INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Contributions to Automatic Legal Document Summarization – Judgements from the Portuguese Supreme Court

Margarida Rebelo Dias

Master in Computer Engineering

Supervisor:
PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor, Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
PhD Helena Sofia Andrade Nunes Pereira Pinto, Assistant Professor, Instituto Superior Técnico, Universidade de Lisboa

September 2024

TECHNOLOGY
AND ARCHITECTURE

Department of Information Science and Technology

Contributions to Automatic Legal Document Summarization – Judgements from the Portuguese Supreme Court

Margarida Rebelo Dias

Master in Computer Engineering

Supervisor:
PhD Ricardo Daniel Santos Faro Marques Ribeiro, Associate Professor, Iscte – Instituto Universitário de Lisboa

Co-Supervisor:
PhD Helena Sofia Andrade Nunes Pereira Pinto, Assistant Professor, Instituto Superior Técnico, Universidade de Lisboa

September 2024

Acknowledgment

First of all, I would like to express my deep gratitude to my supervisors, Professor Ricardo Ribeiro and Professor Helena Sofia Pinto, for their guidance and support. Thank you for always being available and for providing the valuable knowledge that made this thesis possible.

I would also like to thank my family. To my parents, Manuela and Agostinho, thank you for always believing in me and encouraging me to keep going. Without you, this project would not have been possible. And to my sister, Mariana, thank you not only for your constant support and words of encouragement but also for being an example and inspiration throughout my journey.

Margarida Rebelo Dias

Resumo

Com o aumento exponencial das diferentes formas de informação, ultrapassando a capacidade humana de as acompanhar, torna-se crucial desenvolver estratégias que minimizem o tempo gasto tanto na leitura como na compreensão da informação. No meio jurídico, o processo de sumarização tem sido requerido para este fim, no entanto sendo feito manualmente.

Esta dissertação foca-se na avaliação de diferentes modelos de sumarização cujo objetivo é entender a eficácia dos mesmos na automatização do processo de sumarização, especificamente para documentos jurídicos portugueses do Supremo Tribunal de Justiça.

Diferentes modelos de sumarização têm sido desenvolvidos em várias áreas. O meio jurídico apresenta algumas limitações devido não só à extensão dos documentos, mas também ao vocabulário específico utilizado. Neste trabalho, foram desenvolvidos três modelos: um modelo ao nível das frases, um modelo ao nível do sumário e uma abordagem híbrida. Estas implementações tiveram como objetivo perceber as diferenças na geração de sumários usando tanto modelos de sumarização extrativos quanto abstrativos.

Para cada implementação, usámos dois tipos de input: os documentos originais e secções específicas dos documentos. Para a fase de avaliação, usamos as métricas de avaliação ROUGE e BERTscore, onde comparamos os sumários gerados com os de referência.

A análise dos resultados levou-nos a concluir que os modelos extrativos são eficazes na redução do tamanho dos documentos, especialmente no modelo ao nível do sumário e a utilizão de algoritmos abstractivos permite tornar o texto mais fluído. Além disso, verificou-se que a experiência ao nível do sumário teve um impacto substancial no processo de sumarização de documentos jurídicos portugueses.

Palavras Chave: *sumarização de texto automática*, *sumarização de documentos jurídicos*, *sumarização extrativa*, *sumarização abstrativa*, *Português Europeu*

Abstract

As information continues to grow in an exponential way, overtaking humans capacity to reach all of it, it is crucial to develop strategies to minimize the time spent on reading and comprehending information. In the legal field, the process of summarization has been used for this purpose, however, it is still done manually by legal experts.

This dissertation focuses on testing different summarization models in order to understand their efficacy in automating the summarization process, specifically for Portuguese legal documents from the Portuguese Supreme Court of Justice.

Automatic summarization models have been developed in a variety of areas. Conversely, the legal field brings some constraints because of the length of the documents and the particular vocabulary used in them. We implemented three different models: a sentence-level model, a summary-level model, and a hybrid approach to evaluate the generation of summaries using both extractive and abstractive summarization methods.

For each experiment, we used two different input texts: the original documents and specific sections from the original documents. For the evaluation process, we use the ROUGE and BERTscore metrics, where we compare the generated summaries with the reference summaries available for each document.

The analysis of the results made us conclude that the extractive models are effective at reducing document length, particularly with the summary-level approach, and that abstractive techniques can improve summary fluency. Furthermore, it was confirmed that the use of a summary-level approach has a significant effect on the summarization of Portuguese legal documents.

KEYWORDS: *automatic text summarization*, *legal document summarization*, *extractive summarization*, *abstractive summarization*, *European Portuguese*

Contents

List of Figures

List of Tables

List of Acronyms

AI: Artificial Intelligence

NLP: Natural Language Processing

DL: Deep Learning

ML: Machine Learning

TS: Text Summarization

ETS: Extractive Text Summarization

ATS: Abstractive Text Summarization

HTS: Hybrid Text Summarization

SLR: Systematic Literature Review

PRISMA: Preferred Reporting Items for Systematic reviews and Meta-Analyses

ROUGE: Recall-Oriented Understudy for Gisting Evaluation

TF-IDF: Term Frequency-Inverse Document Frequency

LSTM: Long Short-Term Memory

Seq2Seq: Sequence-to-sequence

LLMs: Large Language Models

CHAPTER 1

Introduction

In this era, marked by the exponential growth of data, it has become impossible for individuals to keep pace with the amount of information that is generated every second on the Internet. The legal domain is no exception to this phenomenon. Portuguese legal professionals, such as judges and lawyers, often contend with extensive legal documents in their daily work. Access to faster, more concise, and reliable information is crucial for enhancing efficiency and ensuring effective decision-making. Natural Language Processing (NLP) is an area of Artificial Intelligence (AI) that is contributing to the process of automating how computers can understand and generate human language [1]. Text Summarization (TS) is a task of NLP that focuses on the synthesis of information from a larger text into a concise format. TS is a process that is able to capture important words and information from a text without losing its context [2]. By applying TS to Portuguese legal documents, Portuguese judges could gain access to summaries that allow them to save time and gain a broader perspective on a larger number of cases. This would contribute to a more efficient legal workflow and improved case management.

The TS concept was first introduced by Hans Peter Luhn in the 1950s when he reduced a text by focusing on the importance of sentences based on the frequencies of the words [3]. Since then, improvements in NLP have been made, and the concept of Extractive Text Summarization (ETS) was established in the TS field. ETS approaches focus on linguistic features and statistical models in order to identify the most pertinent elements of the texts, which are then extracted directly from the text to form the summaries. Subsequently, the objective of generating more accurate summaries that could be considered similar to those written by humans led to the development of Abstractive Text Summarization (ATS) models. ATS models were more focused on understanding the context and the relation between the words in a text. In contrast to ETS, these algorithms can generate summaries that may contain words or sentences that are not present in the original document. Nevertheless, the summary generated by these algorithms are capable of capturing the fundamental concepts and ideas present in the original text.

Despite these advancements, TS is not yet at its optimal level, representing one of the most challenging areas of NLP. The complexity of the syntactic and semantic aspects of the text makes it difficult for a machine to understand its meaning, thus providing scope for improvements [4].

The purpose of this study was to explore, implement, and evaluate a range of TS models and approaches with the aim of overcoming existing challenges and contributing to the advancement of knowledge in the field of TS. Our main purpose was to examine

the summarization of legal documents, with a particular focus on Portuguese judgments from the Portuguese Supreme Court Justice.

1.1. Overview

This dissertation explores the application of TS techniques to Portuguese legal documents. A Systematic Literature Review (SLR) methodology [5], following the PRISMA guidelines [6], was employed to identify the most relevant TS models and techniques, as well as to acknowledge the main limitations in the field of legal document summarization that require improvements.

The main goal of this research was to evaluate the performance of different summarization models and techniques. Specifically, we investigated the LexRank algorithm [7], an ETS approach that uses a graph-based ranking strategy mechanism to select the most informative sentences from a document, and the MBART model [8], known for being a Sequence-to-sequence (Seq2Seq) encoder-decoder model pretrained on a multilingual corpus, thus enabling the generation of Portuguese texts from a given input text. Different techniques of summarization were analyzed, including the generation of sentence-level and summary-level summaries.

We used a corpus of Portuguese legal documents from the Portuguese Supreme Court of Justice as input, which allowed us to implement the models. The performance of the models was evaluated using standard summarization techniques, namely Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [9] and BERTscore [10].

The greatest challenges in this process were the application of the summarization techniques to lengthy documents and achieving accurate summaries for the Portuguese language.

This dissertation made significant contributions to the field through the development of a novel hybrid method capable of handling lengthy documents by combining the strengths of both extractive and abstractive models. Additionally, the study reveals that focusing on particular sections of the Portuguese legal documents may result in the generation of more accurate summaries. Future research could be conducted to improve this hybrid approach by incorporating domain-specific language embeddings for the extractive component and fine-tuning the abstractive model for a more contextualized summaries.

In summary, the purpose of this thesis was to expand the comprehension and application of summarization techniques in the legal domain while contributing valuable insights and practical solutions to the field of NLP.

1.2. Motivation

It is important that we can efficiently extract information and knowledge from existing sources without spending too much time. Summaries are useful for quickly understanding the main concepts, ideas and message of documents, which play a crucial role in a diversity of areas, including the legal domain.

Legal documents are formally written texts that describe laws, regulations, contracts, and other important information that in principle, should be easily accessible to everyone. These publications provide legal professionals with insights into past cases and legal precedents that help them prepare for future proceedings [11]. Legal cases also act as crucial learning materials for law students, allowing them to comprehend legal concepts and how they should be implemented. It is also vital that the general public is able to understand these documents, as they provide avenues for comprehending rights and responsibilities [12].

However, the difficulty and complexity of obtaining useful information from legal documents continue to be a challenge for legal experts and non-jurists. The creation of manual summaries has been a solution to this problem. Consequentially, performing it manually can be very time-consuming and cost-inefficient, especially for long documents, as legal documents are, that require a lot of human effort due to the variety and length of the lexica [13]. In response to this manual approach, researchers have been exploring NLP and Machine Learning (ML) techniques, to offer the finest automatic TS model.

Automatic TS has been applied to different types of texts for over two decades, with significant advances emerging through the development of recent Large Language Models (LLMs). LLMs are trained on large datasets and capable of understanding the context and capturing complex semantic relationships within a text. These models significantly increase the quality and coherence of generated summaries, resulting in outputs that are comparable to those created by humans [14]. Despite this, legal documents are a particular type of document, which increases the difficulty of effectively accomplishing this task. Generally, these kinds of documents have an extensive size, presenting a significant challenge for automatic summarization models to include all essential information within the constraints of a limited output length [15]. Furthermore, legal documents have distinct and complex lexica, which require incorporating domain-specific knowledge and improving contextual understanding into NLP models. Additionally, each country has its own legal system, characterized by its own language and legislation, which challenges the adaptation of models that have already demonstrated good results in each country's domain [16]. These requirements create a significant challenge for automated systems to accurately comprehend and summarize legal content.

In the end, the goal behind this research lies in the crucial need to develop or improve automatic summarization techniques that can efficiently reduce the time and effort invested in extracting essential data while also improving access to important legal information.

1.3. Research Questions

The objective of this work is to explore different TS models in the field of legal documents, with focus on Portuguese judgments. In this section, we delineate a set of research questions to guide the exploration of automatic TS techniques for Portuguese legal documents. The referred questions and their objectives are listed below:

- What methods can accurately assess the quality of summaries in terms of capturing relevant information from the original document and the coherence and fluency of the generated summaries? This question aims to explore the TS techniques used to evaluate the quality of summaries and identify the most appropriate ones.
- What impact does the length of summaries have on their quality? This question aims to investigate the impact of different summary sizes. By analyzing different summary lengths, the goal is to understand if there is any optimal length to the Portuguese legal document summaries.
- When dealing with lengthy documents, more information needs to be considered in order to capture the overall context of the documents. How do TS models deal with lengthy input texts and capture the key information of the original document? The purpose of this question is to explore how TS models can process lengthy inputs and the potential effects on summary quality.
- Legal documents are composed of different elements, including a description, facts, and a final decision. Do all the details of the documents contribute equally to the summaries, or are there specific sections that are more relevant? The objective of this question is to determine whether the generation of a summary with specific elements of the document will improve the quality of the summary compared with generating a summary with the entire document.
- What privacy and legal considerations must be taken into account when summarizing Portuguese legal documents that can contain sensitive data? When dealing with sensitive data, it is necessary to pay attention to how NLP technologies and models deal with the data. The goal of this question is to ascertain how TS models process data and whether there is a potential risk associated with the sharing of the data.
- Which extractive and abstractive techniques offer the potential to effectively summarize legal documents while retaining the context and readability of the original document? The purpose of this question is to evaluate summarization methods in terms of their capability to extract crucial information and generate concise yet coherent summaries.

In this work, we investigate how state-of-the-art approaches for other TS processes perform and how we could adapt them to align with our domain. We also focus on identifying the inherent limitations in order to improve our models and overcome the existing techniques. These questions helped us to follow a more systematic approach in our research, enabling us to select the most appropriate models and techniques for the project.

1.4. Document Structure

This document consists of seven chapters, including the **Introduction**. Below is a brief overview of each chapter.

Chapter 2 describes key concepts that are essential for understanding this study. It defines NLP and outlines the main techniques employed in the automatic TS field. This chapter also discusses the fundamental characteristics of Automatic TS. It explores the different input types as well as the potential output summaries, including extractive, abstractive, and hybrid summaries. Additionally, a brief overview is provided of two evaluation metrics that were employed to assess the quality of the generated summaries.

Chapter 3 outlines the research methodologies employed in this work, followed by an overview of the research process carried out in order to obtain relevant information about the work done in this field. Next, a detailed review of the existing literature is presented, in which the main four tasks of TS are described addressing the topics: preprocessing, feature extraction and text representation, summarization algorithms and techniques, and finally evaluation and results. Each section focuses on the methods and algorithms applied to documents and highlights the limitations of current works. Additionally, an overview of the main topic learned is conducted, with the aim of summarizing relevant insights for this research and identifying potential approaches for innovative contributions.

Chapter 4 focuses on detailing the dataset that was used in this study. It begins by describing the structure of the Portuguese legal judgments from the Portuguese Supreme Court of Justice. The chapter proceeds to outline the preprocessing steps carried out on the legal documents and presents an analysis of the structural properties along with their most frequent words.

Chapter 5 outlines the three approaches used in this study: a sentence-level approach, a summary-level approach, and a hybrid approach. For each implementation, we describe in detail the algorithms used and how we applied them based on the characteristics of our dataset. Each of these sections ensures a clear understanding of the methodologies used and the reasons behind their selection.

Chapter 6 presents the results obtained from implementing the approaches explained in Chapter 5. The performance of each model is evaluated via the ROUGE and BERTscore metrics by comparing the generated summaries with the reference summaries. An analysis of the structure and content of the generated summaries is also provided to assess the effectiveness of each algorithm. Overall, this chapter highlights the success and limitations of the implemented models in generating summaries.

Finally, **Chapter 7** discusses the results analyzed in Chapter 6, highlighting their contributions to the domain of the legal field as well as their limitations. It also suggests potential improvements for future research and development in this area.

CHAPTER 2

Background

This chapter outlines the fundamental concepts that are essential for a comprehensive understanding of the project domain. We start by introducing the concept of NLP and its principal techniques. Subsequently, we presented the definition of automatic TS along with an overview of its general characteristics. Finally, the evaluation metrics used to assess the performance of the TS models and the quality of the summaries generated were described.

2.1. Natural Language Processing

According to Khurana, Koli, Khatter, *et al.* [17], NLP is a branch of AI and Linguistics devoted to making computers understand statements or words written in human languages. NLP has the capacity to process large amounts of text data and to give valuable insights from a given input by employing different techniques for understanding and analyzing texts. In recent years, the advancement of NLP has been driven by the introduction of transformer-based models, which have significantly improved the ability to generate coherent texts. These models are able to capture complex relationships within a text, allowing them to produce high-quality summaries, translations and other forms of generating content [18]. Additionally, NLP relies on advanced Deep Learning (DL) and ML models in order to deliver accurate results.

NLP has become important in a wide range of fields by narrowing the gap between human communication and computational systems. Several tasks have emerged to enable computers to interpret and process the human language more accurately. Some key tasks and concepts are listed below.

Tokenization: is the process of splitting words, sentences, or documents, into a smaller units known as tokens. The goal of separating a piece of text is to generate structured data that can be used as a representation of a text in a form that will facilitate further analysis techniques.

Part-of-Speech Tagging: mainly consists of labeling a word in a text with the corresponding grammatical category, such as nouns, verbs, adjectives, and others. This approach helps to understand the sentence syntactic structure, which will be useful in other NLP tasks, such as information extraction, translation, or name entity recognition.

Word Embeddings: are a form of representing words in a continuous vector space. These representations are used for capturing context and similarity between

words, having the ability to compare words or sentences by their semantic meaning.

Text Generation: is the process where texts are written in a coherent and meaningful format by computer systems. Text Generation tasks employ AI models in order to provide an output text that is similar to human language patterns and style.

Topic Modeling: is a technique that can recognize a group of related words in the context of a document. The algorithms attempt to identify latent semantic structures in a database, uncovering patterns that connect words to determine the main topics in a document.

Text Summarization: reduces a longer text into a shorter version, either selecting the essential paragraphs and combining them or generating new sentences while retaining the main idea of the input text.

The application of NLP is critical in this work, as it provides a vast number of techniques that allow us to implement a complete process to generate summaries. The use of NLP techniques will enable the representation of legal documents, the extraction of the main keywords, and the identification of the essential topics in the context of legal documents. Furthermore, it allows for the efficient improvement of the understanding of legal documents as well as the generation of summaries.

2.2. Automatic Text Summarization

Automatic TS is the process of automatically generating text that captures the key concepts and preserves the meaning of an original textual document or documents in a shorter version. With the increasing number of legal documents, the use of these systems becomes appealing to both the general public and all stakeholders involved in the legal domain (legislators, judges, lawyers, students, etc). The main goal when implementing an automatic TS system is to create a robust model that can generate a shorter text that captures the content of the original documents [19].

Nowadays, the requirements to generate a summary are more demanding, and identifying relevant sentences from a document is not enough. The quality and fluency of the summary are also important aspects to take into consideration. The final objective is to reduce the cost, effort, and time required to make a summary as understandable and readable as possible [14].

Every day, different scenarios emerge, and it is necessary to constantly adapt the TS model to the requirements. The specification of each parameter influences the implementation of the models, dictating how algorithms must handle each situation. Several factors need to be considered to ensure that summarization techniques are appropriately tailored to meet this specific requirement, including [20]:

- The type and domain of the input text,
- The length of both the input and output texts,
- The purpose of the summary,

- The language of the input text and final summary,
- The computational resources available to implement the models.

These considerations are critical when selecting the appropriate models and determining which parameters best suit the specific task.

When referring to the type of input text, a summary can be classified into two different strategies depending on the number of documents that are used to generate the summary [13]: single-document summarization and multi-document summarization.

Single-document summarization: In this case, a model is used to create a single summary for a respective document. The goal of this technique is to convert a single document into a shorter version while maintaining the relevant information.

Multi-document summarization: In this case, a summary is generated from a set of documents, which emphasizes the overall context. This approach facilitates the acquisition of general knowledge about the information contained in a range of documents. This approach is valuable when there is a significant amount of common information between a set of documents, and it is not necessary to read all the documents to understand their context.

Depending on the requirements, the output summary can take different forms. Summaries can be either extractive, where sentences are selected based on different criteria, or abstractive, where a new text is written based on the context of the input. In both processes of generating summaries, there are advantages and limitations, sometimes combining them can be the perfect solution to generate a summary that grabs all the advantages of each approach while overcoming their limitations. This type of process is called a hybrid TS method. To better understand the difference between these three TS methods, we describe in more detail each of them below.

Extractive Text Summarization (ETS): involves determining the most relevant sentences from the original document so that they can be combined to generate a summary. This technique aims to preserve the key concepts from the input text and present them in a condensed form [21]. The most important task in the ETS process is the selection of the most significant sentences to be included in the summary. There are several techniques to implement this task, including graph-based methods, linguistic-based methods, cluster-based methods, and methods based on ML and neural network models [22]. While ETS can capture and maintain the original text key information it will also increase the possibility of grammatically unnatural outcomes due to the concatenation of the selected sentences.

Abstractive Text Summarization (ATS): goes beyond simply extracting and reorganizing sentences from the original data. ATS returns a representation of the real text, by first deeply understanding the entire content and then generating a new shorter text often using paraphrasing [23]. This type of summarization

makes use of NLP and DL techniques, which allow the generation of a summary that is similar to the ones made by humans. It can produce grammatically accurate sentences while also considering the semantics of the source text.

Hybrid Text Summarization (HTS): is known for the combination of ETS and ATS methods, designed to leverage the strengths of each model to generate the final summary. Usually, its implementation is divided into two tasks: first, the key concepts and relevant information are selected from the documents using an extractive model that will serve as input for the second task. In this second task, the goal is to use the strength of the ATS methods to generate a fluent and readable summary that captures the main idea of the original text [24].

Due to the use of neural networks, automatic TS has advanced significantly recently, although, it involves complex language modeling, which makes it a difficult task to accomplish [25].

Despite the strong emphasis on developing accurate models, automatic TS still faces several challenges, such as the interpretation and selection of pertinent sentences, particularly within lengthy documents, the preservation of summary quality and fluency, and the evaluation process that fails to align with the desired outcomes.

2.3. Evaluation Metrics

Evaluating the effectiveness and quality of TS techniques is a critical aspect of research in NLP. To assess the performance of summarization approaches and provide quantitative insights into their capabilities, researchers commonly employ the ROUGE metric. BERTscore is another evaluation model available to assess the quality and fluency of the generated summaries. In this section, we describe in detail how these two metrics can be used to evaluate the process of TS:

ROUGE: is a metric used to evaluate TS and translation models. It facilitates a comparison between a generated summary from a model and a reference summary, usually created by humans [9]. The two most commonly used ROUGE metrics in the context of TS are ROUGE-N and ROUGE-L.

ROUGE-N, Eq. 2.1, measures the overlap of N-grams between the generated summary and the reference one, where *gramN* indicates the contiguous sequences of N words. ROUGE-N is often used to evaluate the grammatical correctness and fluency of the generated text.

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{Reference Summaries}} \sum_{gramN \in S} \text{count\_match}(gramN)}{\sum_{S \in \text{Reference Summaries}} \sum_{gramN \in S} \text{count}(gramN)} \tag{2.1}$$

In contrast, ROUGE-L does not require a predefined *gramN* length because it identifies the longest common subsequence between the two summaries. It analyzes the content coverage and the semantic similarity of the generated text.

While ROUGE metrics offer valuable insights into the performance of TS techniques, it is also essential to consider its limitation: ROUGE metrics are evaluated by comparing human generated summaries that require human labor. Also, ROUGE metrics do not consider linguistic qualities or terminology variations as humans do. Finally, the sequential comparison of summaries is not accurate to the number of sentences and their order.

BERTscore: differs from ROUGE metrics because it uses the contextual embeddings of tokens to compute the cosine similarity, capturing the semantic context of words in sentences and providing a more accurate evaluation of text generation tasks. It takes into account recall, which measures the proportion of the reference summary that is covered by the generated summary; precision which evaluates how well the generated summary represents the content of the reference summary; and, F1 score which combines recall and precision, as shown in Eqs. 2.2, 2.3, and 2.4), where x is a reference summary, $\hat{x}$ is a candidate summary, and $x_i^T \hat{x}_j$ is a cosine similarity calculation [10].

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \tag{2.2}$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j \tag{2.3}$$

$$F1_{\text{BERT}} = \frac{2 \cdot P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{2.4}$$

CHAPTER 3

Related Work

This chapter begins by outlining the research methodology employed in this project, as detailed in Section 3.1. It then provides an overview of the research process, guided by the PRISMA 2020 framework, which enables us to collect a range of relevant works in the field, Section 3.2. Subsequently, Section 3.3 describes the phases and methodologies employed in the process of TS necessary to generate a summary from an input text or texts. This section is subdivided into four different sections each one representing a phase of the TS process:

(1) Preprocessing
(2) Feature Extraction and Text Representation
(3) Summarization Algorithms
(4) Evaluation and Results

Section 3.4 concludes this chapter with an overview of all techniques, algorithms and implementations that have been done in the field of text summarization.

3.1. Research Methodology

This study combines two techniques to provide a comprehensive knowledge of the state of the art in legal document summarization. The methodologies, SLR and PRISMA, are described in the following subsections.

3.1.1. Systematic Literature Review

An SLR is a meticulous research methodology that describes the main ideas of how a researcher should collect, select, and analyze all the available research studies, such as books, articles, and documents. Using the SLR methodology, it is possible to obtain a vast number of related research studies and documents about a specific topic [5].

The primary goal is to understand the current knowledge within the domain of the research by assessing and comparing various approaches that have already been undertaken. Simultaneously, it proves valuable in identifying unresolved challenges, providing opportunities for authors to explore these unaddressed aspects.

The SLR must always follow a well-defined protocol that specifies the criteria before starting the review process. According to Neiva and Silva [26], this process involves:

- setting research questions that will guide the author through the entire investigation process;
- defining keywords and search queries to capture relevant studies;

- defining the inclusion and exclusion criteria so that the results are accurate to the topic in discussion, such as the year of article publications, relevant keywords, document types, etc;
- evaluating the quality and validity of selected studies.

When the process is followed correctly and with the least amount of error, the study can produce accurate findings and a reliable conclusion with transparency and rigor, allowing researchers to make the best choices based on the evidence gathered.

3.1.2. Preferred Reporting Items For Systematic Reviews And Meta-Analyses

"The PRISMA statement, published in 2009, was designed to help systematic reviewers transparently report why the review was done, what the authors did, and what they found" [6]. Over the years, the methodologies employed in systematic reviews have improved significantly, leading to advancements and enhancements in review guidelines. This evolution has culminated in the current utilization of PRISMA 2020.

PRISMA 2020 includes an updated 27-item checklist and flow diagrams which assist researchers in identifying, selecting, evaluating, and synthesizing papers in order to answer research questions. It serves as a robust framework for conducting comprehensive and methodical systematic reviews.

PRISMA 2020 does not suffice as a singular tool for conducting systematic reviews – it is only a complementary methodology that brings value to the process of a systematic review by being advantageous during the planning and execution phases of systematic reviews, ensuring accurate information is captured.

3.2. Research Process

Conducting a SLR in this field is critical for understanding the literature landscape and the methods used in legal document summarization, as well as analyzing their limitations and effectiveness. The Scopus[1] repository was used as a knowledge base for the survey because it offers advanced search features and analytical tools to effectively refine and explore search results, which will help to include and/or exclude some criteria as required for one of the SLR steps. It is also known for having a vast database that includes a large number of peer-reviewed articles, which will allow us to obtain accurate and trustworthy studies.

Following a preliminary investigation aimed at comprehending the primary background of the subject, a set of keywords was selected in order to formulate a query that initiated the article selection process. In conclusion, "Legal Document" and "Summarization" were the research's important terms.

The initial approach employed for searching in Scopus involved creating the query as "Legal Document" AND "Summarization", resulting in an exhaustive collection of 721 documents. Considering the size of the database, an effort to refine and focus the search

[1]Scopus: `https://www.scopus.com/`

on more specific content led to the consolidation of the query into "Legal Document Summarization".

This adjustment in the search strategy resulted in a dataset that was significantly reduced to a more manageable set of 148 research articles. This approach made it easier to put together a focused collection that closely connects legal documents to the summarization process, allowing a more in-depth investigation of this particular field.

Making use of the analysis tools that Scopus has to offer, it was evident that there have been an increasing number of research studies since 2018, Figure 3.1. For this systematic review, articles published before the year 2018 were excluded. This decision was made to only obtain the most recent developments and advancements in the state of the art.

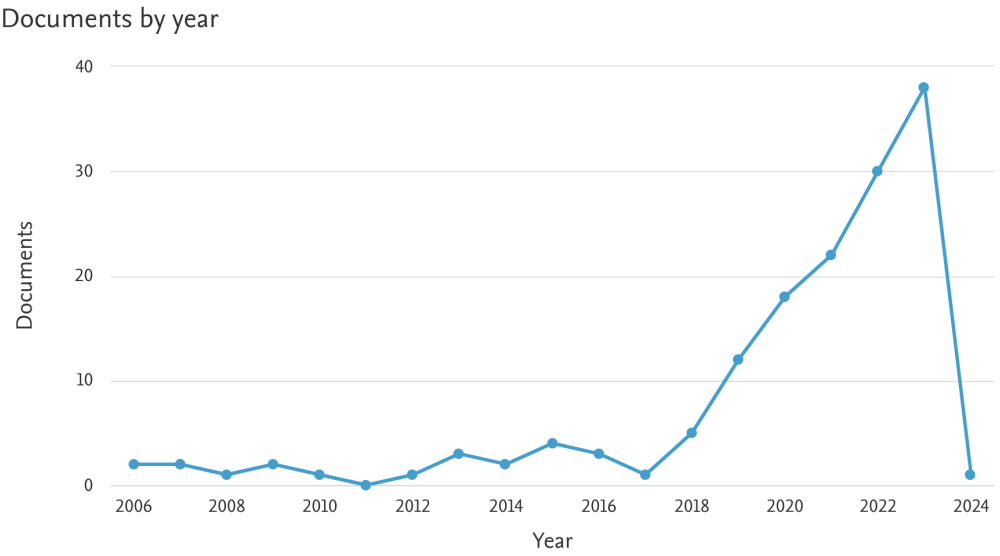


FIGURE 3.1. Distribution of documents by year using the query "Legal Document Summarization"

Additionally, two filters were applied:

- Papers belonging to the types of Conference review, Book, and Book chapter were not considered;
- Subjects inserted within the categories of Business, Management, and Accounting, science and natural resources (Agricultural and Biological Sciences, Chemical Engineering, Physics and Astronomy, Energy), and Medicine, were excluded since they did not satisfy the inclusion requirements.

The screening phase initially started with a set of 103 documents. As shown in the PRISMA flow diagram in Figure 3.2, this phase consists of three distinct stages ("Records screened", "Reports sought for retrieval", "Reports assessed for eligibility"), each designed for the deliberate exclusion of articles based on specific criteria, detailed below.

Step 1: Records were analyzed by assessing both the titles and abstracts of the documents. Out of the total, only 40 records were deemed relevant as they contained pertinent information related to the discussed topic. In Table 3.1, we can see that there are three distinct types of documents represented. Specifically, the

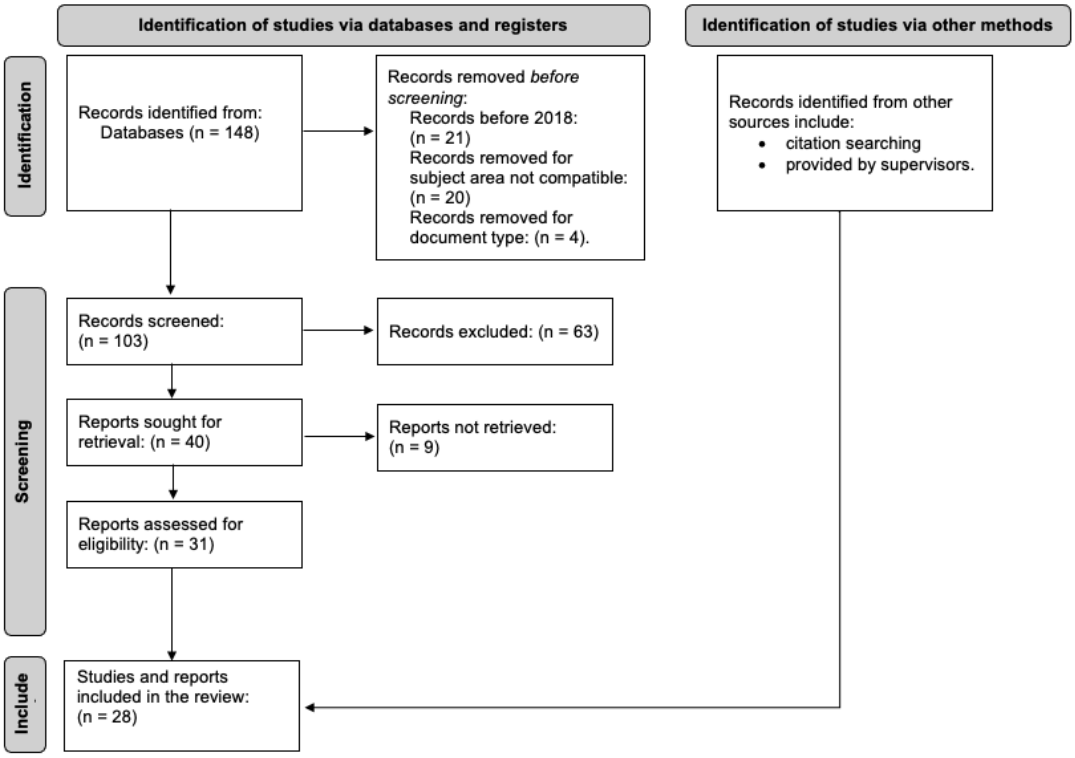


FIGURE 3.2. PRISMA 2020 flow diagram

data reveals that Conference Papers constitute the predominant document type within this dataset.

Document Type	Number of Documents
Conference Paper	24
Article	15
Review	1

TABLE 3.1. Total number of documents for each document type

Step 2: In this step nine of the documents were inaccessible and could not be used in the next phase.

Step 3: At the end of this process only 31 documents were used in the review process.

3.3. Text Summarization Process

This section outlines the steps involved in the TS Process. In order to achieve a final summary, the original documents go through several transformations. We delineate a representation of the sequence of these steps for a more straightforward interpretation (see Figure 3.3). More precisely, the different methods that are useful for creating clean and structured data are described in Section 3.3.1. Section 3.3.2 covers the techniques for extracting relevant features from the texts and ways to represent documents in order for TS models can understand them. Section 3.3.3 explains different ETS, ATS, and HTS models and how they were implemented in state-of-art works. Finally, Section 3.3.4 describes how the evaluation and results are delivered in different works.

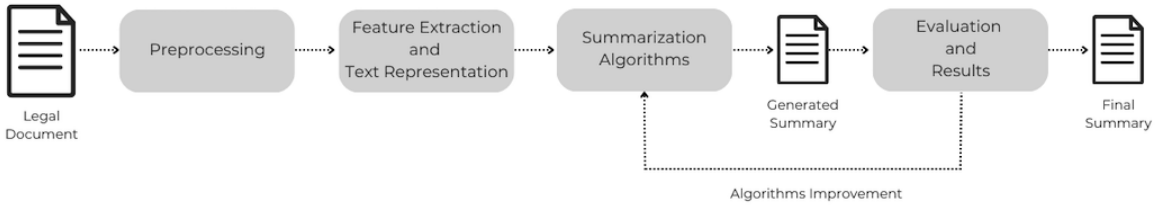


FIGURE 3.3. Summarization Process Representation

3.3.1. Preprocessing

In the field of legal document summarization, preprocessing is a crucial step that refines raw datasets to ensure their potential for subsequent tasks. The goal is to eliminate inconsistencies, noise, and unnecessary information, thereby enhancing the overall quality of the dataset and optimizing computational efficiency.

Multiple studies have emphasized the importance of preprocessing, employing similar methods to prepare datasets, such as the following:

Sentence Segmentation: In multiple works, the authors recognize the necessity for segmenting the text into individual sentences, which helps in structuring the text for further analysis [27], [28]. The most common technique for identifying and splitting a text into individual sentences is Punctuation-based segmentation. With this method, sentences are split based on punctuation marks such as periods (.), exclamation marks (!), and question marks (?) [29].

Noise Removal: Removes unnecessary text from the document, such as the header and footer, that does not add any special information to the model. Additionally, punctuation marks, special characters, and blank lines are also considered noise that must be eliminated. [12], [30], [31]. This practice contributes to minimising the computational cost and improves the quality of the summary.

Document Parsing: Splits the documents into manageable segments and smaller chunks, such as paragraphs or sections. The parsing task can be done by setting a limited number of tokens per chunk or by rule-based approaches. Other approaches consider the context and similarity between sentences before splitting. [27], [32].

Removing Stop-Words: In all languages, there are words present in text like prepositions, articles, and conjugation that do not offer much value in the context of the document. The objective of this technique is to identify and eliminate these words to reduce the input text. Additionally, depending on the summarization algorithm used, it helps to focus on the semantic content of the text [33], [34].

Stemming or Lemmatization: The main approach of these two methodologies is to normalise words to their original form. Stemming is a process that eliminates the prefixes and suffixes of words. Lemmantization, in contrast, considers the morphological syntax of the word and takes the root of it. The application of these strategies results in the creation of a more robust representation of the

documents which can be beneficial when implementing models where the word statistics are significant [35], [36].

The application of these common preprocessing methods indicates a collective agreement in the literature regarding essential foundational steps for successful legal document summarization. Researchers employ these techniques to create refined datasets that enable precise and efficient automatic summarization processes. While each technique offers distinct advantages, it is crucial to understand the potential implications they may have on the data when applying them.

3.3.2. Feature Extraction and Text Representation

Legal documents are characterized by being long texts with a unique language where specific concepts and technical terms are used. When generating a summary, particular care must be taken to ensure that the vocabulary is incorporated seamlessly and coherently while maintaining the context of the original documents. The complexity of legal language contributes to the substantial challenge of the summarization task.

To enhance the incorporation of relevant keywords and features in summaries, previous works have shown that incorporating methods capable of extracting relevant information can improve the task of summarization. Different techniques can be used in order to achieve that goal. For example, the term-frequency, which measures the frequency of each word in a document, helps identify important keywords. Characteristics of the sentences, such as the length or position in the document, are also possible features to consider when selecting relevant sentences that could be included in the final summary. Other works, compared the similarity between sentences which is also a strategy that has been adopted to verify which sentences are more informative [4], [34].

In order to effectively process and analyze documents, it is necessary to transform raw data into a form that models and algorithms can handle. Text representation is a crucial step in the TS process since it translates words into numbers, allowing a computer to understand the human language. Depending on the type of summary, different textual representation techniques can be used. Some methods are more simple and represent the frequency of words. Others go further and use embedding systems to represent words based on their context within the document.

One of the simplest methods used to represent words is Term Frequency-Inverse Document Frequency (TF-IDF). TF-IDF assigns a weight for each word based on its occurrence in a document [37].

$$\text{TF} = \frac{\text{number of times a term appears in the document}}{\text{total number of terms in the document}} \tag{3.1}$$

$$\text{IDF} = \log\left(\frac{\text{number of documents in the corpus}}{\text{number of documents in the corpus that contain the term}}\right) \tag{3.2}$$

$$\text{TF-IDF} = \text{TF} \times \text{IDF} \tag{3.3}$$

Although it can extract the keywords of a text, it neglects both the order and the semantics of sentences in a document.

To overcome the limitations of simple vector representations like TF-IDF, word embedding models have been developed. These models preserve syntactic and semantic relationships between words. Word2Vec and GloVe are two models that fit into this category. Word2Vec uses neural networks to learn word associations, predicting the context of each word in the document. In this method, a word can only have one representation regardless of its location in the document [38]. GloVe (Global Vectors for Word Representation) combined matrix factorization with context-based learning. This approach allows GloVe to capture both local and global statistical information, resulting in robust word vectors [39]. Several studies employ Word2Vec and GloVe for input representation and feature extraction, as evidenced in the work of Schraagen, Bex, Luijtgaarden, *et al.* [24], Anand and Wagh [28], and Rani and Lobiyal [33].

More recent models, like BERT and transformer-based embeddings, have been introduced as improved embedding systems, showing better performance in providing a more accurate representation of words to the TS models.

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a revolutionary language model created by Devlin, Chang, Lee, *et al.* [40]. BERT uses surrounding text to establish context in order to understand the meaning of natural language. The BERT model is pretrained using a masked language model task, which randomly masks tokens in a sentence and then predicts the original terms based on the context. The use of transformers significantly increases the capacity of BERT to understand context and ambiguity in language.

Variations of BERT have been explored by others and applied to this task: Jain, Borah, and Biswas [41] use Legal BERT [42] to obtain the individual sentence representation of a document; Sun, Yang, Wang, *et al.* [43] use BERTSUM for sentence-level encoding and obtaining sentence representations for documents, which allow modeling the relationship between sentences and summaries. Models like T5 (Text-to-Text Transfer Transformer) and GPT (Generative Pre-trained Transformer) represent the new generation of text representation. T5 is based on encoder-decoder layers. The encoder section is the one associated with the input representation. It starts by reading the input text and transforming it into a high-dimensional representation. T5 is able to capture the meaning of each individual word and the relationships between the words in the context of the entire input sequence [44]. GPT also uses a transformer-based architecture to generate context-aware embeddings. In this method, words are represented in smaller units called tokens, representing a common sequence of characters, which are then accessed through multiple layers of transformers. GPT adopted a self-attention mechanism and feed-forward neural

networks to enable the model to focus on different parts of the documents at different stages of the process [45].

The way we represent vectorized documents and the embedding model used in the different tasks is also very relevant and can greatly influence the results. In the context of summarizing documents in Portuguese, Souza, Nogueira, and Alencar Lotufo [46] created BERTimbau, an adaptation of BERT designed specifically for the Brazilian Portuguese language. BERTimbau has demonstrated effectiveness in large, pretrained language models for Portuguese.

3.3.3. Summarization Algorithms and Techniques

There are a variety of algorithms and techniques that can be used to perform large document summarization. The primary purpose of this subsection is to identify those that better adjust to the goal of this work and that can achieve better results. As described in Section 2.2 there are two main methods to generate summaries: extractive and abstractive approaches. In this section, we discuss different proposed models and techniques implemented in other works for both approaches. Additionally, we also give some examples of how hybrid approaches were developed using both ETS and ATS methods.

3.3.3.1. *Extractive Text Summarization Approaches*

The idea behind extractive algorithms is to generate a summary by concatenating sentences extracted from the original document. It is important that these sentences contain keywords that are easily associated with the main topic and that they can express the overall meaning of the input text. In this section, we presented several examples of how these algorithms have been used in the field of TS and in the context of the legal domain.

Some techniques include graph-based algorithms, where graphs are used to represent the sentences of a document. The input text is represented as a network, where each node is a sentence. With sentence-similarity algorithms, the relationship between each node is calculated, and a link between similar nodes is established. Finally, with a configured network, the sentences are ranked based on specific criteria, and the top sentences are selected to be included in the summary. TextRank [47] and LexRank [7] are examples of graph-based algorithms. Improvements to these algorithms have been performed by Jain, Borah, and Biswas [31], which implemented a Bayesian Optimization-based strategy to optimize the TextRank algorithm. Their work revealed the effectiveness of improving the TextRank algorithm through hyperparameter tuning.

Other authors developed cluster-based algorithms. In this type of algorithm, similar sentences are grouped into different clusters, and then the most relevant sentence from each cluster is selected to be included in the summary. This approach ensures that each sentence covers different topics in the document and that the output summary is less redundant. Rani and Lobiyal [33] applied the K-means algorithm to partition semantically closer sentences into the same cluster. They feed average sentence embedding to the algorithm to capture the semantic dissimilarity between sentences. The final summary is the result of the clustering algorithm that returns the clusters formed based on the

semantic closeness of the sentences. DCESumm is a sentence scoring approach that uses LEGAL BERT [42] to find the summary relevance scores where the scores are improved via a clustering approach. In the end, if a sentence has a higher level of relevance and also belongs to a cluster that contains other relevant sentences, it has a higher probability of being chosen for the summary [41].

There are also other techniques using DL approaches; one of the most common these days is BERT. BERT has become a foundational model in TS because of its bidirectional nature for context understanding. Researchers have adopted the BERT approach to their projects in a variety of ways. The main goal of Klaus, Hecke, Naini, *et al.* [48] was to understand the essence of legal texts relayed on a BERT architecture. This was accomplished through tasks involving sentence representation and classification, aiming to understand the relevance of individual sentences in the texts. Different variations of BERT have emerged to overcome some of its limitations. For example, Sun, Yang, Wang, *et al.* [43] propose a new method called BERTSLCA based on the BERTSUM model, a variant of BERT. BERTSLCA was developed to address complexity and training time issues, as well as the neglect of sentence interrelationships in multisentence documents in the BERTSUM model.

A very specific algorithm developed particularly for the legal document domain is the DELSumm (Domain-adaptive Extractive Legal Summarizer), an innovative unsupervised extractive summarization algorithm developed by Bhattacharya, Poddar, Rudra, *et al.* [49]. Unlike many existing algorithms, DELSumm considers rhetorical segments within legal case documents and subsequently identifies which parts of the segments to include in the summary, according to the guidelines set forth by legal experts. DELSumm stands out due to the significant limitations in existent methods to incorporate domain knowledge that specifies the vital information that should be present in the output. Some improvements to DELSumm algorithm were made by incorporating document-specific catchphrases [50]. These catchphrases are not only legal domain-specific terms but also terms or phrases that have document-specific importance.

The majority of the ETS models presented here used sentence-level methods where they extracted sentences based on certain criteria to generate the final summary. However, these algorithms tend to select highly generalized sentences. In contrast, approaches based on summary-level methods have yielded significant results and improved the quality of extractive models. In these types of methods, the top sentences of a document are extracted and combined to generate a range of different candidate summaries. Subsequently, by using a text-matching model the best summary is chosen. For example, MatchSum is a model where the similarity between all generated candidates and the original text is calculated. The candidate with the highest score is selected for the final summary [51]. A more recent improvement, SeburSum, introduces a contrastive learning framework to train the model. Also instead of using the original text they just compare the candidates between them instead, which improved the computational performance of the model [52].

3.3.3.2. *Abstractive Text Summarization Approaches*

ATS approaches have gained considerable attention in recent days. Being able to generate a summary similar to a human-written summary has intrigued researchers. To create an abstractive summary, it is necessary to have a language model that understands and rewrites actual text from scratch. In this subsection, we describe models capable of formulating texts and how researchers implemented ATS methods into their works.

The development of Seq2Seq models enabled the development of modern ATS methods. The Seq2Seq model was introduced by Sutskever, Vinyals, and Le [53]. This model is known for having an encoder-decoder architecture where it can process an input sequence and transform it into a different output. By leveraging the power of neural networks, Seq2Seq models are a fundamental framework for NLP tasks, including TS.

Long Short-Term Memory (LSTM) networks are commonly used to implement Seq2Seq models, addressing the challenge of capturing long-term dependencies in sequences. LSTM has a special memory unit that allows it to learn long-term dependencies in sequential data, which enables the model to understand complex relationships within the text, leading to a more accurate and coherent summary. In [24] sentences are generated by using a bi-directional LSTM encoder that transforms the sentence representation into a contextual representation and a LSTM decoder that computes the extraction probability of a sentence based on the contextual embedding.

The introduction of an attention mechanism was motivated by the need for models to focus on specific parts of the input texts when generating the output. This innovation later became a foundational component of Transformer models. By using self-attention, Transformers models can selectively pay attention to the most informative segments without losing important information. Different approaches were designed with the aim of text generation. For example, BART is a bidirectional and auto-regressive transformer proposed by Lewis, Liu, Goyal, *et al.* [54]. It implements a bidirectional encoder, like BERT, and a left-to-right decoder, like GPT. It has proven to be a remarkable model for text generation, especially when fine-tuned for that specific task. The authors from [24] implemented the BART model which showed good performance in rewriting and shortening sentences.

T5 is another framework that is showing significant results for this type of task. Developed by Google researchers, T5 is a large-scale transformer-based language model that has achieved state-of-the-art results on various NLP tasks, including TS. T5 is an encoder-decoder model that is pre-trained on a mixture of unsupervised and supervised tasks in a multi-task setting. T5 converts each task into a text-to-text format, enabling to work on a variety of tasks out of the box. As the model is pre-trained on a mixture of unsupervised and supervised tasks, it has the potential to generalize well to new tasks. By providing the text to be summarized with the prefix "summarize:", T5 can generate a concise summary that captures the essence of the original document [55].

Recent advances in large language models, such as GPT-3 and GPT-4, have demonstrated superior ability in understanding and generating coherent texts. Their capacity to capture complex relationships between sentences and texts renders GPT models relevant to the task of summarization. These models are trained on a diverse range of datasets comprising multiple languages, which offers an advantage when dealing with non-English texts, such as Portuguese ones (see, for example, the work of Zhao, Wang, Abid, *et al.* [56]).

In the case of sensitive documents that contain personal information, like Portuguese legal documents, it is essential to consider how summarization models manage the data provided as input. For instance, models such as GPT retain all the information provided, which represents a risk if a document that contains sensitive data is provided to these models. In the case of Portuguese Supreme Court of Justice judgments, it is necessary to create summaries based on judgments that have not been anonymized, as the judge may require the use of sensitive data to create a well-founded and accurate summary. While GPT models have achieved state-of-the-art results in the NLP field, it is essential to consider the risk of violating data privacy. Another factor to take into consideration when implementing transformer-based models, especially in the context of legal documents, is the number of tokens a model can process. When dealing with lengthy documents, like legal documents, this limitation is of significant concern in a way that challenges the model's ability to capture the full context of the document.

3.3.3.3. *Hybrid Text Summarization Approaches*

Hybrid summarization approaches can offer different solutions to the limitations of the ETS and ATS models. By combining both strategies, hybrid models can leverage the strengths of each approach and improve performance in summarizing documents. Especially when dealing with texts like legal documents that are long documents with very specific domain language.

Huang, Sun, Han, *et al.* [32] adopted a hybrid approach where they first annotated which sentences were more relevant to be included in the final summary using BERT and LSTM models and added a BiLSTM attention mechanism to identify relevant keywords in the document. Finally, to generate the abstractive summary, they use the selected sentences and extracted keywords as input for the UNILM based model [57]. This approach was fundamental to capturing important and relevant information from lengthy legal texts.

In order to handle the lengthy nature of the document and the limited available training data, Jain, Borah, and Biswas [15] also propose a hybrid approach. Their methods involve creating different summaries with a maximum number of tokens from the original document; for each extractive summary, they use a different extractive algorithm. This new set of summaries is then used to build a new training set for fine-tuning a BART model. This approach handled the limited amount of data. Finally, the fine-tuned model is used with test documents, where each document is divided into chunks of a limited number

of tokens so it can generate the summary of each one and concatenate each generated summary into the final one. By limiting each text to an exact number of tokens, the texts could be processed by the BART model, overcoming the lengthy document limitation.

3.3.4. Evaluation and Results

This section presents how different authors evaluate the performance of their models and which parameters are crucial to take into consideration to assess the quality of the generated summaries.

For the evaluation process, there are two main techniques that are used to evaluate the performance of the models. The most common ones are the automated metrics, which we described in Section 2.3. These metrics are employed by comparing the generated summaries with reference summaries. Others still prefer human evaluation techniques since existing metrics do not lead to a perfect evaluation in terms of the coherency and fluency of a generated text. Human evaluators can provide qualitative insights that complete the quantitative automated metrics [11], [24].

In order to evaluate the performance of TS models authors explore a selection of parameters, such as the length of the summaries, relevance, and redundancy.

Length of the summaries: The length of the generated summaries is set to different values in order to find the optimal length for them.

Relevance: Measures how much information is retained from the original document.

Redundancy: Checks if the information contained in the summary covers different topics of the original document or if it is repeated in the summaries.

After obtaining the initial results from the implemented baseline models and conducting a rigorous analysis, it is essential to make iterative improvements. By progressively incorporating more sophisticated methodologies and techniques, the goal is to enhance the model's performance. Each iteration involves evaluating the improved models, analyzing the results, and identifying their limitations for further improvement. This continuous cycle of evaluation and upgrading aims to achieve increasingly better summarization results over time.

3.4. Overview

This section aims to provide an overview of the work done in the field of TS based on our systematic review. This review covers the four key tasks of the TS process.

The preprocessing phase involves techniques such as tokenization, noise removal, and document or sentence segmentation, which proved to be essential for preparing the documents for further processing. Some studies emphasize the importance of this step in maximizing the performance of summarization models and algorithms.

In the text representation and word extraction phase, simple methods like TF-IDF or Word2Vec were commonly used. Additionally, advanced embedding models and transformers have shown to be capable of retaining more informative representations of the

relationship between words. These methodologies facilitate getting a structured representation of the text that computers can understand.

Regarding the summarization algorithms, there were two main approaches: ETS and ATS. ETS methods proved to be more simple and do not require that much computational power, however, they lack the coherency and fluency of the generated summaries. ATS algorithms tend to be more complex, but they can provide more accurate summaries with a better representation of the original document. ATS algorithms still have many limitations that need to be faced, especially when dealing with long documents.

Finally, the evaluation task is still a step in progress since human evaluation is still needed. Existing metrics like ROUGE and BERTscore can measure the similarity between the generated and reference summaries; however, it is still difficult to accurately evaluate some qualitative aspects of the summaries.

Despite the advancements in the field of TS, the complexity of the legal domain adds a significant challenge. Dealing with lengthy documents with particular language and structure requires the development of personalized algorithms and techniques to ensure accurate summaries.

This investigation allowed us to gain a more contextualized knowledge of the work done so far, providing valuable insights into the diversity of models that we could use to investigate algorithms and metrics with better performance on our datasets.

CHAPTER 4

Dataset

Datasets play a significant role in the improvement of TS models. Access to a representative dataset is crucial to evaluate the performance of the algorithms and other methods and to identify improvements that can be made to them.

One of the motivations for this study was the availability of a set of legal documents from the Portuguese Supreme Court of Justice. These documents provide the opportunity to explore and improve the automation of summary generation for Portuguese legal documents.

A legal document from the Portuguese Supreme Court of Justice is typically organized into three top level, distinct, sections:

- Report (*relatório* in Portuguese), which outlines the main topics of the judgment, identifies the different parties involved, and specifies the decision to be made;
- Grounds (*fundamentação* in Portuguese), where a first subsection describes the facts of the case (*matéria de facto* in Portuguese) and a second part that integrates all the details that have been addressed so far that could be taken into consideration to make the final decision (*fundamentação de direito* in Portuguese);
- Decision (*decisão* in Portuguese) that corresponds to the final decision of the judges.

However, in some legal documents, these sections are not directly identified, which limits their comprehension and creates a challenge in identifying the placement of each sentence within the document structure.

Our data is composed of 5000 legal documents from the Portuguese Supreme Court of Justice, representing the "Cível" (*Civil* in English), "Criminal" (*Criminal* in English), and "Social" (*Social* in English) areas. Additionally, for each document, an official judge provided a written summary. These summaries serve as reference summaries in our study, allowing for a comparison with the automatically generated summary, facilitating the evaluation process.

4.1. Dataset Preparation and Creation

Data preparation is a fundamental step in the development of TS models, guaranteeing that the data is clean and normalized to be prepared for subsequent phases.

In the preprocessing phase, we start by cleaning and structuring each legal document and reference summary. First, we remove all HTML tags by using "BeautifulSoup"[1].

[1] `https://pypi.org/project/beautifulsoup4/`

Secondly, we divide each text into sentences. Additionally, we removed all lines that were empty or that contained only numbers or punctuation marks.

Two datasets were created, one with the original 5000 legal documents and their respective reference summaries, referred to as "Dataset 1", and a second dataset, where instead of using the original documents, we only use the parts of the documents extracted from the sections *"Fundamentação de direito"* and *"Relatório"*, which we call "Dataset 2".

"Dataset 2" was created with the aim of investigating whether certain sections of Portuguese legal documents were more relevant to summarization than using the entire document. The sections *"Fundamentação de direito"* and *"Relatório"* were chosen since they were the ones that offer the more important facts and foundations of the case. In order to identify these two sections from the legal documents, we execute the segmenting model created by Zanatti, Ribeiro, and Pinto [58]. This Segmenting Judgment Model employs a Bidirectional LSTM with a Conditional Random Field (CRF) model that is able to recognize and assign section names specific to judgments. Given that the original documents do not always contain a uniform structure, it was only possible to identify 3552 documents that contain the sections *"Fundamentação de direito"* and *"Relatório"*.

4.2. Dataset Analysis

In this section, we analyze the statistical aspects and lexical properties of both datasets. A deeper understanding of the characteristics of the legal documents allowed us to adapt our models to specific features of the data. We further explore the average number of sentences and the average length of tokens in the documents and in the reference summaries. Additionally, we analyze the most frequent words in the legal documents and their respective summaries.

TABLE 4.1. Statistical properties of legal documents for each area and section

	Areas			**Sections**	
Statistical Property	***Cível***	***Criminal***	***Social***	***Fundamentação***	***Relatório***
# of Documents	2862	1444	694	3393	159
Avg. tokens/doc	8136.00	13009.21	9857.20	2985.73	4585.47
Avg. sent/doc	141.21	209.41	174.83	43.60	73.10
Avg. tokens/sentences	58.18	61.24	56.70	66.93	61.78
Avg. tokens/sum	300.96	692.22	329.88	304.76	236.74
Avg. sent/sum	5.94	9.47	5.69	6.05	5.52
Avg. tokens/sentences	53.31	75.77	59.69	53.38	45.69

Table 4.1 summarizes the statistical properties of our datasets for the different areas and sections. From them, we could conclude:

- The documents from the "Criminal" area have the highest average number of tokens and sentences per document, indicating this type of document can be more detailed in the description of the judgment.

- "Cível" and "Social" documents are similar, with fewer tokens and sentences on average when compared to "Criminal" documents.
- The documents that only contain the sections "Fundamentação de direito" ("Fundamentação" in Table 4.1) and "Relatório" show a big discrepancy in the document length, with "Fundamentação de direito" having a significantly lower average number of tokens per document compared to "Relatório".
- With the exception of "Criminal" area summaries, all properties are consistent across the different areas and sections, indicating a similar level of complexity. This suggests that the generated summaries should range between 230 and 350 tokens, typically consisting of five to six sentences.

In order to verify the most frequent words in legal documents and their reference summaries we employ the "pt_core_news_sm" model from Spacy[2]. During the tokenization process, we did not consider stopwords, punctuation, and digits. Furthermore, we also perform lemmatization to enhance the accuracy of the extraction. Figure 4.1 shows the ten most frequent words that appear in the 5000 legal documents, while Figure 4.2 represents the ten top words in the reference summaries. In both figures, the words are presented in their original Portuguese form and with an English translation.

The words represented in the figures indicate a strong focus on the legal field, suggesting that the documents discuss factual details and rights. The analysis of the most frequent terms is useful to evaluate the quality of the generated summaries since these terms are important in the context of legal documents. The word "article" has a big influence in the summaries showing more than 10000 appearances, reflecting that the content of the summaries is very related to the articles present in the documents. Additionally,an equilibrium is visible between the words present in the documents and in the summary, where only four words: "Court", "Judgement", "Crime" and "Contract" do not appear in both of the graphs.

This investigation helps to understand the nature of the datasets and guides the design of our summarization models, allowing us to better handle the specific characteristics of each document type.

[2] `https://spacy.io/models/pt`

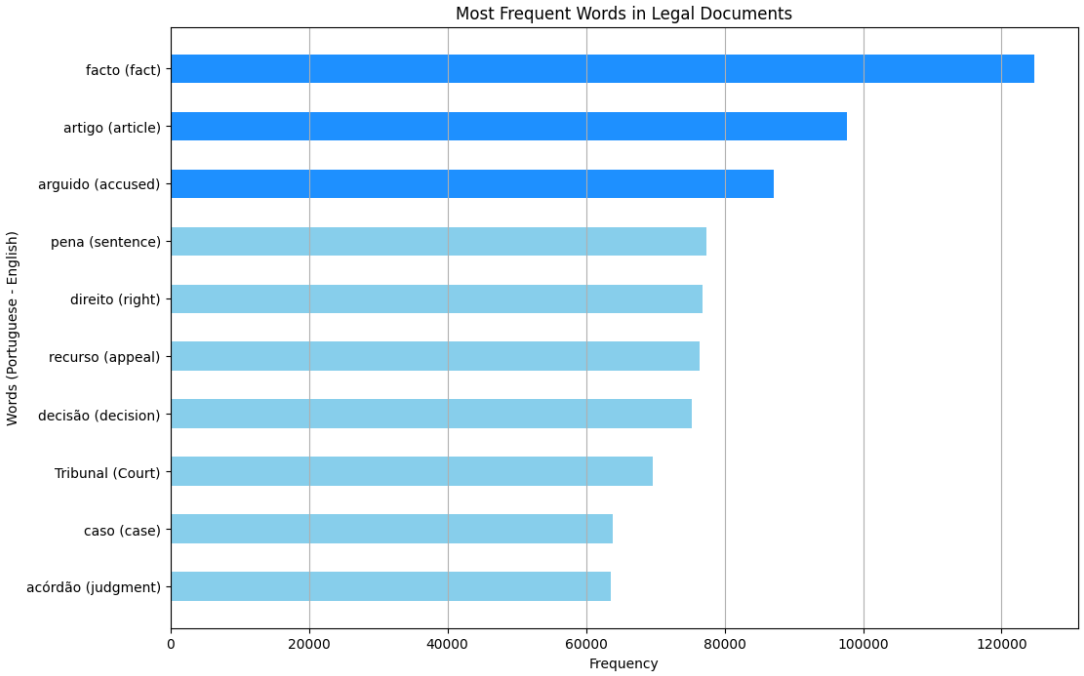


FIGURE 4.1. Ten most frequent words in legal documents

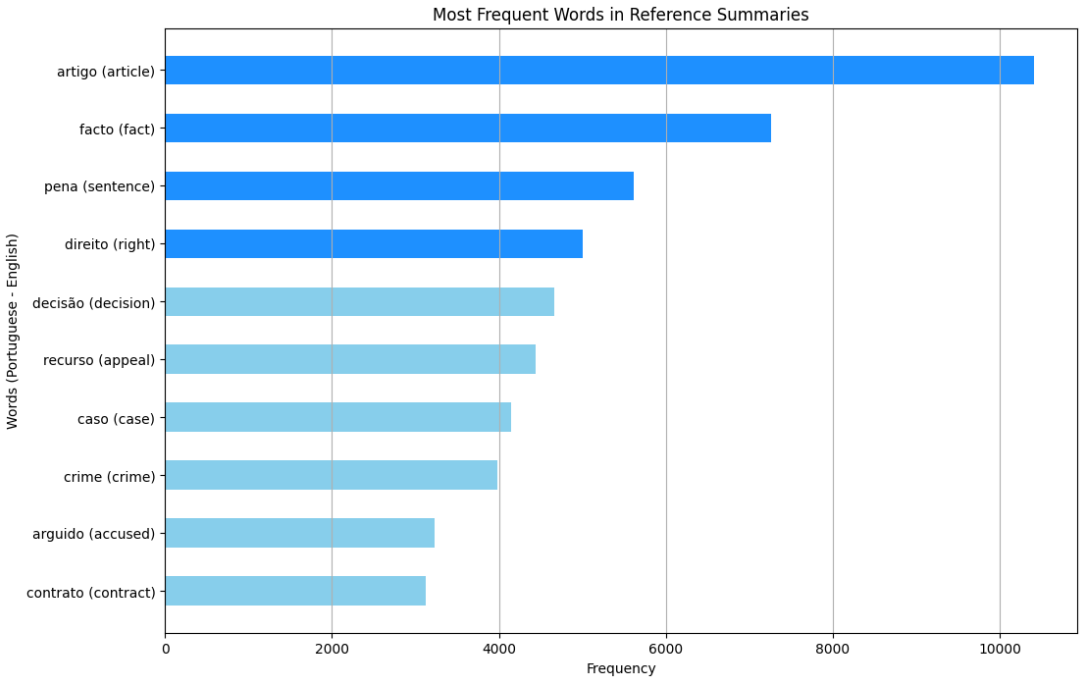


FIGURE 4.2. Ten most frequent words in legal reference summaries

CHAPTER 5

Summarization Approaches

In this chapter, we describe in detail the models and algorithms we implemented in order to assess the performance of summarization. The approaches conducted in this work have the primary goal of understanding the effectiveness of generating a summary from a Portuguese legal judgment. In this chapter we delve into three different approaches where we describe each of them and why they were chosen to be incorporated into this work. They are a sentence-level approach, a summary-level approach and a hybrid implementation that includes a combination of the LexRank algorithm and the MBART model.

5.1. Sentence-level Approach

In order to implement this sentence-level approach, the LexRank algorithm was selected. LexRank is an unsupervised ETS algorithm. It was chosen for this study due to its effectiveness in capturing the most relevant sentences within a document. LexRank simplicity makes it a solid algorithm to serve as a starting point for this research. The goal was to have LexRank as a baseline model that can be used for comparing the performance of more complex implementations. Additionally, LexRank also serves as an extracting algorithm in the other implementations.

LexRank is a graph-based algorithm that computes the similarity between sentences and uses graph centrality to determine the most important sentences in a document. The process begins by representing each sentence with a TF-IDF vector in order to calculate cosine similarity between all pairs of sentences. Then, the base of the graph is constructed, where the sentences represent nodes and the edges the similarity scores between the sentences. Finally, the centrality of each sentence in the document is calculated, and the sentences are ranked based on the score. The higher the score, the more significant and representative the sentence is in the document. The summary is then generated with the sentences that obtained the highest centrality scores. The LexRank algorithm was designed with the possibility of selecting the number of sentences that can be included in the generated summary and adjusting the threshold parameter. The threshold parameter gives the versatility to control the degree of similarity required for sentences to be linked. Setting a higher threshold results in fewer sentences being connected, since it is necessary for sentences to have higher similarity. Conversely, lowering the threshold leads to more sentences being connected.

For this sentence-level implementation, we applied the classical LexRank algorithm [1]. This approach requires a set of documents that will serve as the primary source of information for the algorithm. By providing an initial set of documents related to the ones to be summarized, the classical LexRank algorithm can identify the key topics and themes by analyzing the relationships between words and sentences across the documents. This will give the algorithm a deep understanding of the context of the documents, which will help rank the sentences accurately by importance. For this first step, we divide both datasets, where 80% of the original judgments were the initial set of documents and the other 20% were the ones that were summarized.

Secondly, we decided to set a fixed size for the generated summaries. By setting a fixed size, it was possible to evaluate the effects of different summary lengths on the generated summaries. We got results for sizes ranging from three to ten for "Dataset 1" and from three to seven sentences in "Dataset 2". These values were chosen on the basis of the mean size of the reference summaries, as seen in Table 4.1. Additionally, the threshold parameter was set to 0.1 in order to eliminate connections between sentences that have minimal common ground.

5.2. Summary-level Approach

Implementing an ETS approach suggests that the generation of a summary relies on the extraction of the most significant sentences from the document. However, when applying this method, the selected sentences tend to be very general.

Summary-level approaches can be a way to deal with this problem. In this type of approach, different summaries are generated. Instead of only generating a single summary with a set of top relevant sentences, as represented in Figure 5.1, this approach creates different combinations of candidate summaries. Figure 5.2 shows an example of the generation of a set of candidate summaries with the top three sentences, extracted from the legal document, where the length of the candidate summaries could vary between two to three sentences. In the final step all four possible summary combinations are visible according to the detailed specifications.

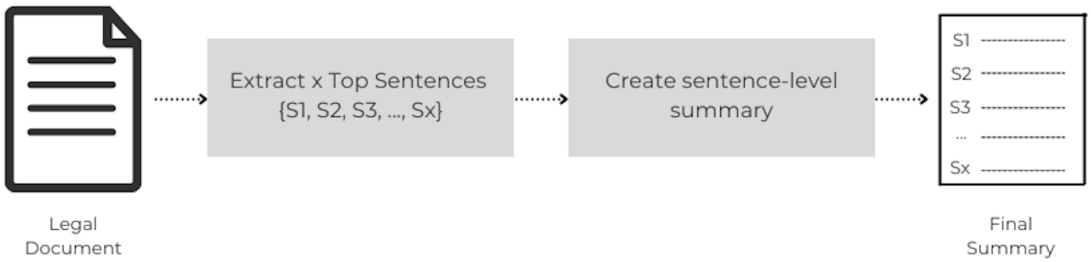


FIGURE 5.1. Example of a sentence-level summary creation

[1] https://github.com/crabcamp/lexrank

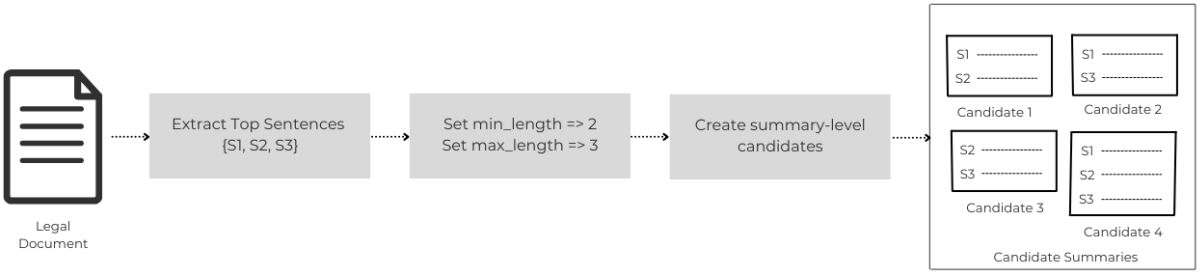


FIGURE 5.2. Example of summary-level candidate summaries creation with extracted_sentences=3, minimum_tokens=2 and maximum_tokens=3

The final summary in summary-level approaches can be one composed with sentences that are not made of only the best sentences but with the sentences that better capture relevant information without losing the context of the document.

Different techniques can be applied to choose the best summary among all the candidate summaries. Comparing the candidates with the original document and determining which candidate is more similar to it or comparing all candidate summaries between them and verifying which one has the highest similarity are two different approaches that can be used.

For this experiment, we followed some techniques from the SeburSum model implementation. The SeburSum is an extractive summary-level approach that showed good results and improved the quality of extractive models. SeburSum distinguishes itself by only comparing candidate summaries that do not have sentences in common instead of comparing the candidate summaries with the original document or between all candidate summaries. Our approach is to follow this same process of comparing candidates since it helped to overcome some limitations on summary-level approaches, such as: reducing the time of computation because it was not necessary to compare the candidates with the original documents, it also helped reducing the tendency of the models to select longer candidates and avoid redundancy by not comparing candidates with equal sentences.

Our summary-level approach consists of the subsequent steps to generate a summary from a document, D:

(1) We start ordering the sentences in D by relevance. To rank the sentences, we use classical LexRank.

(2) The next step involved creating a set of candidate summaries, $C = (C_1, C_2, C_3, \ldots, C_m)$. To create the candidates, it was necessary to select the top N sentences from D and the minimum, min, and maximum, max, number of sentences a candidate could comprise. The candidate summaries are then created by generating all possible combinations of sentences for every size between the min and max, inclusive. This means generating combinations of all possible sizes from the set of selected top N sentences.

(3) For each candidate, it is necessary to create embedding for the full text. By creating an embedding of the full text, it is possible to capture the relation between

sentences and words across the entire text, leading to a better understanding of the overall context of the candidate summary. It also simplifies the comparison between the candidates since it only needs to compare two single vectors. In order to compare the candidate summaries, we calculate the cosine similarity between the embeddings of the summaries, which were generated BERTimbau embeddings. It is important to acknowledge that BERTimbau was primarily trained on Brazilian Portuguese datasets. Despite this limitations, BERTimbau has demonstrated strong performance in Portuguese NLP tasks, making a sutable choice for this study and ensuring consistency. Additonally, BERTimbau computational requirements are relatively modest compared to larger transformers models. Its efficiency enables it to be used locally, even on machines with limited processing power, making it the most practical option given the computational constraints of this study.

(4) This next step involves computing the similarity scores for all the candidates that do not contain equal sentences in them.

(5) Finally, the selection of the candidate summary that achieved the best score can be executed.

For this experiment, we wanted to test if the method selected for choosing the best candidate had an influence on the performance of the model. Usually, the selected candidate summary is one that achieved the highest similarity score. When following this approach, all the other candidates that were compared will not have any influence on the decision of the best summary. We delineate three different methods to select the best candidate:

Max Score: In this method, we use the max version of the SeburSum. For each candidate summary, the best similarity score is saved among all the candidate summaries that were compared. Finally, the selected summary is the one with the highest similarity scores among all candidates.

Mean Scores: In this method, we compute the mean of the similarity scores obtained from comparing each candidate summary with the respective mutually exclusive candidate summaries. The best summary will be the one with the highest mean of all the candidate summaries.

Similarity score greater than x%: For this method, we count how many mutually exclusive candidates obtain a similarity score greater than x%, where x is a value between zero and 100. The candidate with the highest number of candidates with a similarity score greater than x is selected to be the final generated summary.

Additionally, when multiple candidate summaries were identified as optimal due to achieving the highest results, the selected summary was the one containing the most relevant sentences defined earlier by the LexRank algorithm.

For these implementations, the parameters for the number of extracted sentences (N) and the minimum (min) and maximum (max) length for each dataset are represented in Table 5.1.

TABLE 5.1. Summary-level parameters

Dataset	N extracted sentences	min summary length	max summary length
Dataset 1	12	4	6
Dataset 2	10	3	5

5.3. Hybrid Approach

The main goal of this experiment is to contribute new ideas and approaches that could help automate the process of summarizing legal documents. When writing a summary for this specific type of document, it is necessary to take into consideration the entire context of the document as well as the key terms that are crucial in the field. Besides that, it is also important to create a summary that is coherent, fluent, and comprehensible to both legal professionals and the general public.

Generating a summary using an ETS approach results in a set of sentences that are concatenated without any type of connector between them. Consequently, when reading a text of this nature, it will be more difficult to comprehend the information due to the lack of fluency and coherence. This limitation of ETS can be overcome by using ATS approaches. ATS models can generate logical and coherent texts. By giving an input, such as sentences, texts, or even questions, these models learn how to generate a new text from the context of the input, almost mimicking how humans write.

To evaluate the performance of an abstractive approach, we implemented the MBART (Multilingual BART) model to generate summaries. MBART is a transformer-based designed for different NLP tasks, such as translation, summarization, and text generation. MBART is based on the BART model, which features a Seq2Seq architecture with a bidirectional encoder and an unidirectional decoder. This specification allows the model to understand the input text and generate accurate outputs.

BART is pre-trained on a corpus of English texts, while MBART leverages the BART architecture by handling multilingual texts, including Portuguese. This extension makes MBART a suitable model for our task of summarizing Portuguese legal judgments. When configuring MBART, it is possible to define the target language and set the maximum length of generated summaries, providing flexibility and control over the output. Additionally, MBART is an open-source model, which ensures accessibility and flexibility for adjusting it to specific tasks. Another important consideration is its computational efficiency: MBART's requirements are manageable, enabling us to use the model in our local machines. This is particular important, as it allow us to avoid more resource-intensive transformers, ensuring that the computational demands of the study remain within reasonable limits. Furthermore, MBART allow us to use the model without concerns about

compromising data confidentiality, as it does not store or retain any used data. All this characteristics of MBART highly suitable for the objective of this dissertation.

ATS methods are relatively recent, they have shown significant progress with the advancement of DL and transformer-based models. Therefore, there are inherent limitations when implementing them. In the context of this work, the length of the legal judgments presented a challenge. When giving a text that contains more than 1024 tokens as input for the MBART model, it can not process due to its limit of tokens. One possible solution for reducing the texts would be to truncate the input texts to the MBART token limit. However, this would result in the first sentences of the texts being the only ones to be processed by the MBART model. Consequently, it can not be guaranteed that the first sentences were the most relevant ones to be present in the final summary. To address this issue, we implemented a hybrid approach where we start by using an ETS method to obtain a set of the most relevant sentences in the documents that do not exceed the token limit of MBART model, and a second phase where we use the set of sentences as input for MBART model. By applying this strategy, we could guarantee that we provide the highest amount of content in the MBART model, so it could generate a more informative summary.

The hybrid implementation was designed with an architecture comprising two distinct phases. In the extractive phase, we start by ranking the sentences from each document by relevance using the base LexRank algorithm. Secondly, the sentences that achieved the highest scores were selected until the sum of tokens in all sentences reached the limit of 1024 tokens. To count the number of tokens present in each sentence, we encoded each sentence with 'MBART50TokenizerFast"[2]. Finally, the selected sentences were grouped together to create a new input text. The sentences were rearranged in the same order as they appeared in the original document. For the second phase, the new input text is passed to the MBART model to generate the final summary. We configured the target language to Portuguese and the "max_tokens" parameter to ensure that the generated summary will be shorter than the input text. Figure 5.3 illustrates the hybrid implementation in a schematic format.

For this implementation, we applied two different values for the "max_tokens" parameters for each area and section. In the first experiment, we set the "max_tokens" to 600 tokens for every summary, in order to have a more general evaluation of the performance of the model. And a second experiment where we set the "max_tokens" based on the mean number of tokens per document. For the "Criminal" area the "max_tokens" was set to 700 tokens and for the other areas and sections was set to 350 tokens.

[2]https://huggingface.co/docs/transformers/model_doc/mBART

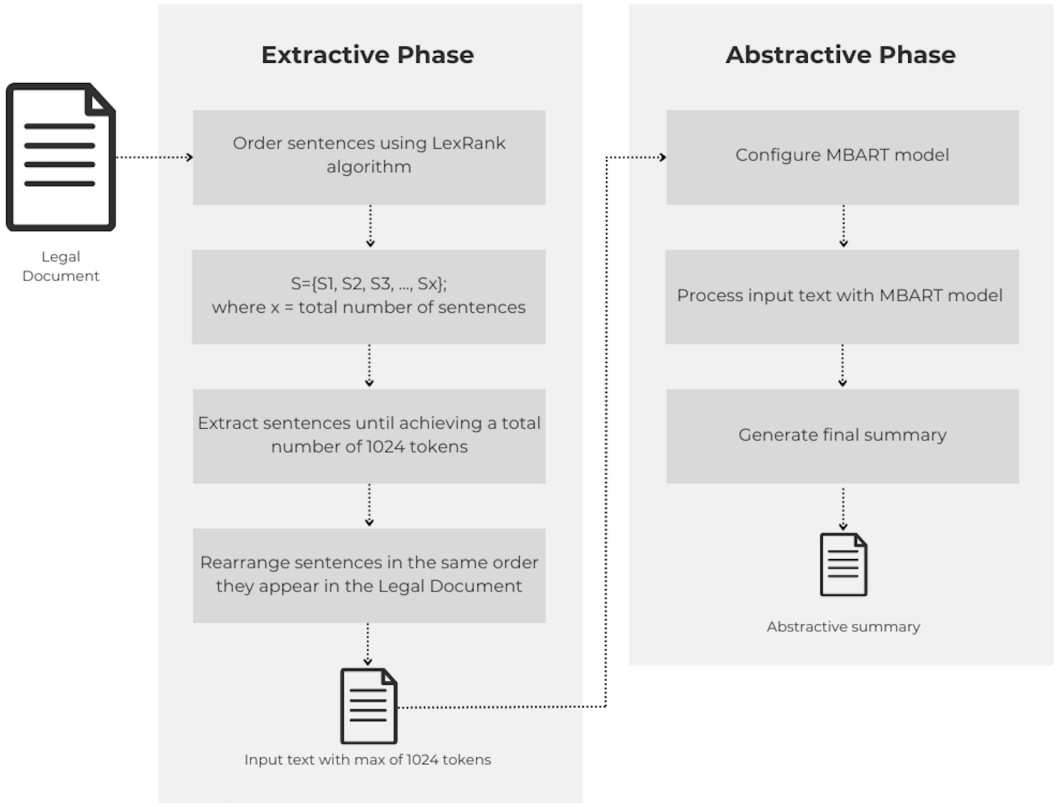


FIGURE 5.3. Hybrid architecture representation

CHAPTER 6

Evaluation and Results

In this chapter, we delve into the analyses of the results obtained from the three implementations explained in Chapter 5. This chapter is divided into four different sections: Section 6.1 describes the results obtained for the sentence-level model; Section 6.2 outlines the outcomes achieved for the summary-level implementation; Section 6.3 presents the findings for the hybrid approach; Finally, Section 6.4 compares the overall results between the three implementations.

In this study, both manual and automatic evaluation methods were employed to assess the quality of the generated summaries. The automatic evaluation process compared the generated summaries with the reference summaries using ROUGE and BERTscore metrics. To achieve a more balanced evaluation that accounts for both coverage and conciseness, we report the results using the ROUGE F1 metric, which combines precision and recall to provide an overall measure of performance. ROUGE-Recall metrics quantify how well the generated summary captures key sentences and keywords from the original document by measuring lexical similarity. Conversely, ROUGE-Precision evaluates the proportion of relevant content within the generated summary, ensuring the avoidance of irrelevant or excessive information. Additionally, we present the results for four specific ROUGE F1 metrics:

- **ROUGE-1:** Measures the proportion of unigrams (single words) that appear in both the generated and reference summaries.
- **ROUGE-2:** Evaluates the proportion of bigrams (two-word sequences) shared between the two summaries.
- **ROUGE-L:** Assesses the longest sequence of words shared between the summaries while maintaining the same order.
- **ROUGE-Lsum:** Compares the longest common sequence of words that appear in the same order within corresponding sentences of the summaries.

BERTscore, on the other hand, evaluates the contextual similarity of words, making it capable of capturing more complex meanings. This metric ensures that the summaries are semantically accurate. In this case we used the BERT base embeddings for Portuguese texts BERTimbau [46], in order to create the representation of the generated and reference summaries. For this evaluation, we employed the BERTimbau embeddings [46], a BERT-based model designed for Portuguese texts, to represent both the generated and reference summaries. The decision to use BERTimbau embeddings for this purpose is consistent with the reasoning outlined in Section 5.2 and ensures methodological consistency across all approaches used in this study.

6.1. Sentence-level Approach Results

In this section, the results of the sentence-level approach were represented individually for each area and section, followed by the discussion of the outcomes.

Table 6.1 shows the results from the sentence-level approach for "Dataset 1". This table represents the summaries generated for each area, with different summary sizes ranging from three to ten sentences.

In general, the results are very consistent, independent of the area or the number of extracted sentences. ROUGE achieved lower results, ranging from 0.128 to 0.368, compared with BERTscore that varies between 0.672 and 0.714. Additionally, it is evident that ROUGE-2 was the metric with the worst results. Also, it was common in every area that the best results for the recall metric were always summaries comprised of eight, nine, or ten sentences. This outcome is probably due to the fact that longer summaries have more content, allowing for the reference summary to have more information represented in the generated summary. These findings can be an indicator that the first extracted sentences, classified as the best ones, do not contain all the information that is represented in the reference summaries, missing out on some specific details.

For the "Cível" and "Social" areas, summaries with fewer sentences achieved the best results, with scores of 0.200 for the ROUGE-L and 0.276 and 0.274 for the ROUGE-Lsum respectively. While these scores are relatively low, they indicate that summaries above six sentences can have more redundant information that is not essential to the overall summary quality, even though the chosen sentences were not the most representative. As expected, the summaries that stood out in the "Criminal" area were the ones with eight and nine sentences showing better ROUGE and BERTscore precision results.

In Table 6.2 we show the results for the generated summaries with three to seven sentences extracted with the LexRank algorithm for "Dataset 2".

For the "Fundamentação de direito" section both ROUGE and BERT-score scores surpass all results for other areas and the "Relatório" section in their respective parameters. With a ROUGE-1 score of 0.408 and ROUGE-Lsum score of 0.350, this section showed a higher overlap of unigrams and the longest common subsequence between the generated and reference summaries.

Contrarily the "Relatório" section had slightly lower results compared to the results in other areas. These lower results may be related to the fact that this was the section with fewer documents in both datasets. For this section, it was clear that the best number of sentences to be extracted would be three or four since they achieved the highest scores compared with the other number of sentences extracted.

In "Dataset 2" the recall results had the same behavior as "Dataset 1" where the highest values correspond for the summaries with seven sentences, with values of 0.734 and 0.716 to the "Fundamentação de direito" and "Relatório" sections, respectively.

Overall the evaluation of the sentence-level approach shows the limitation of LexRank algorithm to extract the exact top sentences from the original document, due to the

TABLE 6.1. Sentence-level approach Evaluation Dataset 1

Area	Extracted Sentences	ROUGE 1	ROUGE 2	ROUGE L	ROUGE Lsum	P_{BERT}	R_{BERT}	F_{BERT}
Cível	3	0.360	0.133	0.200	0.272	0.680	0.691	0.684
	4	0.365	0.138	0.200	0.276	0.676	0.700	0.687
	5	0.361	0.140	0.197	0.274	0.674	0.706	0.689
	6	0.352	0.141	0.192	0.268	0.673	0.709	0.690
	7	0.343	0.140	0.186	0.261	0.673	0.712	0.691
	8	0.333	0.142	0.183	0.257	0.673	0.714	0.692
	9	0.321	0.140	0.177	0.249	0.672	0.714	0.692
	10	0.311	0.140	0.173	0.243	0.673	0.714	0.692
Criminal	3	0.315	0.132	0.179	0.247	0.680	0.675	0.676
	4	0.340	0.147	0.189	0.268	0.680	0.688	0.683
	5	0.352	0.155	0.193	0.279	0.680	0.696	0.687
	6	0.358	0.161	0.196	0.287	0.681	0.703	0.691
	7	0.361	0.164	0.197	0.290	0.681	0.705	0.692
	8	0.364	0.167	0.196	0.292	0.682	0.708	0.694
	9	0.364	0.170	0.196	0.294	0.682	0.709	0.694
	10	0.361	0.170	0.194	0.292	0.682	0.709	0.695
Social	3	0.365	0.136	0.200	0.273	0.678	0.692	0.684
	4	0.368	0.142	0.198	0.274	0.677	0.702	0.688
	5	0.366	0.146	0.196	0.274	0.675	0.709	0.691
	6	0.353	0.146	0.189	0.265	0.673	0.712	0.691
	7	0.340	0.143	0.184	0.255	0.672	0.713	0.691
	8	0.325	0.137	0.177	0.245	0.672	0.714	0.692
	9	0.309	0.132	0.170	0.234	0.672	0.714	0.692
	10	0.297	0.128	0.163	0.225	0.672	0.714	0.692

ROUGE scores being low. However, when analyzing the BERT precision scores we can say that the algorithm can extract sentences that are semantically similar to those in the reference summaries, demonstrating that the algorithm can capture the essence of the original documents (note, however, that this behavior of BERT-score is well-known and might also have an impact on the achieved results since they can show high similarity

TABLE 6.2. Sentence-level approach Evaluation Dataset 2

Section	Extracted Sentences	ROUGE 1	ROUGE 2	ROUGE L	ROUGE Lsum	P_{BERT}	R_{BERT}	F_{BERT}
Fundamentação de direito	3	0.400	0.190	0.233	0.335	0.701	0.713	0.706
	4	0.408	0.203	0.234	0.348	0.700	0.723	0.710
	5	0.404	0.210	0.233	0.350	0.698	0.729	0.712
	6	0.396	0.211	0.226	0.346	0.697	0.732	0.713
	7	0.387	0.212	0.222	0.340	0.696	0.734	0.714
Relatório	3	0.307	0.120	0.181	0.250	0.661	0.700	0.679
	4	0.304	0.126	0.175	0.251	0.664	0.710	0.686
	5	0.287	0.120	0.166	0.239	0.659	0.711	0.683
	6	0.272	0.120	0.158	0.230	0.659	0.713	0.685
	7	0.261	0.119	0.158	0.222	0.660	0.716	0.686

scores even when the actual words differ, as long as the sentences express the same idea). When analyzing the difference between the number of extracted sentences, we verify that it does not have a significant impact on the performance since the results do not show a significant difference.

6.2. Summary-level Approach Results

The goal of using a summary-level implementation was to determine if there were benefits to creating different candidate summaries in order to choose the best one, as opposed to only extracting the best sentences from the original document. Additionally, different approaches to choosing the most relevant summary among the candidate summaries were tested. In Tables 6.3 and 6.4 we show the results obtained for "Dataset 1" and "Dataset 2", respectively. In Figures 6.1 and 6.2 we present statistics of the length and the content of the summaries. It is important to take into consideration that these results exclude the documents for which this method was unable to generate summaries due to a lack of sentences that could be employed in calculating the similarity between candidates without overlap sentences.

For all the metrics applied in "Dataset 1", choosing the most relevant summary from the candidates allows us to infer that the "Mean" metric showed the most favorable results, despite demonstrating minimal variation in the outcomes. When comparing the results of the "Mean" and the sentence-level approach, the "Mean" approach achieved a ROUGE-Lsum score of 0.298. All the other metrics never overcame the sentence-level approach results, however, the highest difference was for ROUGE-2 of 0.030 for the "Criminal" area.

TABLE 6.3. Summary-level approach Evaluation Dataset 1

Select Best Candidate by:	ROUGE 1	ROUGE 2	ROUGE L	ROUGE Lsum	P_{BERT}	R_{BERT}	F_{BERT}
max	0.349	0.137	0.196	0.292	0.672	0.691	0.680
similarity $>$ 80%	0.342	0.129	0.195	0.283	0.671	0.685	0.677
similarity $>$ 90%	0.346	0.129	0.194	0.285	0.672	0.688	0.679
mean	0.353	0.140	0.196	0.298	0.670	0.697	0.683

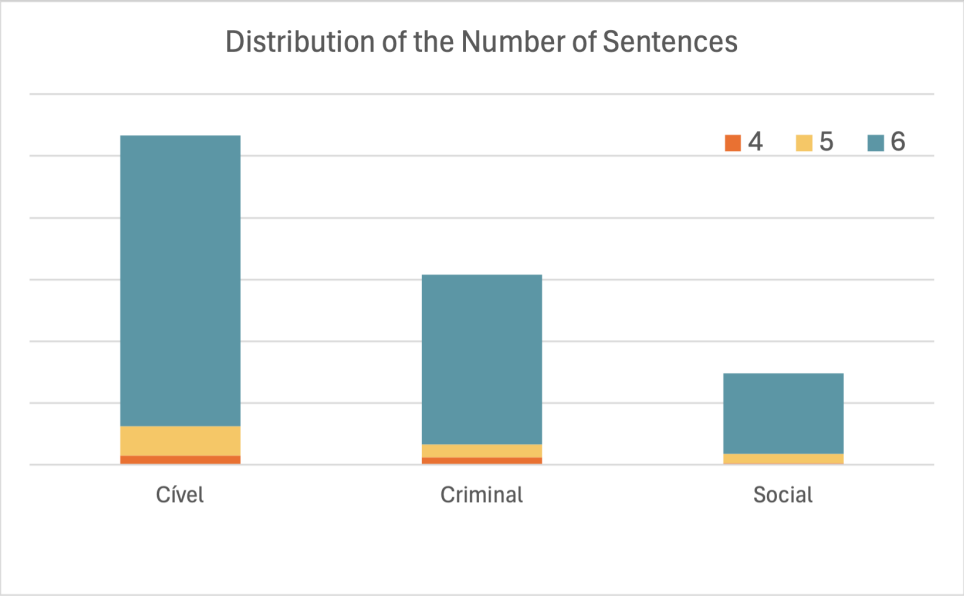
For "Dataset 2" also the "Mean" metric was the one that stood out the most. Only the BERT precision metric had better results with a value of 0.702 for the "similarity score >80%". Additionally, in opposition to the results of the "Dataset 1" the outcomes from the summary-level implementation overcome the ROUGE-1, ROUGE-L, ROUGE-Lsum, and BERT precision metrics, when compared with the sentence-level results.

One more time the "Dataset 2" demonstrates higher efficiency than "Dataset 1". Additionally, both datasets showed that using a "similarity score greater than X%" metric resulted in better performance as the chosen percentage increased. As represented in Tables 6.3 and 6.4 the "similarity score $>$ 90%' showed a small increase in the performance compared to the "similarity score $>$ 80%" metric.

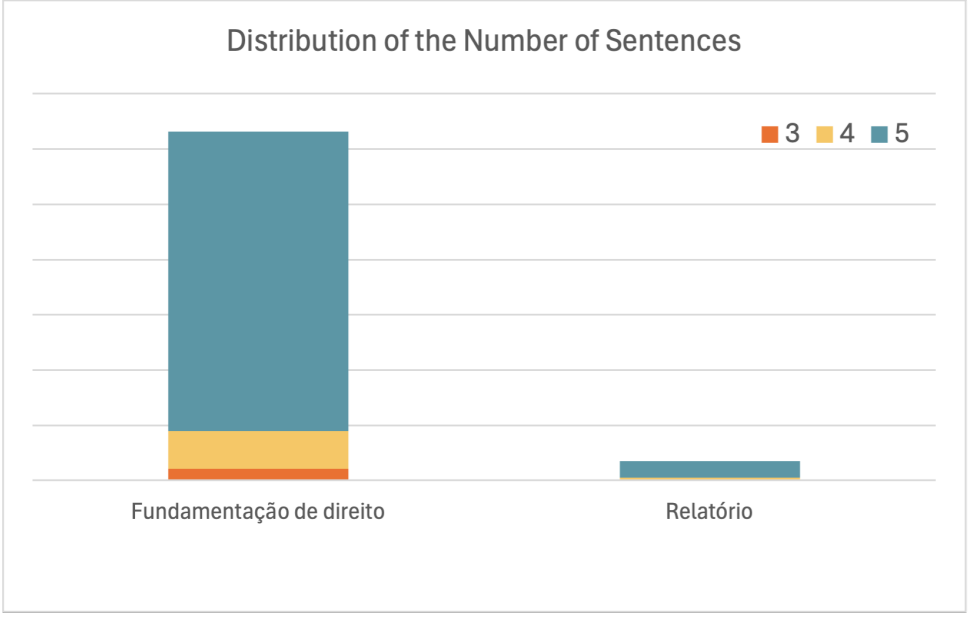
TABLE 6.4. Summary-level approach Evaluation Dataset 2

Select Best Candidate By:	ROUGE 1	ROUGE 2	ROUGE L	ROUGE Lsum	P_{BERT}	R_{BERT}	F_{BERT}
max	0.401	0.191	0.239	0.343	0.699	0.717	0.707
similarity $>$ 80%	0.390	0.178	0.237	0.327	0.701	0.707	0.703
similarity $>$ 90%	0.407	0.192	0.249	0.344	0.702	0.714	0.707
mean	0.413	0.207	0.250	0.358	0.699	0.727	0.712

Analyses to the structure of the generated summaries were also conducted. We use the results obtained from the "Mean" metric since it was the one that achieved the best results for both datasets. Figure 6.1 offers a clear view of the distributions of the number of sentences from the generated summaries. We verify the length of the generated summaries by examining the number of sentences included in the final summaries. For "Dataset 1" it is clear that the majority of the generated summaries comprise six sentences for all areas(see Figure 6.1a). And in "Dataset 2" the most prevalent number of sentences was five (see Figure 6.1b).



(A) Dataset 1



(B) Dataset 2

FIGURE 6.1. Distribution of the number of sentences using the summary-level "Mean" results

In order to understand if there was a difference between the content of summaries generated using a summary-level approach versus a sentence-level approach (in our case LexRank) approaches, we analyzed the percentage of the sentences in the summary-level implementation that matched those in a summary of the same size generated by the LexRank algorithm. We found that only 33% to 60% of sentences in a summary generated by the summary-level approach were the same as those in a summary generated by the sentence-level approach for "Dataset 1", and between 40% and 60% for "Dataset 2" (see Figure 6.2). Based on these results it visible the difference in the sentences chosen from the summary-level approach and the sentence-level approach.

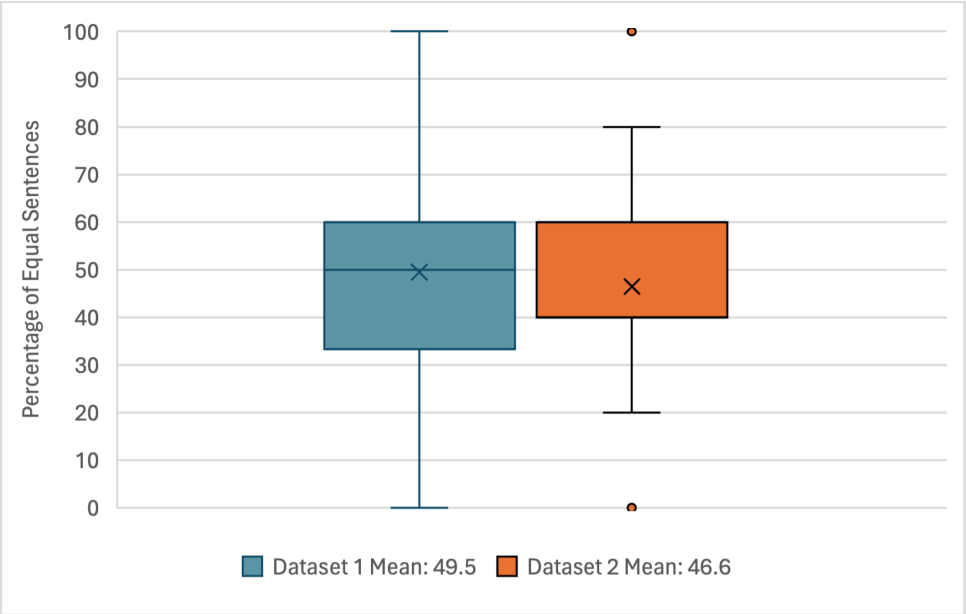


FIGURE 6.2. Percentage of equal sentences between summaries generated with a sentence-level and a summary-level approach

6.3. Hybrid Approach Results

In this section, we present the discussion of the results for the ROUGE and BERT-score metrics for the hybrid implementation. The results are presented in Tables 6.5 and 6.6 for "Dataset 1" and "Dataset 2", respectively.

Independent of the area or section the results never overcome the sentence-level or the summary-level approaches. Showing a decrease of performance on the generation of the summaries using an ATS model. The decrease in the ROUGE score can be explainable since we used a text generation model. Which may create a final summary with similar words instead of the actual words in the input text. However, the BERT-score also had a reduction in the score, meaning that the content generated from the hybrid approach is not as similar to the ones generated by the extractive approaches. Another possible reason for this effect was that when analyzing the content of the summaries manually, we observed that some of them did not comprise an actual text but a repetition of words or letters (see Figure 6.3).

S. SGPS, S.A., S.A., S.A., S., S.A., S., S.A., S., S.A., S., S., S.A., S., S.A., S., S.A., S., S.A., S., S.A., S., S.A., S., S., S., S.A., S., S., S., S., S., S., S., S.,[...], S., S., S., S., S.

FIGURE 6.3. Example of a repeated pattern in a generated summary using the hybrid approach

When comparing the generation of summaries, based on the parameter max_tokens, the summaries generated with a maximum of 600 tokens prove to be more effective in summarization, achieving higher scores across all metrics, with the exception of the "Relatório" section.

TABLE 6.5. Hybrid approach Evaluation Dataset 1

Areas	max tokens	ROUGE 1	ROUGE 2	ROUGE L	ROUGE Lsum	P_{BERT}	R_{BERT}	F_{BERT}
Cível	600	0.270	0.078	0.154	0.209	0.597	0.626	0.610
	350	0.249	0.070	0.142	0.192	0.578	0.613	0.594
Criminal	600	0.239	0.077	0.140	0.197	0.599	0.625	0.610
	700	0.237	0.076	0.138	0.194	0.589	0.620	0.602
Social	600	0.267	0.076	0.154	0.203	0.598	0.623	0.609
	350	0.257	0.073	0.147	0.197	0.587	0.616	0.600

TABLE 6.6. Hybrid approach Evaluation Dataset 2

Sections	max tokens	ROUGE 1	ROUGE 2	ROUGE L	ROUGE Lsum	P_{BERT}	R_{BERT}	F_{BERT}
Fundamentação de direito	600	0.318	0.130	0.193	0.255	0.636	0.655	0.644
	350	0.307	0.127	0.185	0.248	0.624	0.647	0.634
Relatório	600	0.199	0.056	0.118	0.151	0.546	0.592	0.566
	350	0.234	0.070	0.139	0.183	0.565	0.606	0.583

The ten most frequent words were analyzed with the objective of determining whether the key terms of the generated summaries resembled those of the reference summaries (see Figure 4.2) when employing a text generation model. For the development of the graphs, the generated summaries that had a number of "max_tokens" of 600 tokens were used, and to count the frequency of words, the same process was followed as in Chapter 4. Figure 6.4 represents the top ten terms of "Dataset 1" while Figure 6.5 illustrates the top keywords for "Dataset 2".

For both datasets, the top word was "article" with a frequency higher than 1600 tokens, matching the top word identified in the reference summaries (see Figure 4.2). Additionally,

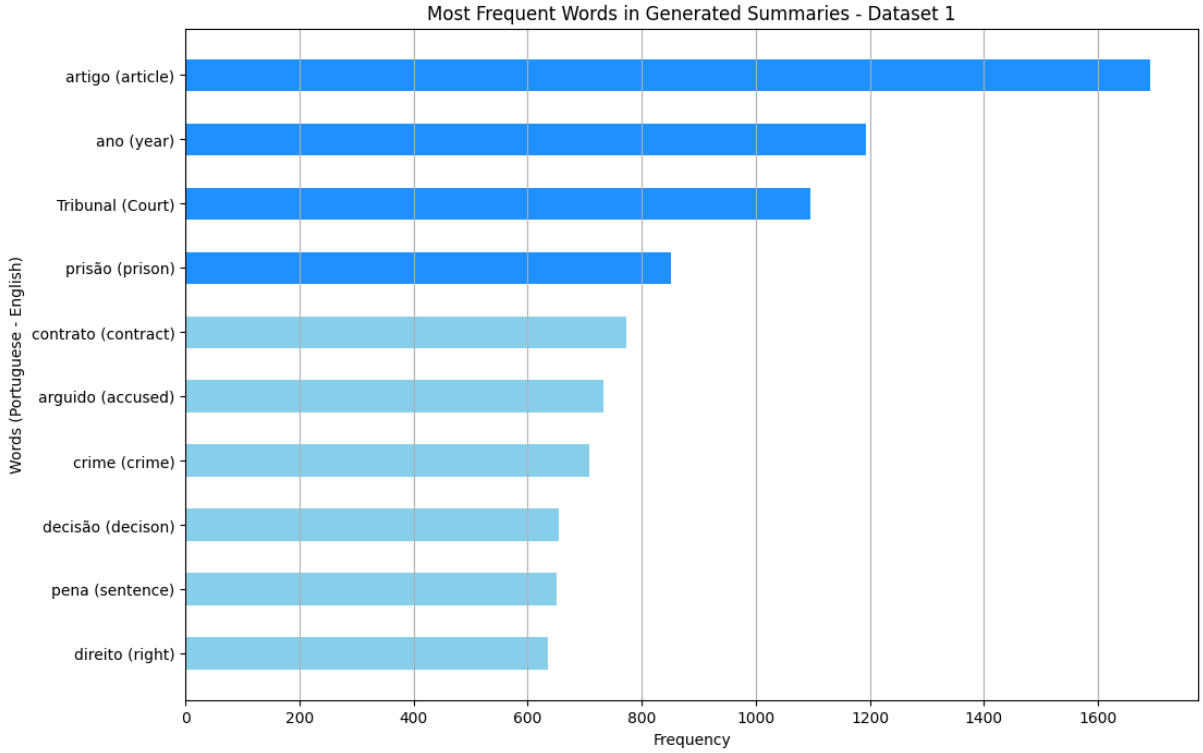


FIGURE 6.4. Ten most frequent words in hybrid generated summaries for Dataset 1

the words "Court" and "prison" in "Dataset 1" and "Civil", "proof", "code" and "Court" in "Dataset 2" were not identified in the reference summaries as top terms. Although they still have significant matters in the legal domain.

In the reference summaries, the words "fact", "sentence", and "right" were identified as the most frequent ones besides the term "article". However, the word "fact" in "Dataset 1" is not even represented, and the other two terms are positioned at the bottom of the graph. Contrarily, the summaries from "Dataset 2" place a strong emphasis on the words "right" and "fact", but the term "sentence" is not as prominent. This difference in word frequency highlights potential variations in generated summaries using the entire document or just one section of it.

One common characteristic in both figures is the range of the frequency axis. Even though "Dataset 1" has more generated summaries than "Dataset 2" the frequency statistics are very similar, meaning that using only a specific section rather than the entire document can enhance the performance of the models since it can capture more relevant words.

6.4. Discussion

In this section, we compare the performance of the sentence-level, summary-level, and hybrid summarization models, highlighting the strengths and limitations of each approach.

For both ETS models, sentence-level and summary-level approaches, the number of sentences that should be included in the summary was a defined parameter from the

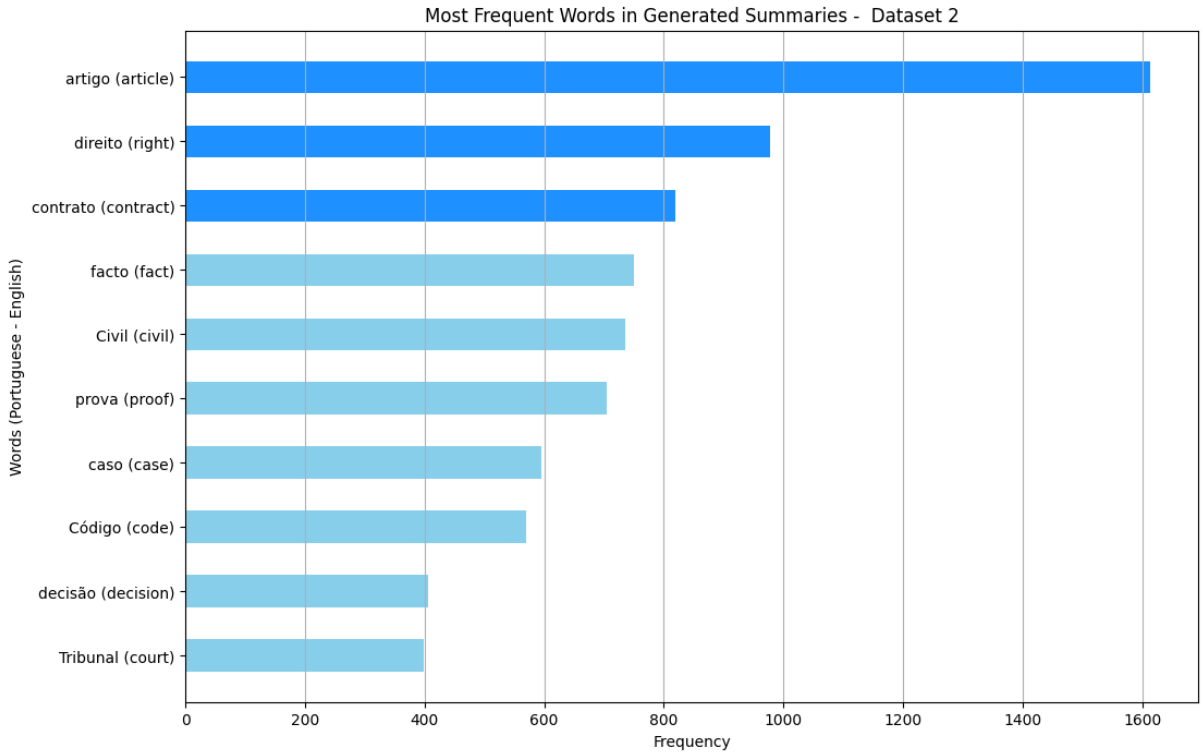


FIGURE 6.5. Ten most frequent words in hybrid generated summaries for Dataset 2

beginning. When comparing the performance of these two approaches we could verify that the "Mean' metric for the summary-level approach surpassed the sentence-level implementation across both datasets. These results suggest that creating a set of candidate summaries may slightly improve the performance of the ETS algorithms, as it allows for the selection of sentences that more closely match the reference summaries. Another comparison done between the two implementation focused on the difference in content selected for the summaries, as showed in Figure 6.2. We observed that the top sentences selected in the sentence-level approach differed from those included in the summary-level approach. This highlights a significant point, particularly within the "Criminal' area: for the summary-level approach, the most effective summaries were those comprising four to six sentences, whereas in the sentence-level approach, summaries containing eight to nine sentences achieved similar results. Based on these findings, we conclude that the summary-level approach produced more concise summaries while maintaining the context.

Additionally, the ROUGE and BERTscore results were similar across all three models (sentence-level, summary-level and hybrid), with only minor discrepancies. However, the ROUGE results were consistently lower, indicating that the models struggled to select all the exact sentences from the original documents. This demonstrates a lack of performance, particularly for the sentence-level and summary-level approaches due to their nature of extracting sentences from the original document, which should produce a higher ROUGE score if they selected the correct sentences. In contrast, the BERTscore results

were slightly better, suggesting that the generated summaries were more closely aligned with the reference summaries in terms of the topics covered. This metric holds particular relevance for the hybrid approach due to its potential to generate new words or sentences. Despite this, the hybrid approach demonstrated the lowest results among the three implementations, indicating that improvements are necessary in the abstractive component of this approach.

In conclusion, extractive models are effective in reducing the length of long documents, especially when using a summary-level approach. This method demonstrated proficiency in identifying the most informative content with fewer sentences. Also, combining extractive and abstractive techniques can further enhance the fluency of the generated summaries. However, the issue of hallucination in abstractive models must be addressed to achieve better results.

CHAPTER 7

Conclusions and Future Work

This chapter presents the conclusions regarding the experiments done in the field of automatic TS for Portuguese legal documents from the Portuguese Supreme Court of Justice. We describe the main contributions of our study as well as the limitations we found throughout the project. Additionally, we address some possible improvements for future work in the automatic TS domain.

7.1. Scientific Contributions

Through this dissertation, it was possible to make several contributions to the field of automatic TS, with special focus on Portuguese documents from the Portuguese Supreme Court of Justice. The most significant contributions are as follows:

- A comprehensive review of recent techniques and algorithms used in the process of automatic TS, with a particular focus on approaches specialized for legal document summarization.
- The investigation and implementation of different TS approaches, including sentence-level, summary-level, and hybrid approaches, to evaluate their effectiveness in summarizing Portuguese legal documents.
- The proposal of a new hybrid approach combining the LexRank algorithm and the MBART model, surpassing the limitations related to the maximum number of tokens of MBART models.
- The identification that the using a summary-level approach plays a significant role in the summarization of Portuguese judgments from the Portuguese Supreme Court of Justice.
- A publication at the SLATE'24 conference on the topic of Human-Human Language. This publication culminates the key aspects of the work conducted on this dissertation.

7.2. Conclusions

The main goal of this dissertation was to investigate different TS approaches in order to comprehend their effectiveness in summarizing Portuguese legal documents. In this work, we investigate both ETS and ATS approaches.

In this research, the evaluation of the models was conducted by comparing the generated summaries with reference summaries mainly using two automatic evaluation techniques: ROUGE and BERTscore. After obtaining and analysing all results we conclude that the ROUGE metric was accurate on capturing overlap terms and sentences between

two texts. Also, the BERTscore metric was accurate in capturing the semantic similarity between two texts. This explains the higher results for the BERTscore metrics when compared to the ROUGE ones. However, both metrics have inherent limitations, as they do not fully capture the nuances of summary quality on their own, which are essential for a comprehensive evaluation.

When generating summaries, one of the key parameters to consider is the length of the output text, whether to produce a more concise summary that captures the general information or a longer one that includes more detailed content. Independently of the length, the summary should always cover the main context of the original document. In this research, we used a sentence-level approach with the LexRank algorithm as a starting point, with the main goal of evaluating how different summary sizes impact the overall quality of the summary. The results showed that the summaries from the "Criminal" area had slightly better results for summaries comprising eight and nine sentences, while all the other areas and sections showed maximum results for summaries composed of three to five sentences. However, the variation in performance across different lengths was minimal, making it difficult to conclude that these lengths represent the optimal summary length for Portuguese legal documents.

With the summary-level approach, we try different methods to choose the most relevant summary among all possible candidates. The results showed that verifying the mean of the similarity scores of all mutually exclusive candidates obtained higher results. Additionally, the ROUGE-Lsum metric obtained a score of 0.298 in "Dataset 1" indicating that it was better at selecting the most significant sentences when compared with the sentence-level approach. For "Dataset 2", Figure 6.4 shows that this implementation actually achieved higher results than the sentence-level approach. Implementing a summary-level approach allows us to conclude that it is possible to obtain a similar level of information with a summary that only includes 30-60% of the top sentences selected as the most informative by a sentence-level summary, LexRank.

Our findings indicate that the length of the documents is not a limiting factor when implementing ETS approaches using the LexRank algorithm since it was capable of processing the entirety of the document independently of its length. Conversely, when dealing with the MBART model, an ATS, we encountered a limitation concerning the number of tokens that could be used, leading to the contribution of a new hybrid implementation. By creating a two phase model, where it first extracts the most relevant content of a document until it reaches the MBART limit of tokens and then passes through an abstractive phase in order to generate a more fluent summary. The results showed a decrease of performance on using an HTS model compared with the other two models. This decline could be explained due to the model experiencing hallucinations, where in some cases it was unable to produce a structured and accurate summary but instead presented repetitions of words.

The creation of two different datasets, one comprised of the original documents and the other with only some specific sections of the documents, allowed us to verify the differences between generating a summary using each approach. We found that summaries generated from the "Fundamentação de Direito" section yielded promising results. However, the limitation of extracting this section from the original documents affected the efficiency of the process, as we were unable to consistently extract the sections from all judgments. This resulted in the inability to conduct a meaningful comparative analysis between the two approaches.

Although not directly addressed in the experimental results, the selection of the algorithms and models employed in this study were also based on the privacy and legal implications. Summarizing Portuguese legal documents involve dealing with sensitive data. Legal cases summaries need to be very precise and all the details are necessary to generate a summary from a legal document. Subsequently, given the importance of maintaining the privacy of sensitive data, it was imperative to prioritize the selection of algorithms that could ensure the security of the data and generate a summary with quality.

In summary, the investigation of both extractive and abstractive techniques revealed that each approach has its own strengths and weaknesses. Both sentence-level and summary-level approaches were effective in extracting sentences that captured the general context of the original document. However, they lacked the capacity to maintain fluency between sentences and required improvement when selecting sentences that provide more specific details about the cases. The abstractive model, MBART, faced challenges in processing the entire legal document due to its token limitation, which led to the implementation of a hybrid approach. Despite these challenges, the hybrid method demonstrated promising directions for future research, particularly when applied to lengthy documents where fluency and context preservation are crucial.

7.3. Future Work

This section presents several adjustments that could be implemented in future work to enhance the performance of the models and techniques used in this research.

Conducting this study revealed some adaptations that could improve the effectiveness of the summarization models, including:

- **Portuguese embeddings for LexRank:** In the LexRank algorithm, integrating Portuguese embeddings can provide a more accurate representation of the documents, improving the accuracy of sentence-level and summary-level summaries.
- **Fine-tuning MBART for Portuguese legal documents:** Adapting MBART to the domain of legal documents could enhance its ability to generate more precise summaries.

- **Larger Dataset:** Increasing the dataset to include a more diverse collection of Portuguese legal documents would allow the models to better understand the patterns of the legal documents.
- **Hybrid model improvement:** We propose the development of a model that combines the three summarization approaches. This model would include three phases:
  (1) The initial step, where the sentences would be ranked by LexRank based on their importance.
  (2) In a second phase, the best sentences would be selected using a summary-level approach, where the metric to select the most relevant summary among all candidates would be the "Mean".
  (3) Finally, the selected candidate from the previous phase would be the input to the MBART model, with the goal of obtaining a structured summary that maintains the context of the original document.

References

[1] E. D. Liddy, "Natural language processing," 2001.

[2] I. Awasthi, K. Gupta, P. S. Bhogal, S. S. Anand, and P. K. Soni, "Natural language processing (nlp) based text summarization-a survey," in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, IEEE, 2021, pp. 1310–1317.

[3] H. P. Luhn, "The automatic creation of literature abstracts," *IBM J. Res. Dev.*, vol. 2, no. 2, pp. 159–165, 1958. DOI: 10.1147/RD.22.0159. [Online]. Available: https://doi.org/10.1147/rd.22.0159.

[4] W. B. Demilie, "Comparative analysis of automated text summarization techniques: The case of ethiopian languages," *Wireless Communications & Mobile Computing (Online)*, vol. 2022, 2022.

[5] W. Mengist, T. Soromessa, and G. Legese, "Method for conducting systematic literature review and meta-analysis for environmental science research," *MethodsX*, vol. 7, p. 100 777, 2020.

[6] M. J. Page, J. E. McKenzie, P. M. Bossuyt, *et al.*, "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *International journal of surgery*, vol. 88, p. 105 906, 2021.

[7] G. Erkan and D. R. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *CoRR*, vol. abs/1109.2128, 2011.

[8] Y. Liu, J. Gu, N. Goyal, *et al.*, "Multilingual denoising pre-training for neural machine translation," *Trans. Assoc. Comput. Linguistics*, vol. 8, pp. 726–742, 2020. DOI: 10.1162/TACL_A_00343. [Online]. Available: https://doi.org/10.1162/tacl%5C_a%5C_00343.

[9] M. Barbella and G. Tortora, "Rouge metric evaluation for text summarization techniques," *Available at SSRN 4120317*, 2022.

[10] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with BERT," in *ICLR*, OpenReview.net, 2020.

[11] M. Koniaris, D. Galanis, E. Giannini, and P. Tsanakas, "Evaluation of automatic legal text summarization techniques for greek case law," *Inf.*, vol. 14, no. 4, p. 250, 2023. DOI: 10.3390/INFO14040250. [Online]. Available: https://doi.org/10.3390/info14040250.

[12] N. Begum and A. Goyal, "Analysis of legal case document automated summarizer," in *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, IEEE, 2021, pp. 533–538.

[13] A. Kanapala, S. Pal, and R. Pamula, "Text summarization from legal documents: A survey," *Artif. Intell. Rev.*, vol. 51, no. 3, pp. 371–402, 2019.

[14] H. Jin, Y. Zhang, D. Meng, J. Wang, and J. Tan, "A comprehensive survey on process-oriented automatic text summarization with exploration of llm-based methods," *CoRR*, vol. abs/2403.02901, 2024. DOI: 10.48550/ARXIV.2403.02901. arXiv: 2403.02901. [Online]. Available: https://doi.org/10.48550/arXiv.2403.02901.

[15] D. Jain, M. D. Borah, and A. Biswas, "Summarization of lengthy legal documents via abstractive dataset building: An extract-then-assign approach," *Expert Syst. Appl.*, vol. 237, no. Part B, p. 121 571, 2024. DOI: 10.1016/J.ESWA.2023.121571. [Online]. Available: https://doi.org/10.1016/j.eswa.2023.121571.

[16] D. Jain, M. D. Borah, and A. Biswas, "Summarization of legal documents: Where are we now and the way forward," *Comput. Sci. Rev.*, vol. 40, p. 100 388, 2021. DOI: 10.1016/J.COSREV.2021.100388. [Online]. Available: https://doi.org/10.1016/j.cosrev.2021.100388.

[17] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *Multim. Tools Appl.*, vol. 82, no. 3, pp. 3713–3744, 2023. DOI: 10.1007/S11042-022-13428-4. [Online]. Available: https://doi.org/10.1007/s11042-022-13428-4.

[18] A. Clark, "" enhancing and exploring the use of transformer models in nlp tasks,"

[19] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A survey of automatic text summarization: Progress, process and challenges," *IEEE Access*, vol. 9, pp. 156 043–156 070, 2021.

[20] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," *Expert Syst. Appl.*, vol. 165, p. 113 679, 2021.

[21] G. R. Sundar, C. S. Nageswari, J. A. Vijay, S. R. Devi, N. Basker, and N. Pushpa, "Machine learning based extractive text summarization techniques," in *2023 International Conference on Integrated Intelligence and Communication Systems (ICIICS)*, IEEE, 2023, pp. 1–6.

[22] A. K. Yadav, Ranvijay, R. S. Yadav, and A. K. Maurya, "State-of-the-art approach to extractive text summarization: A comprehensive review," *Multim. Tools Appl.*, vol. 82, no. 19, pp. 29 135–29 197, 2023.

[23] R. K. B. Lenka, T. Coombs, S. Assi, *et al.*, "Evaluation of extractive and abstract methods in text summarization," in *Data Science and Emerging Technologies - Proceedings of DaSET 2022, Virtual Event, 20-21 December, 2022*, Y. B. Wah, M. W. Berry, A. Mohamed, and D. Al-Jumeily, Eds., ser. Lecture Notes on Data Engineering and Communications Technologies, vol. 165, Springer, 2022, pp. 535–546. DOI: 10.1007/978-981-99-0741-0_38. [Online]. Available: https://doi.org/10.1007/978-981-99-0741-0%5C_38.

[24] M. Schraagen, F. Bex, N. V. D. Luijtgaarden, and D. Prijs, "Abstractive summarization of dutch court verdicts using sequence-to-sequence models," in *NLLP@EMNLP*, Association for Computational Linguistics, 2022, pp. 76–87.

[25] M. Kavitha and K. Akila, "An exploratory study of abstractive text summarization using a sequence-to-sequence model," in *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, IEEE, 2023, pp. 1–5.

[26] F. Neiva and R. Silva, "Systematic literature review in computer science - a practical guide," no. 1, 2016.

[27] Y. Zhong and D. J. Litman, "Computing and exploiting document structure to improve unsupervised extractive summarization of legal case decisions," in *Proceedings of the Natural Legal Language Processing Workshop, NLLP@EMNLP 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 8, 2022*, N. Aletras, I. Chalkidis, L. Barrett, C. Goanta, and D. Preotiuc-Pietro, Eds., Association for Computational Linguistics, 2022, pp. 322–337. [Online]. Available: https://aclanthology.org/2022.nllp-1.30.

[28] D. Anand and R. S. Wagh, "Effective deep learning approaches for summarization of legal texts," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 5, pp. 2141–2150, 2022. DOI: 10.1016/J.JKSUCI.2019.11.015. [Online]. Available: https://doi.org/10.1016/j.jksuci.2019.11.015.

[29] R. Wicks and M. Post, "A unified approach to sentence segmentation of punctuated text in many languages," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds., Association for Computational Linguistics, 2021, pp. 3995–4007. DOI: 10.18653/V1/2021.ACL-LONG.309. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.309.

[30] M. C. C. Medina, L. M. D. S. Oliveira, J. F. C. Ferreira, *et al.*, "Classification of legal documents in portuguese language based on summarization," in *2022 IEEE Latin American Conference on Computational Intelligence (LA-CCI), Montevideo, Uruguay, November 23-25, 2022*, IEEE, 2022, pp. 1–6. DOI: 10.1109/LA-CCI54402.2022.9981852. [Online]. Available: https://doi.org/10.1109/LA-CCI54402.2022.9981852.

[31] D. Jain, M. D. Borah, and A. Biswas, "Fine-tuning textrank for legal document summarization: A bayesian optimization based approach," P. Majumder, M. Mitra, S. Gangopadhyay, and P. Mehta, Eds., pp. 41–48, 2020. DOI: 10.1145/3441501.3441502. [Online]. Available: https://doi.org/10.1145/3441501.3441502.

[32] Y. Huang, L. Sun, C. Han, and J. Guo, "A high-precision two-stage legal judgment summarization," *Mathematics*, vol. 11, no. 6, p. 1320, 2023.

[33] R. Rani and D. K. Lobiyal, "A weighted word embedding based approach for extractive text summarization," *Expert Syst. Appl.*, vol. 186, p. 115 867, 2021.

[34] A. P. Widyassari, E. Noersasongko, A. Syukur, *et al.*, "An extractive text summarization based on candidate summary sentences using fuzzy-decision tree," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 7, 2022.

[35] M. Kondath, D. P. Suseelan, and S. M. Idicula, "Extractive summarization of malayalam documents using latent dirichlet allocation: An experience," *J. Intell. Syst.*, vol. 31, no. 1, pp. 393–406, 2022. DOI: 10.1515/JISYS-2022-0027. [Online]. Available: https://doi.org/10.1515/jisys-2022-0027.

[36] D. O. Cajueiro, A. G. Nery, I. Tavares, *et al.*, "A comprehensive review of automatic text summarization techniques: Method, data, evaluation and coding," *CoRR*, vol. abs/2301.03403, 2023. DOI: 10.48550/ARXIV.2301.03403. arXiv: 2301.03403. [Online]. Available: https://doi.org/10.48550/arXiv.2301.03403.

[37] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *J. Documentation*, vol. 60, no. 5, pp. 493–502, 2004.

[38] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR (Workshop Poster)*, 2013.

[39] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, ACL, 2014, pp. 1532–1543.

[40] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds., Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/V1/N19-1423. [Online]. Available: https://doi.org/10.18653/v1/n19-1423.

[41] D. Jain, M. D. Borah, and A. Biswas, "A sentence is known by the company it keeps: Improving legal document summarization using deep clustering," *Artificial Intelligence and Law*, pp. 1–36, 2023.

[42] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: the muppets straight out of law school," *CoRR*, vol. abs/2010.02559, 2020. arXiv: 2010.02559. [Online]. Available: https://arxiv.org/abs/2010.02559.

[43] Y. Sun, F. Yang, X. Wang, and H. Dong, "Automatic generation of the draft procuratorial suggestions based on an extractive summarization method: Bertslca," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–12, 2021.

[44] C. Raffel, N. Shazeer, A. Roberts, *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, 140:1–140:67, 2020.

[45] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[46] F. Souza, R. F. Nogueira, and R. de Alencar Lotufo, "Bertimbau: Pretrained BERT models for brazilian portuguese," in *BRACIS*, ser. Lecture Notes in Computer Science, vol. 12319, Springer, 2020, pp. 403–417.

[47] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *EMNLP*, ACL, 2004, pp. 404–411.

[48] S. Klaus, R. V. Hecke, K. D. Naini, I. S. Altingovde, J. Bernabé-Moreno, and E. Herrera-Viedma, "Summarizing legal regulatory documents using transformers," in *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, and G. Kazai, Eds., ACM, 2022, pp. 2426–2430. DOI: 10.1145/3477495.3531872. [Online]. Available: https://doi.org/10.1145/3477495.3531872.

[49] P. Bhattacharya, S. Poddar, K. Rudra, K. Ghosh, and S. Ghosh, "Incorporating domain knowledge for extractive summarization of legal case documents," in *ICAIL '21: Eighteenth International Conference for Artificial Intelligence and Law, São Paulo Brazil, June 21 - 25, 2021*, J. Maranhão and A. Z. Wyner, Eds., ACM, 2021, pp. 22–31. DOI: 10.1145/3462757.3466092. [Online]. Available: https://doi.org/10.1145/3462757.3466092.

[50] A. Mandal, P. Bhattacharya, S. Mandal, and S. Ghosh, "Improving legal case summarization using document-specific catchphrases," in *JURIX*, ser. Frontiers in Artificial Intelligence and Applications, vol. 346, IOS Press, 2021, pp. 76–81.

[51] M. Zhong, P. Liu, Y. Chen, D. Wang, X. Qiu, and X. Huang, "Extractive summarization as text matching," in *ACL*, Association for Computational Linguistics, 2020, pp. 6197–6208.

[52] S. Gong, Z. Zhu, J. Qi, W. Wu, and C. Tong, "Sebursum: A novel set-based summary ranking strategy for summary-level extractive summarization," *J. Supercomput.*, vol. 79, no. 12, pp. 12 949–12 977, 2023.

[53] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 3104–3112. [Online]. Available: https://proceedings.neurips.cc/paper/2014/hash/a14ac55a4f27472c5d894ec1c3c743d2-Abstract.html.

[54] M. Lewis, Y. Liu, N. Goyal, *et al.*, "BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *ACL*, Association for Computational Linguistics, 2020, pp. 7871–7880.

[55] M. Guo, J. Ainslie, D. C. Uthus, *et al.*, "Longt5: Efficient text-to-text transformer for long sequences," in *Findings of the Association for Computational Linguistics: NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, M. Carpuat, M. de Marneffe, and I. V. M. Ruíz, Eds., Association for Computational Linguistics, 2022, pp. 724–736. DOI: 10.18653/V1/2022.FINDINGS-NAACL.55. [Online]. Available: https://doi.org/10.18653/v1/2022.findings-naacl.55.

[56] J. Zhao, T. Wang, W. Abid, *et al.*, "Lora land: 310 fine-tuned llms that rival gpt-4, A technical report," *CoRR*, vol. abs/2405.00732, 2024. DOI: 10.48550/ARXIV.2405.00732. arXiv: 2405.00732. [Online]. Available: https://doi.org/10.48550/arXiv.2405.00732.

[57] L. Dong, N. Yang, W. Wang, *et al.*, "Unified language model pre-training for natural language understanding and generation," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, Eds., 2019, pp. 13 042–13 054. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/c20bb2d9a50d5ac1f713f8b34d9aac5a-Abstract.html.

[58] M. Zanatti, R. Ribeiro, and H. S. Pinto, "Segmenting model for portuguese judgments," in *Progress in Artificial Intelligence - 23nd EPIA Conference on Artificial Intelligence, EPIA 2024, Viana do Castelo September 3-6, 2024*, 2024.