base across different learning rates reveals some notable trends. For the RoBERTaLexPT-base model, the highest average F1-scores were achieved with learning rates of 1e-5 and 9e-6, both yielding a mean of 0.8857, with standard deviations of 0.001854 and 0.001708, respectively, indicating a high level of consistency. As the learning rate increased to 2e-5, 3e-5, and 4e-5, a gradual decline in the average F1-score was observed, with values of 0.8840, 0.8817, and 0.8801, respectively. The increasing standard deviation, reaching up to 0.002468, suggests greater variability in performance with higher learning rates.

For the bert-base-portuguese-cased model, the best performance was observed with lower learning rates, with a rate of 9e-6 yielding an average F1-score of 0.8694 and relatively low variation (standard deviation of 0.003259). As the learning rate increased, the average F1-score decreased, with the lowest value being 0.8655 at 4e-5. The standard deviation also fluctuated significantly across learning rates, ranging from 0.002242 to 0.003755, indicating variability in results.

Finally, the deberta-v3-base model demonstrated the lowest average F1-scores among the three models tested. The best performance was achieved with a learning rate of 4e-5 (mean of 0.8567 and standard deviation of 0.002873), while the worst performance was observed with a learning rate of 1e-5 (mean of 0.845968 and a higher standard deviation of 0.024663), indicating considerable variability, likely due to the presence of extreme values.

Overall, the statistical analysis indicates that RoBERTaLexPT-base exhibited the best performance, with lower variability and consistent results across different learning rates, particularly at lower rates. The bert-base-portuguese-cased model showed intermediate performance, while the deberta-v3-base model had the lowest results, with greater variability, especially at higher learning rates.

When considering "ARGUMENTATION" as the focal class, based on 6.1, 6.2 and 6.3, analysis reveals that RoBERTaLexPT-base achieves the highest performance with an accuracy rate of 88.86%, notably outperforming bert-base-portuguese-cased at 87.99% and deberta-v3-base at 86.29%. This margin suggests that RoBERTaLexPT-base is the most suitable model for tasks prioritizing precise "ARGUMENTATION" classification, likely due to its fine-tuning on Portuguese legal texts.

Given the critical role of "ARGUMENTATION" in this work, the performance difference here is statistically significant; even a few percentage points can greatly enhance classifier reliability for this task. The other two models, particularly deberta-v3-base, may not offer the same level of specificity and could benefit from further training on domain-specific datasets to improve performance in complex classes like "ARGUMENTATION."

Overall, RoBERTaLexPT-base's strength in this category, along with its stable performance across most other classes, solidifies its position as the optimal model for this study, especially where "ARGUMENTATION" accuracy is paramount.

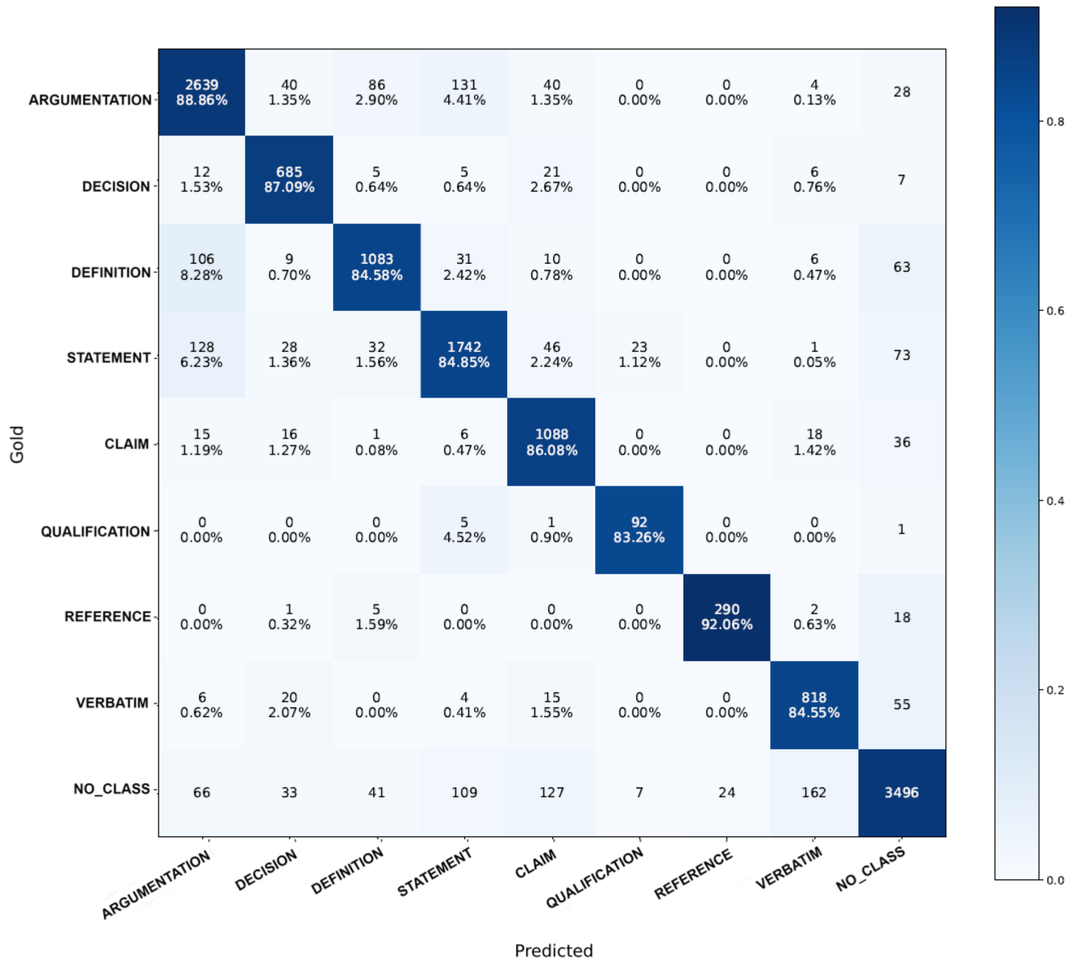


Figure 6.1: Confusion matrix by class from RoBERTaLexPT-base's results.

Taking Example 6.1 as a reference, we can observe that the model's highest error rates, in terms of both false positives and false negatives, occurred in the target class 'ARGUMENTATION' and were most frequently confused with the class 'STATEMENT.' A closer examination of the test dataset reveals that the sentence structures in these classes are very similar. Figures 6.4 and 6.5 show, respectively, a high concentration of sentences exceeding 50 words for each of these classes. In contrast, different patterns are observed for other classes. 'REFERENCE' and 'DECISION' have fewer instances of lengthy sentences, as shown in Figures 6.6 and 6.7. This may be a contributing factor to the model's increased confusion between 'ARGUMENTATION' and 'STATEMENT'.

Continuing with the analysis of the reasons behind the high confusion rate between the "ARGUMENTATION" and "STATEMENT" classes, we observe that both share a similar writing structure. While the content in 'STATEMENT' aims to detail how the facts occurred, sentences in "ARGUMENTATION" are intended to strengthen the narrative of the described events. This explains their similarity, and Table 6.2 presents

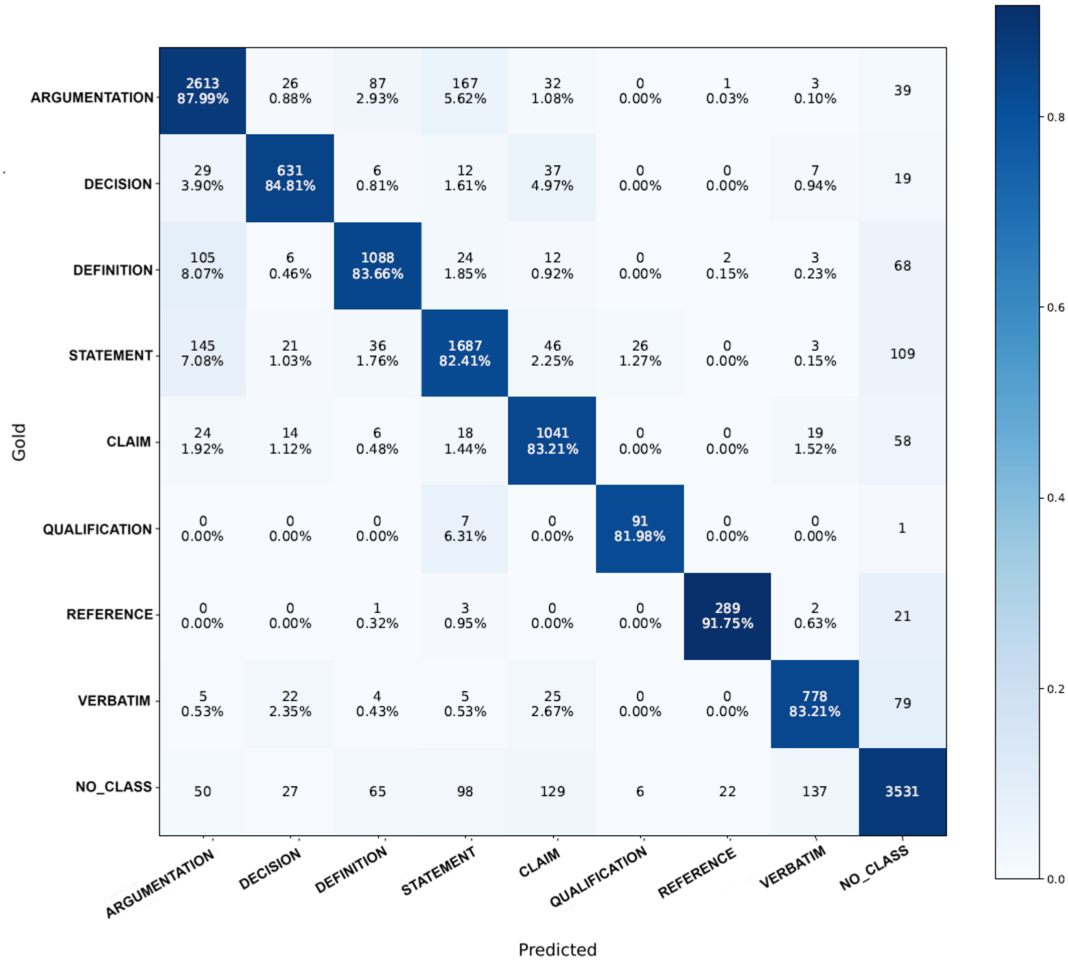


Figure 6.2: Confusion matrix by class from bert-base-portuguese-cased's results.

some examples.

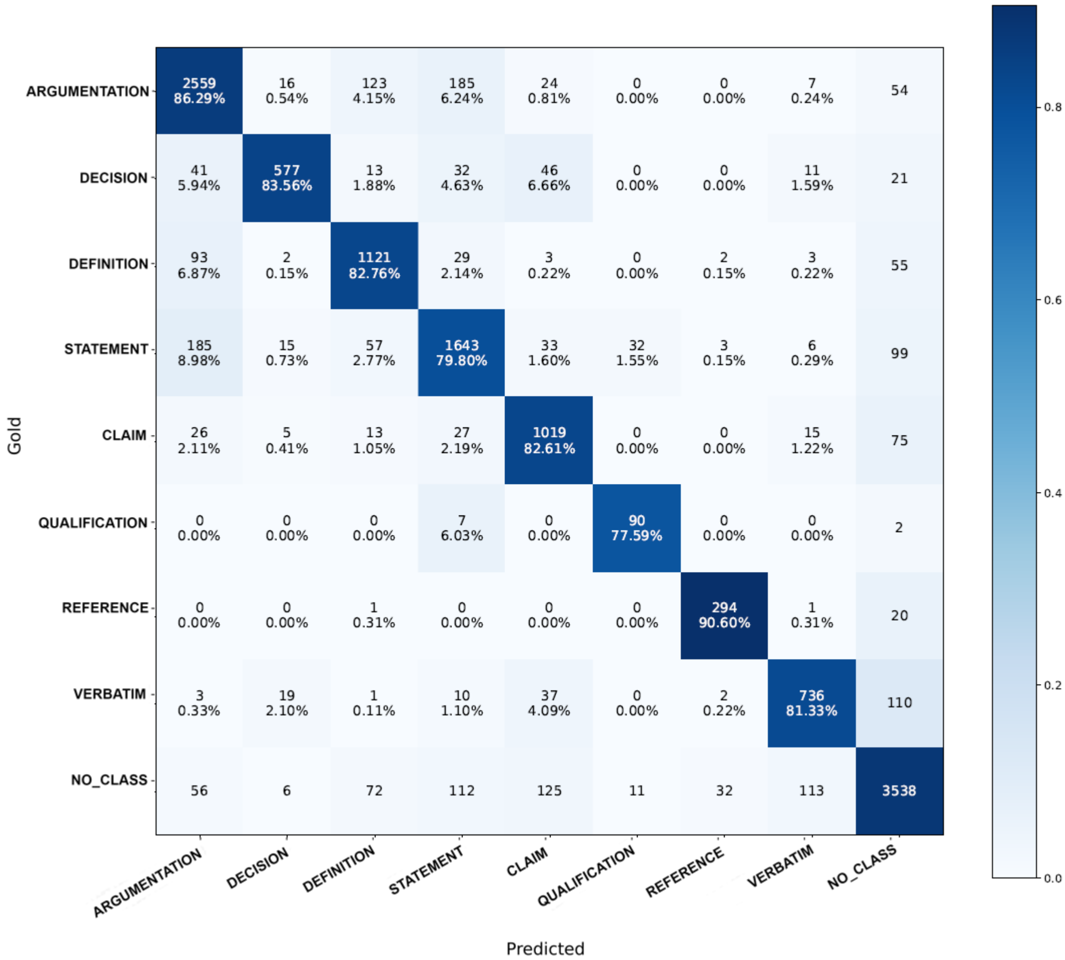


Figure 6.3: Confusion matrix by class from deberta-v3-base's results.

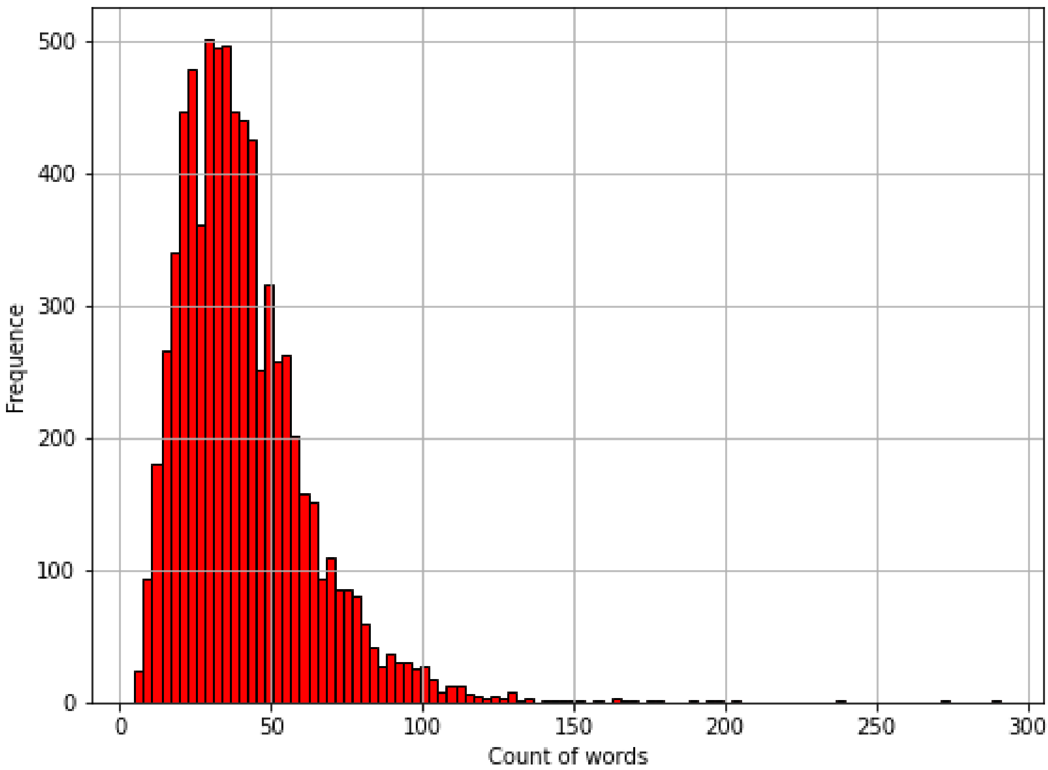


Figure 6.4: Distribuition of size sentences for class ARGUMENTATION

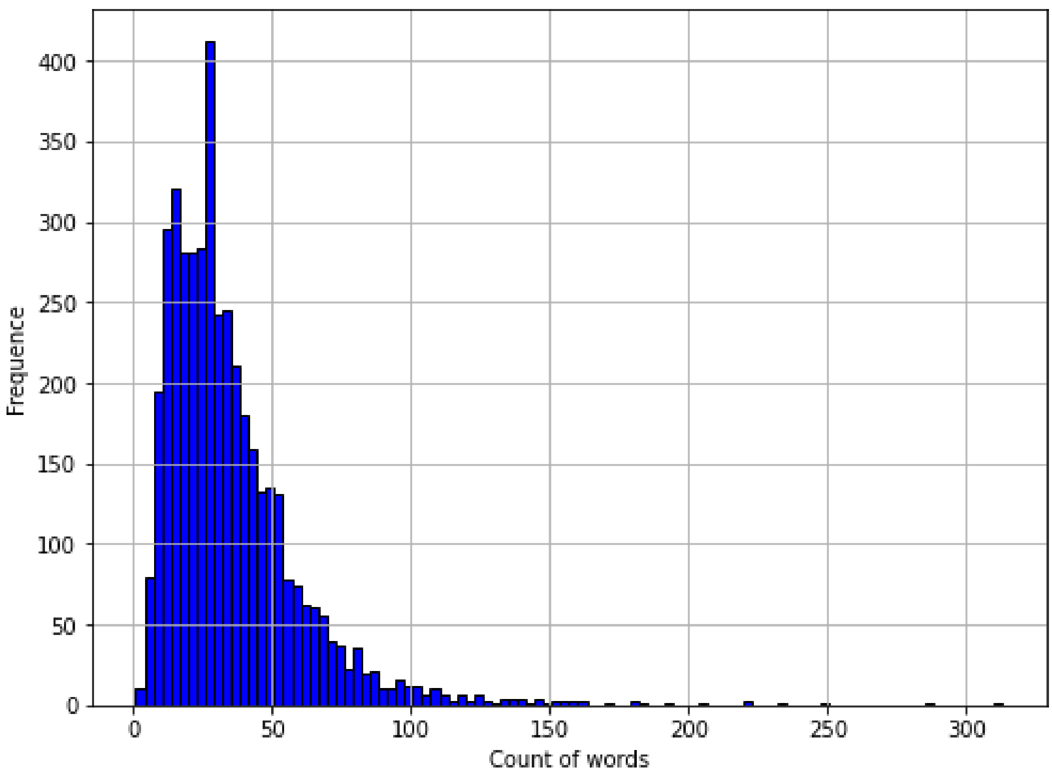


Figure 6.5: Distribuition of size sentences for class STATEMENT

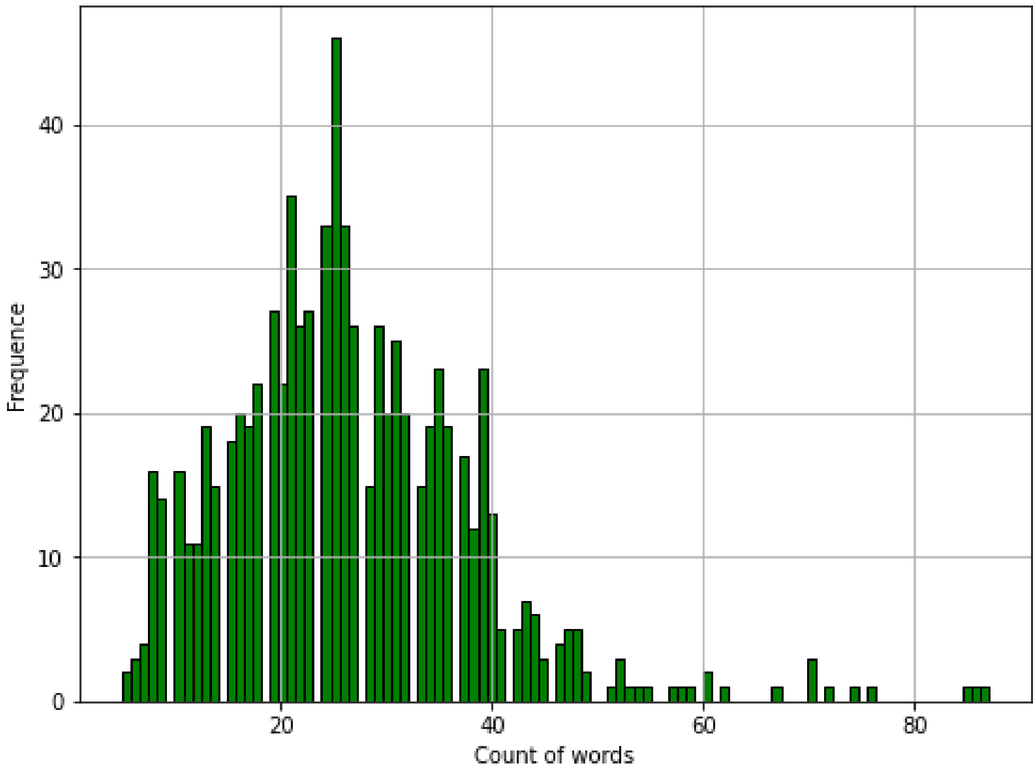


Figure 6.6: Distribuition of size sentences for class REFERENCE

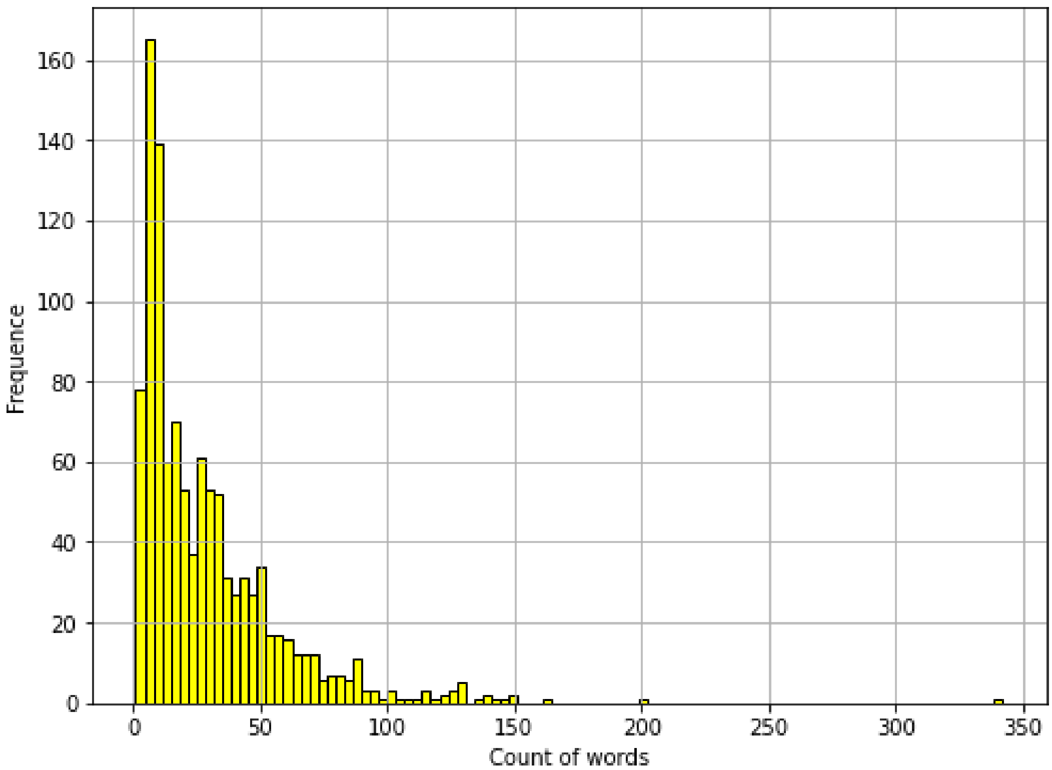


Figure 6.7: Distribuition of size sentences for class DECISION

Table 6.2: Some examples of texts labeled as ARGUMENTATION and STATEMENT

| **Text** | **Gold Label** | **Predicted** |
|---|---|---|
| Se não bastasse, a única testemunha ouvida - Sr. José Ricardo Marques Monteiro - não corroborou as ponderações tecidas na peça de ingresso. Muito pelo contrário, demonstrou que, além de não haver o acúmulo de função narrado na petição inicial, o serviço realizado pela reclamante não era pesado como a autor descreveu na exordial, conforme já enfrentado no tópico precedente. | **ARGUMENTATION** | STATEMENT |
| O reclamante afirma uma jornada de trabalho que efetivamente não existiu e que conforme já se afirmou não condiz com a realidade da atividade desenvolvida pela reclamada. | **ARGUMENTATION** | STATEMENT |
| Nos termos do acórdão regional , foi atribuída ao ente público , tomador de serviços , a responsabilidade subsidiária pelo adimplemento das obrigações trabalhistas devidas pela empresa prestadora de serviços , com fundamento no fato de que ele , a quem incumbiria o ônus da prova , por força do princípio da aptidão para a prova , não logrou demonstrar a efetiva fiscalização do contrato firmado com a prestadora . | STATEMENT | **ARGUMENTATION** |

CHAPTER 7

# Conclusion

In this work, the performance of the *Transformers* models was evaluated in the sentence classification task for the Portuguese language and legal domain. With the classifier it is possible to identify argumentative excerpts, the main focus of this work. To this end, a language model trained in a specific domain in legal documents (RobertaJur) and a *Corpus* created specifically to label, among other classes, argumentative excerpts were built.

Given the methodology applied, the results were quite satisfactory for us, given the initial state of the research. There are still experiments that need to be applied to compare this results with another model languages.

Although the research is in the initial results phase, it is already possible to consider the results promising. This is because the task is already able to be executed end to end with the proposal presented and also because it demonstrates that the hypothesis of building a specific domain corpus has a considerable contribution to training results.

As main contributions, we have:

- *Corpus* for the Portuguese Language built in the legal domain for Labor Justice;
- Model trained in a specific domain capable of classifying sentences from legal documents for the Portuguese language;

As future work we can mention the classification of identified arguments, given in related works, in Section 3 as the next stage of argument mining. As mentioned above, the classification of arguments is an old source of studies and was proposed by [Hardin 1959] and branched out by several other authors, such as [Freeman 2011].

Creating a corpus, as explored in Section 4, is no trivial task and requires specialized labor. Given that the research field involves legal documents, it is ideal for lawyers to perform the annotations to avoid potential misinterpretations of legal terminology. However, this research does not have sponsorship or financial support that would allow for hiring such resources, which led us to seek a partnership with the private sector. Through this partnership, the company would provide expert annotators; the corpus produced could then be used as a resource in this research, although its dissemination would remain restricted. Thus, it is worth noting that, although the annotated data are not publicly accessible, the partnership with the private sector enabled the production of specialized annotations at a pace more suited to the needs of this research. While the annotation process followed the company's proprietary protocol rather than the formal standard established by [Antoine, Villaneau e Lefeuvre 2014], this collaboration allowed for greater efficiency in corpus creation. It is believed that the results obtained may still significantly contribute to the advancement of the field, even if the corpus itself is restricted to this study.

The primary hypothesis of this research—whether arguments in Portuguese legal documents can be effectively identified through a sentence classification model using the Transformers architecture—has shown encouraging preliminary support. Our approach leveraged a domain-specific language model, RoBERTaLexPT [Garcia et al. 2024], trained on a specialized legal corpus tailored to the Portuguese language. The results indicate that the model is not only capable of identifying argumentative segments within legal texts but also performs well despite the inherent challenges posed by language-specific nuances and the complexity of legal terminology. These findings support our hypothesis and suggest that the use of a customized corpus in the legal domain contributes significantly to model accuracy, reinforcing the importance of a focused dataset for training language models within specialized domains.

# Bibliography

[Adnan et al. 2022]ADNAN, M. et al. Utilizing grid search cross-validation with adaptive boosting for augmenting performance of machine learning models. *PeerJ Computer Science*, PeerJ Inc., v. 8, p. e803, 2022.

[Antoine, Villaneau e Lefeuvre 2014]ANTOINE, J.-Y.; VILLANEAU, J.; LEFEUVRE, A. Weighted krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In: *EACL 2014*. [S.l.: s.n.], 2014. p. 10–p.

[Aragy, Fernandes e Caceres 2021]ARAGY, R.; FERNANDES, E. R.; CACERES, E. N. Rhetorical role identification for portuguese legal documents. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2021. p. 557–571.

[Ba, Kiros e Hinton 2016]BA, J. L.; KIROS, J. R.; HINTON, G. E. *Layer Normalization*. 2016.

[Basile 2012]BASILE, C. R. O. *Processo do trabalho: justiça do trabalho e dissídios trabalhistas*. [S.l.]: Saraiva, 2012.

[Chelba et al. 2013]CHELBA, C. et al. One billion word benchmark for measuring progress in statistical language modeling. *arXiv preprint arXiv:1312.3005*, 2013.

[Código de Processo Civil 2015]CóDIGO de Processo Civil. 2015. URL: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2015/lei/l13105.htm.

[Devlin et al. 2019]DEVLIN, J. 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, 2019. p. 4171–4186. Disponível em: <https://aclanthology.org/N19-1423>.

[Freeman 2011]FREEMAN, J. B. *Argument Structure:: Representation and Theory*. [S.l.]: Springer Science & Business Media, 2011.

[Garcia et al. 2024]GARCIA, E. A. S. et al. Robertalexpt: A legal roberta model pretrained with deduplication for portuguese. In: *Computational Processing of the Portuguese Language*. [S.l.]: Association for Computational Linguistics, 2024.

[Hardin 1959]HARDIN, C. L. Stephen e. toulmin the uses of argument. new york: Cambridge university press, 1958. 264 pp. 4.50.*Philosophy of Science*, $CambridgeUniversityPress, v. \quad 26, n. \quad 2, p.160 - -163, 1959$.

[He et al. 2020]HE, P. et al. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*, 2020.

[Klie et al. 2018]KLIE, J.-C. et al. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In: . [S.l.: s.n.], 2018.

[Liu et al. 2019]LIU, Y. et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019.

[Matthew et al. 2018]MATTHEW, E. P. et al. Deep contextualized word representations.(2018). *Access mode: https://arxiv. org/abs*, 2018.

[Mochales e Moens 2008]MOCHALES, R.; MOENS, M.-F. Study on the structure of argumentation in case law. In: . [S.l.: s.n.], 2008. p. 11–20.

[Mochales e Moens 2011]MOCHALES, R.; MOENS, M.-F. Argumentation mining. *Artif. Intell. Law*, Kluwer Academic Publishers, USA, v. 19, n. 1, p. 1–22, mar 2011. ISSN 0924-8463. Disponível em: <https://doi.org/10.1007/s10506-010-9104-x>.

[Nascimento e Fernandes 2023]NASCIMENTO, E. D.; FERNANDES, S. R. T. A inaÇÃo forÇada como prÁtica de assÉdio moral: O olhar da jurisprudÊncia trabalhista brasileira. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, v. 9, n. 5, p. 371–389, maio 2023. Disponível em: <https://www.periodicorease.pro.br/rease/article/view/9767>.

[Poudyal et al. 2020]POUDYAL, P. et al. Echr: Legal corpus for argument mining. In: *Proceedings of the 7th Workshop on Argument Mining*. [S.l.: s.n.], 2020. p. 67–75.

[Prasad et al. 2008]PRASAD, R. et al. The penn discourse treebank 2.0. In: *LREC*. [S.l.: s.n.], 2008.

[Qiu et al. 2020]QIU, X. et al. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, Springer Science and Business Media LLC, v. 63, n. 10, p. 1872–1897, Sep 2020. ISSN 1869-1900. Disponível em: <http://dx.doi.org/10.1007/s11431-020-1647-3>.

[Reale 2001]REALE, M. *Lições Preliminares de Direito. 2. ª tiragem*. [S.l.]: Saraiva, 2001.

[Rocha et al. 2018]ROCHA, G. et al. Cross-lingual argumentative relation identification: from english to portuguese. In: *Proceedings of the 5th Workshop on Argument Mining, 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*. [S.l.: s.n.], 2018.

[Rocha 2016]ROCHA, G. F. da. *ArgMine: Argumentation Mining from Text*. Dissertação (Mestrado) — Universidade do Porto (Portugal), 2016.

[Savelka et al. 2017]SAVELKA, J. et al. Sentence boundary detection in adjudicatory decisions in the United States. *Traitement Automatique des Langues*, ATALA (Association pour le Traitement Automatique des Langues), France, v. 58, n. 2, p. 21–45, 2017. Disponível em: <https://aclanthology.org/2017.tal-2.2>.

[Silveira et al. 2023]SILVEIRA, R. et al. Legalbert-pt: A pretrained language model for the brazilian portuguese legal domain. In: SPRINGER. *Brazilian Conference on Intelligent Systems*. [S.l.], 2023. p. 268–282.

[Souza, Nogueira e Lotufo 2020]SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Bertimbau: pretrained bert models for brazilian portuguese. In: SPRINGER. *Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23, 2020, Proceedings, Part I 9*. [S.l.], 2020. p. 403–417.

[Stab e Gurevych 2014]STAB, C.; GUREVYCH, I. Identifying argumentative discourse structures in persuasive essays. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. [S.l.: s.n.], 2014. p. 46–56.

[Sutskever, Vinyals e Le 2014]SUTSKEVER, I.; VINYALS, O.; LE, Q. V. Sequence to sequence learning with neural networks. In: GHAHRAMANI, Z. et al. (Ed.). *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2014. v. 27.

[Toutanova et al. 2003]TOUTANOVA, K. et al. Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of the 2003 human language technology conference of the north american chapter of the association for computational linguistics*. [S.l.: s.n.], 2003. p. 252–259.

[Vaswani et al. 2017]VASWANI, A. et al. Attention is all you need. In: GUYON, I. et al. (Ed.). *Advances in Neural Information Processing Systems*. [S.l.]: Curran Associates, Inc., 2017. v. 30.

[Walton 2009]WALTON, D. Argumentation theory: A very short introduction. In: ____. *Argumentation in Artificial Intelligence*. Boston, MA: Springer US, 2009. p. 1–22. ISBN 978-0-387-98197-0. Disponível em: <https://doi.org/10.1007/978-0-387-98197-0$_1$>.

[You et al. 2020]YOU, Y. et al. *Large Batch Optimization for Deep Learning: Training BERT in 76 minutes*. 2020.

APPENDIX A

# Appendix

**THE ARGUMENT TEXT**

*<Name of person anonymized> conceptualizes the Writ of Mandamus as a constitutional remedy, with the nature of a civil action, made available to holders of clear and certain rights, injured or threatened with injury, by act or omission of a public authority or agent of a person legal in the exercise of Public Power attribution.*
*(Course of positive constitutional law. São Paulo: Malheiros, 2003, p. 446)*
*In conjunction with the constitutional guarantees of habeas corpus and habeas data, of non-defeasibility of jurisdiction and full access to Justice, the writ of mandamus represents one of the greatest tools for protecting the individual and collective rights of citizens against the discretion of authorities holding power.*

**ANNOTATIONS**

**DEFINITION**
*<Name of person anonymized> conceptualizes the Writ of Mandamus as a constitutional remedy, with the nature of a civil action, made available to holders of clear and certain rights, injured or threatened with injury, by act or omission of a public authority or agent of a person legal in the exercise of Public Power attribution.*

**REFERENCE**
*(Course of positive constitutional law. São Paulo: Malheiros, 2003, p. 446)*

**ARGUMENTATION**
*In conjunction with the constitutional guarantees of habeas corpus and habeas data, of non-defeasibility of jurisdiction and full access to Justice, the writ of mandamus represents one of the greatest tools for protecting the individual and collective rights of citizens against the discretion of authorities holding power.*

Figure A.1: Example of annotation of the resulting corpus.