UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

LEONARDO BONALUME DE ANDRADE

BB25HLegalSum: a method for legal document summarization that leverages BM25 and BERT-based clustering

Thesis presented in partial fulfillment
of the requirements for the degree of
Master of Computer Science

Advisor: Profa. Dra. Karin Becker

Porto Alegre
January 2024

CIP — CATALOGING-IN-PUBLICATION

de Andrade, Leonardo Bonalume

BB25HLegalSum: a method for legal document summarization that leverages BM25 and BERT-based clustering / Leonardo Bonalume de Andrade. – Porto Alegre: PPGC da UFRGS, 2024.

62 f.: il.

Thesis (Master) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação, Porto Alegre, BR–RS, 2024. Advisor: Karin Becker.

1. Text summarization. 2. Legal documents. 3. BERT. 4. Multiple color highlighting. 5. Multiple criteria highlighting. I. Becker, Karin. II. Título.

ACKNOWLEDGMENT

I want to begin expressing my gratitude to my family for the presence and support.

I am grateful to my advisor, Karin Becker, because of her teachings and guidance throughout this work.

I am also grateful to André, Bruno, Igor, Lillian, Talita for the insights and support.

I want to thank the members of the Informatics Institute at the UFRGS for all the insights and shared knowledge.

ABSTRACT

Legal document summarization aims to provide a clear understanding of the main points and arguments in a legal document, contributing to the efficiency of the judicial system. In this work, we propose BB25HLegalSum, a method that combines BERT clusters with the BM25 algorithm to summarize legal documents and present them to users with highlighted important information. The process involves selecting unique sentences from the original document, clustering them to find sentences about a similar subject, scoring clusters and sentences to generate a summary according to three strategies, and highlighting them to the user in the original document. Legal workers positively assessed the highlighted presentation.

Keywords: Text summarization. Legal documents. BERT. Multiple color highlighting. Multiple criteria highlighting.

LIST OF ABBREVIATIONS AND ACRONYMS

BERT	Bidirectional Encoder Representations from Transformers
BM25	Best Match 25
candSum	Candidate summaries
COLIEE	Competition on Legal Information Extraction/Entailment
desc	Legal description
GloVe	Global Vectors for Word Representation
GPT	Generative Pre-trained Transformer
GSum	Generated summary
KLSumm	Kullback-Leibler Algorithm
KL	Kullback-Leibler (KL)
LSTM	Long Short-Term Memory
NLP	Natural Language Processing
refSum	Reference summary
repSum	Representative summary
RNNs	Recurrent Neural Networks
SSE	Sum of the Squared Error
TF-IDF	Term Frequency-Inverse Document Frequency
Word2Vec	Word to Vector
XLNet	Transformer-XL Network

LIST OF FIGURES

Figure 2.1 Main elements of ruling number 8036 in the RulingBR dataset ... 14
Figure 2.2 Main elements of bill number 87 in BillSum dataset ... 16
Figure 2.3 BERT input representation ... 18

Figure 4.1 BB25HLegalSum overview ... 30
Figure 4.2 Bill 1104: Reference summary and highlighted bill according to the three strategies in a subsidiary manner. ... 38
Figure 4.3 Subsidiary highlighting in colors scheme ... 39
Figure 4.4 Intersectional highlighting in colors scheme ... 39

Figure 5.1 PoSum scores (US test Data) ... 44
Figure 5.2 FoSum scores (US test Data) ... 44
Figure 5.3 RoSum scores (US test Data) ... 45
Figure 5.4 PoSum scores (RulingBR) ... 46
Figure 5.5 FoSum scores (RulingBR) ... 46
Figure 5.6 RoSum scores (RulingBR) ... 47
Figure 5.7 Bill 1678: Reference summary and highlighted bill according to the three strategies. ... 50
Figure 5.8 Ruling 224: Reference summary and highlighted ruling according to the three strategies. ... 51

Figure A.1 Visão geral do método BB25HLegalSum ... 61

LIST OF TABLES

Table 3.1 Works proposing summarization solutions for legal documents.....................28

Table 4.1 Sentences and their similarity with *refSum* - Bill 1104 (US test data)............32
Table 4.2 Clusters from bill number 1104 within US test dataset 33
Table 4.3 Output scores from bill number 1104 within US dataset for top-20 sentences34

Table 5.1 Performance on US test data 42
Table 5.2 Performance on CA test data...43
Table 5.3 Performance on RulingBR test data 43

CONTENTS

1 INTRODUCTION........9
2 THEORETICAL FOUNDATION........12
2.1 Summarization approach........12
2.2 Structure of the legal documents........13
2.3 Embeddings (word embeddings, contextual embeddings)........15
2.4 BERT........17
2.5 BM25........18
2.6 Clustering........19
2.7 Evaluation metrics........20
2.7.1 Precision, Recall, and F1-Measure........20
2.7.2 ROUGE Evaluation Metrics........22
3 RELATED WORK........24
3.1 Legal document summarization techniques........24
3.2 Presentation of the summary........26
3.3 Readability assessment........27
3.4 Final considerations........28
4 FRAMEWORK FOR SUMMARIZATION OF LEGAL DOCUMENTS........29
4.1 Overview........29
4.2 Extracting BERT Unique Sentences........30
4.3 Scoring sentences and clusters........31
4.3.1 Cluster related sentences........32
4.3.2 Scoring Clusters and Sentences........33
4.4 Generation and Selection of Candidate Summaries........35
4.5 Highlighting and distinct showcasing summaries in the legal document........36
5 EXPERIMENTS........40
5.1 Research questions........40
5.2 Datasets and baselines........40
5.3 Experiment 1........41
5.3.1 Method........41
5.3.2 Results - BillSum........41
5.3.3 Results - RulingBR........42
5.4 Experiment 2........43
5.4.1 Method........43
5.4.2 Results and discussion - BillSum........44
5.4.3 Results and discussion - RulingBR........46
5.5 Experiment 3........47
5.5.1 Method........47
5.5.2 Results and Discussion - BillSum........49
5.5.3 Results and Discussion - RulingBR........52
6 CONCLUSIONS AND FUTURE WORK........54
REFERENCES........55
APPENDIX A — SUMÁRIO EXPANDIDO........59

1 INTRODUCTION

Ongoing legal proceedings are a common concern impacting legal systems across the globe. The quantity of unresolved cases can differ substantially based on the size of the population, the legal structure, and the accumulation of pending cases. While certain nations might have just a few thousand unsettled cases, others could have millions. This scenario motivates the research of computational techniques that can help accelerate judicial analysis, select similar cases for judging in batches, or identify patterns that could lead to better decision-making.

Implementing an automated system for highlighting key information in legal documents could significantly alleviate the burden on legal professionals, making their reading tasks more enjoyable and less arduous, potentially enhancing the efficiency in the judicial analysis process. The automatic summarization of legal documents to synthesize their essence is critical in this context.

The objective of automated text summarization is to produce summaries similar to those created by humans (ALLAHYARI et al., 2017). This proves to be a challenging task due to the intricacies and subtleties inherent in natural language. The algorithms for text summarization must take into account the target audience, the objective of the summary, and also the genre and layout of the original text. Text summarization finds utility in diverse applications, such as news aggregation, document management, and legal document summarization.

The majority of works in the legal domain employ extractive summarization for the creation of summaries, a concept elucidated in (ANAND; WAGH, 2019) as "the generation of a summary containing a sentence subset of the original text after identifying the important sentences". Several techniques were explored for extractive legal text summarization, including word relevance (POLSLEY; JHUNJHUNWALA; HUANG, 2016), graph-based ranking models (DALAL; SINGHAL; LALL, 2023; JAIN; BORAH; BISWAS, 2023a), statistical models (JAIN; BORAH; BISWAS, 2022; MERCHANT; PANDE, 2018), and deep learning (ANAND; WAGH, 2019). More recently, Bidirectional Encoder Representations from Transformers (BERT) (DEVLIN et al., 2018) has been leveraged in the legal area (FURNITUREWALA et al., 2021), inspired by state-of-the-art results achieved in general extractive text summarization (LIU, 2019).

An alternative strategy in the legal documents area is Best Match 25 (BM25), a ranking function commonly utilized in information retrieval to determine the similarity of a document concerning a search query (ROBERTSON; ZARAGOZA et al., 2009). The combined use of BERT and BM25 is recurrent for information retrieval in legal documents (ASKARI et al., 2022; ALTHAMMER et al., 2021); however, it is still in its preliminary stages for the legal document summarization task. The strengths of these techniques can be joined to yield high-quality summaries and assist in overcoming challenges associated with traditional methods, such as feature engineering and lengthy documents. BERT functions as a powerful language model that captures intricate relationships among words and sentences, whereas BM25 operates as an effective information retrieval algorithm for document ranking.

According to Jain, Borah e Biswas (2021), more analysis on the readability of the generated summaries, and how to present them is required. Within the legal area, summary presentation is addressed using highlighting (LICARI et al., 2023) and heatmaps (POLSLEY; JHUNJHUNWALA; HUANG, 2016) representing the relevance of sentences within the original document. However, the relevance of a sentence may be a secondary aspect for legal workers, who might be more interested in the main arguments within their context.

In this dissertation, we propose BB25HLegalSum (BERT + BM25 + Highlighting Legal Documents Summarization), a novel method for the extractive summarization of legal documents. It leverages BERT and BM25 to identify relevant sentences in a legal document and combine clusters of sentences to generate candidate summaries, which are selected using metrics against a reference summary. Our premise is that, for legal workers, the most important aspect of legal document summarization is the extraction of the most relevant arguments and the ability to identify their importance within a context. Hence, an existing reference summary may synthesize the document, but it does not necessarily provide all the useful information they need.

We generate summaries using three strategies to identify the best parts of a document, focused on the precision of the selected sentences, their coverage of the text (recall), and a trade-off between these two criteria. Another distinctive feature of the method is the presentation of the generated summary. We propose a highlighting approach that, by using different colors, represents the sentences contained in the summaries generated according to each strategy. In this way, the user can identify and distinguish in their original context the relevant sentences of the document according to distinct points of view that emphasize precision, coverage, or both. Preliminary results are discussed in (BONALUME; BECKER, 2023).

To assess the proposed method, our experiments aim to answer the following research questions:

- RQ1: How does the performance of BB25HLegalSum compare to baseline methods for legal document summarization?
- RQ2: How does the length of the reference summary impact the recall and precision of the generated summary?
- RQ3: Which type of summary is more suitable in the legal documents context concerning its readability: focused on precision, recall, or f-measure?

Our assessments revealed encouraging results. Our method outperformed baseline works in two legal datasets, namely BillSum (JAIN; BORAH; BISWAS, 2021) and RulingBR (FEIJÓ, 2021). Additionally, the length of the reference summary impacts the recall and precision of the generated summaries, with the proposed approach performing better for larger reference summaries. Finally, a qualitative experiment showed that, in a legal document context, completeness is the most important criterion to summarize, compared to conciseness, since it is more important overall to avoid missing relevant information. Thus, the highlighting with distinct colors enables to identify different types of information captured by each strategy.

The main contributions of this dissertation are:

- (1) a method that leverages BERT and BM25 to generate legal document summaries. It outperforms baselines (ANAND; WAGH, 2019; MIHALCEA; TARAU, 2004; ERKAN; RADEV, 2004; FEIJÓ, 2021) in two different legal datasets (BillSum and RulingBR);
- (2) a presentation method for the generated summaries using different colors that highlights in their original context the importance of sentences according to distinct points of view (precision vs. coverage). Legal workers positively assessed this presentation.

The remaining of this work is structured as follows. Chapter 2 presents the theoretical foundation necessary to understand concepts underlying our research. Chapter 3 reviews related work in this area. Chapter 4 details the proposed summarization method and enlightens its use. Chapter 5 describes the configuration, method, and results of the experiments performed. Finally, Chapter 6 draws conclusions and points out to future work.

2 THEORETICAL FOUNDATION

In this section, we discuss the automatic summarization approaches that are relevant to this work. We also examine the concept of embeddings (word, contextual) that is leveraged in this work. Additionally, we examine the readability assessment and summary presentation, and provide an overview of the metrics utilized to evaluate our summarization method.

2.1 Summarization approach

In automatic summarization, the summary is composed by concatenating the most important selected sentences or by paraphrasing (EL-KASSAS et al., 2021). Document summarization techniques can be broadly classified into two categories (NENKOVA; MCKEOWN et al., 2011): extractive and abstractive summarization. *Extractive summarization* involves selecting sentences or phrases from the original document that are deemed most important and representative of the content. These selected sentences are then combined to form a summary. In other words, extractive summarization form summaries by selecting and concatenating the most important spans in a document, typically sentences (LIU, 2019). Extractive methods often utilize statistical and machine learning approaches to identify salient sentences based on features such as sentence length, term frequency, and position in the document.

On the other hand, abstractive summarization methods generate novel words and phrases that do not feature in the source text (HUANG et al., 2020). They aim to produce summaries that are more concise and coherent, similar to what a human summarizer would create. There are some limitations to abstractive summarization usage in legal documents: legal documents are longer and have citations that cannot be ignored, as well as the fact that the meaning of the legal documents can be altered (JAIN; BORAH; BISWAS, 2021).

In this work, we chose extractive summarization for the legal documents summarizer due to its compatibility with the unique characteristics of legal texts. Legal documents are often highly technical, containing specific legal terminology. Extractive summarization, which involves selecting and rearranging existing sentences from the source document, allows for the preservation of the original wording, maintaining the precision and accuracy of the legal language. By retaining the original sentences, extractive summarization helps to maintain the legal context and integrity necessary for a comprehensive understanding of the legal documents.

2.2 Structure of the legal documents

There are some differences between dealing with a regular document and a legal one in terms of text summarization. These differences may play a major role in the summarization strategy. For example, there is little or no hierarchy in a general document of, say, news genre. Conversely, a legal document usually follows a structure that can not be ignored (KANAPALA; PAL; PAMULA, 2019). As an example of the structure of legal documents, we will examine the structure of legal documents in the datasets RulingBR and BillSum.

RulingBR is the largest Brazilian dataset containing Brazil's Supreme Court (STF) decisions as the main source (FEIJÓ; MOREIRA, 2018). It is composed of 10574 rulings, each one of them having a summary ("ementa"), a report, a vote and a judment section. As we can see in Figure 2.1, the summary section contains the main topics discussed in each case and how the judges have decided. The report section is a compilation of the main arguments and events that happened during the trial. The vote section may contain one or more votes. The judgment section is, in general, short and compiles the outcome as granted or denied (FEIJO; MOREIRA, 2019).

In general, one can ignore references/citations in text summarization, but that may not be possible in the case of legal texts (KANAPALA; PAL; PAMULA, 2019). Therefore, anything that is said in a legal document can be considered crucial and should not be discarded right away. One must analyze the purpose of the summarization task at hand to decide which parts should be provided as input to produce a system summary. For example, if we were to produce summaries for the legal document in the Brazilian Supreme Court structure with the aim of summarizing only the main arguments, we could opt out the judgment section from the text summarization treatment.

The structure of the legal documents may differ from country to country, so whenever a text summarization solution is used within a specific legal documents dataset, its structure should be taken into consideration.

BillSum is a dataset of legislative bills, composed by a title, a text (legal description), and a summary for each bill. According to (KORNILOVA; EIDELMAN, 2019), it is the first dataset for summarization of US Congressional and California state bills.

Figure 2.1: Main elements of ruling number 8036 in the RulingBR dataset

Structure element	**Content**
Summary	Agravo regimental em recurso extraordinário com agravo. 2. Direito do Trabalho. 3. Jornada de trabalho de advogado empregado. 4. Revolvimento do acervo fático-probatório dos autos. Súmula 279 do STF. 5. Matéria infraconstitucional. Ofensa à Constituição, se existente, seria reflexa. 6. Agravo regimental a que se nega provimento.
Report	O SENHOR MINISTRO GILMAR MENDES (RELATOR): Trata-se de agravo regimental interposto contra decisão que conheceu do agravo e negou provimento ao recurso extraordinário, com base na jurisprudência do Supremo Tribunal Federal. Eis um trecho dessa decisão: "Ademais, no que diz respeito à apontada violação do inciso XIII do artigo 7º da CF, melhor sorte não assiste a parte recorrente, pois o tribunal de origem assentou que a carga horária semanal de trabalho do advogado empregado demanda a interpretação do Regulamento Geral da OAB. Assim, para acolher a pretensão recursal e superar o entendimento do acórdão recorrido seria necessário a análise da legislação infraconstitucional de regência assim como o reexame do conjunto fático-probatório constante dos autos, o que inviabiliza o prosseguimento do recurso extraordinário, pois incide o óbice da Súmula 279 do STF". (eDOC 35) No agravo regimental, sustenta-se a não incidência da Súmula 279 desta Corte, ao argumento de que a parte agravada não foi condenada com base no acervo fático-probatório dos autos. Alega-se que, não obstante a possibilidade de a jornada de trabalho exceder as vinte horas semanais tratar-se de exceção, a contratação do advogado com dedicação exclusiva deve obedecer ao artigo 7º, XIII, da Constituição Federal, que limita a referida jornada a 8 horas diárias e 44 semanais. Nesses termos, afirma-se que apenas a hora que ultrapassar esse tempo poderia ser considerada como extra. Aduz ainda que o Estatuto da OAB, em seu artigo 20, não faz nenhuma alusão à jornada de cinco dias na semana (ou quarenta horas) para os casos ali excepcionados (acordo ou convenção coletiva ou em caso de dedicação exclusiva). É o relatório.
Vote	O SENHOR MINISTRO GILMAR MENDES (RELATOR): No agravo regimental, não ficou demonstrado o desacerto da decisão agravada. Verifico que as alegações da parte são impertinentes e decorrem de mero inconformismo com a decisão adotada por este Tribunal, uma vez que a parte agravante não trouxe argumentos suficientes a infirmá-la, visando apenas à rediscussão da matéria já decidida de acordo com a jurisprudência pacífica desta Corte. Conforme consignado na decisão agravada, o Tribunal de origem decidiu acerca da jornada de trabalho de advogado empregado, no caso dos autos, com base nos fatos e provas dos autos, bem com na legislação infraconstitucional aplicável. Assim, divergir desse entendimento demandaria o reexame do conjunto fático-probatório, providência vedada na via extraordinária, a teor do disposto na Súmula 279 desta Corte. Além disso, a matéria debatida no Tribunal de origem restringe-se ao âmbito infraconstitucional, de modo que a ofensa à Constituição, se existente, seria reflexa ou indireta, o que inviabiliza o processamento do presente recurso. Nesse sentido, destaco o seguinte precedente: "AGRAVO REGIMENTAL NO RECURSO EXTRAORDINÁRIO. ADVOGADO EMPREGADO. JORNADA DE TRABALHO. LEI N. 8.906/94 E MP 1.522/96. AUSÊNCIA DE PREQUESTIONAMENTO DOS ARTIGOS 5º, XXI, E 7º, XII E XXVI, DA CONSTITUIÇÃO. SÚMULA N. 282 DO SUPREMO TRIBUNAL FEDERAL. MATÉRIA INFRACONSTITUCIONAL. ALEGAÇÃO DE AFRONTA AO ART. 5º, XXXVI, DA CF. OFENSA INDIRETA. REEXAME DE FATOS E PROVAS. IMPOSSIBILIDADE EM RECURSO EXTRAORDINÁRIO. INCIDÊNCIA DA SÚMULA N. 279 DO STF (...)". (RE-AgR 610.184, Rel. Min. Luiz Fux, Primeira Turma, DJe 17.6.2011) Ante o exposto, nego provimento ao agravo regimental.
Judgment	Vistos, relatados e discutidos estes autos, acordam os ministros do Supremo Tribunal Federal, em Segunda Turma, sob a presidência do ministro Dias Toffoli, na conformidade da ata de julgamento e das notas taquigráficas, por unanimidade, negar provimento ao agravo regimental, nos termos do voto do Relator.

Figure 2.2 displays an example of document within BillSum dataset. The structure of bills is much more simple compared to legal documents of Brazilian STF. It is composed of three elements: title, summary and text. In whichever dataset we work with, there should at least one text to be used as reference and another text document that should be used as candidate summary.

2.3 Embeddings (word embeddings, contextual embeddings)

This section explores the evolution from word embeddings to contextual embeddings.

Word embeddings, such as GloVe (Global Vectors for Word Representation), have revolutionized natural language processing (NLP) tasks by representing words as fixed-dimensional vectors based on co-occurrence statistics. These embeddings assign fixed representations to words regardless of context, in the sense that all senses of a polysemous word have to share the same representation (ETHAYARAJH, 2019).

For example, GloVe is a global log-bilinear regression model for the unsupervised learning of word representations (PENNINGTON; SOCHER; MANNING, 2014). It constructs word embeddings by analyzing the co-occurrence statistics of words in a large corpus. By considering the overall occurrence patterns of words, GloVe generates embeddings that capture semantic similarities and differences between words. For instance, it captures analogy relationships such as "man is to woman as king is to queen" (PENNINGTON; SOCHER; MANNING, 2014).

Word embeddings capture relationships between words, enabling various applications. However, they treat each word as an independent entity and fail to capture the contextual nuances of language. This limitation manifests in several ways. Firstly, traditional word embeddings assign a single vector representation to each word, regardless of its context, thereby failing to capture the distinct meanings: polysemic words must share a single vector, which is a problem in word embeddings (ETHAYARAJH, 2019).

To overcome the limitations of traditional word embeddings, researchers have developed contextual word embeddings that generate word representations based on the surrounding words and their relationships in a given sentence throughout the document. Two prominent approaches have emerged: the first approach is based on RNNs (Recurrent Neural Networks) - they are able to capture the information about the past occurrences and take that into consideration when processing the current input (ANAND; WAGH,

Figure 2.2: Main elements of bill number 87 in BillSum dataset

Structure element	**Content**
Summary	Railroad Hours of Service Act of 2010 - Extends railroad hours of services requirements and limitations to cover yardmaster employees who supervise and coordinate the activities of workers engaged in railroad traffic operations, including making up or breaking up trains and switching inbound or outbound traffic. Revises the prohibition against a railroad carrier's requiring or allowing a train employee to remain or go on duty unless that employee has had at least 10 consecutive hours off duty during the prior 24 hours. Prohibits requiring or allowing an employee from initiating an on duty period unless the employee has had at least 10 consecutive hours off duty immediately prior to going on duty. Directs the Secretary of Transportation (DOT) to prescribe regulations to: (1) require all deadhead transportation in excess of a specific number of hours to be counted as time on duty. And (2) reset the calendar day clock. Revises the rule that an interim period available for at least 4 hours rest at a place with suitable facilities for food and lodging is not time on duty. Repeals the current list of causes for prevention of a return to duty. Requires a train employee to be notified before going off duty whether such period off duty is an interim release. Prohibits a railroad carrier from requiring or allowing an employee to exceed 2 hours in deadhead transportation per each tour of duty. Revises the limitations on the duty hours of signal employees. Specifies that time on duty spent performing any service for the railroad carrier during a 24-hour period in which the employee is engaged in installing, repairing, or maintaining signal systems includes all work where there is a potential to interact or otherwise come into contact with safety-critical devices or circuits. Treats as service covered by hours of duty limitations the operation by signal employees of motor vehicles requiring a commercial driver's license while on duty. Extends to yardmaster employees certain limitations on the duty hours of dispatching service employees. Declares that all commingle service involving yardmaster service and dispatcher service mixing with freight service shall be covered by the limitations on the duty hours of signal employees. Extends to yardmaster employees, when an emergency exists, the same limitation that applies to the hours of dispatching service employees in an emergency.
Title	To amend title 49, United States Code, with respect to hours of service rules for railroad employees.
Text	SECTION 1. SHORT TITLE. This Act may be cited as the "Railroad Hours of Service Act of 2010". SEC. 2. REDESIGNATIONS. Chapter 211 of title 49, United States Code, is amended by redesignating sections 21101 through 21109 as sections 21102 through 21110, respectively. SEC. 3. PURPOSE. Chapter 211 of title 49, United States Code, is further amended by inserting before section 21102 (as so redesignated by section 2 of this Act) the following: "Sec. 21101. Purpose "Railroad employees covered by this chapter shall be provided predictable and defined work and rest periods.". SEC. 4. DEFINITIONS. Section 21102 (as so redesignated by section 2 of this Act) of chapter 211 of title 49, United States Code, is amended– (1) in paragraph (5), by inserting "and yardmaster employee" before the period; and (2) by adding at the end the following: "(6) 'yardmaster employee' means an employee who supervises and coordinates the activities of workers engaged in railroad traffic operations, including making up or breaking up trains and switching inbound or outbound traffic.". SEC. 5. NONAPPLICATION, EXEMPTION, AND ALTERNATE HOURS OF SERVICE REGIME. Section 21103(c) (as so redesignated by section 2 of this Act) of chapter 211 of title 49, United States Code is amended– (1) in paragraph (1)(A), by striking "21109(b)" and inserting "21110(b)"; (2) in paragraph (3), by striking "21109(b)" and inserting "21110(b)"; (3) by striking subparagraph (C) of paragraph (4); (4) by redesignating subparagraph (D) of paragraph (4) as subparagraph (B); and (5) by striking "new section 21103" each place it appears and inserting "section 21104". SEC. 6. LIMITATIONS ON DUTY HOURS OF TRAIN EMPLOYEES. Section 21104 (as so redesignated by section 2 of this Act) of chapter 211 of title 49, United States Code, is amended– (1) in subsection (a)– (A) in paragraph (3)– (i) by striking "remain or go on duty unless" and inserting "initiate an on duty period unless"; and (ii) by striking "during the prior 24 hours; or" and inserting "immediately prior to going on duty; or"; (B) in paragraph (4)(A)– (i) in clause (i), by striking "work" and inserting "initiate an on duty period"; and (ii) in clause (ii), by striking "works" and inserting "initiates an on duty period on"; and (C) in the matter after paragraph (4) by inserting "For purposes of paragraph (4)(A) and (B), within 12 months after the date of enactment of the Railroad Hours of Service Act of 2010, the Secretary shall prescribe regulations to require all deadhead transportation in excess of a specific number of hours to be counted as time on duty and shall reset the calendar day clock." before "The Secretary may waive"; (2) in subsection (b)(7), by striking "when the employee is prevented" and all that follows through "employee left the designated terminal." and inserting ". A train employee shall be notified before going off duty whether such period off duty is an interim release."; and (3) in subsection (c)(1)– (A) in subparagraph (A)(ii), by striking "and" at the end; (B) in subparagraph (B)(ii), by striking "21109." and inserting "21110; and"; and (C) by adding at the end the following new subparagraph: "(C) to exceed 2 hours in deadhead transportation per each tour of duty.". SEC. 7. LIMITATIONS ON DUTY HOURS OF SIGNAL EMPLOYEES. Section 21105 (as so redesignated by section 2 of this Act) of chapter 211 of title 49, United States Code, is amended– (1) in subsection (b)(2), by inserting ", including all work where there is a potential to interact or otherwise come into contact with safety-critical devices or circuits," before "is time on duty"; (2) in subsection (e), by adding at the end the following: "Signal employees operating motor vehicles requiring a commercial driver's license while on duty shall be considered covered service."; and (3) by adding at the end the following new subsection: "(f) Safety-Critical Devices or Circuits.–Time on duty shall include all work where there is a potential to interact or otherwise come into contact with safety-critical devices or circuits.". SEC. 8. LIMITATIONS ON DUTY HOURS OF DISPATCHING SERVICE EMPLOYEES AND YARDMASTER EMPLOYEES. Section 21106 (as so redesignated by section 2 of this Act) of chapter 211 of title 49, United States Code, is amended– (1) in the section heading by inserting "and yardmaster employees" after "service employees"; (2) in subsection (a)– (A) by striking "21103 or 21104" and inserting "21104 or 21105"; and (B) by inserting "or yardmaster employee" after "service employee"; (3) in subsection (b), by inserting "or yardmaster employee" after "a dispatching service employee"; (4) in subsection (c), by adding at the end the following: "All commingle service involving yardmaster service and dispatcher service mixing with freight service shall be covered under the provisions of section 21104."; and (5) in subsection (d), by inserting "or yardmaster employee" after "dispatching service employee". SEC. 9. CLERICAL AMENDMENT. Chapter 211 of title 49, United States Code, is further amended by amending the table of sections at the beginning of the chapter to read as follows: "Sec. "21101. Purpose. "21102. Definitions. "21103. Nonapplication, exemption, and alternate hours of service regime. "21104. Limitations on duty hours of train employees. "21105. Limitations on duty hours of signal employees. "21106. Limitations on duty hours of dispatching service employees and yardmaster employees. "21107. Limitations on employee sleeping quarters. "21108. Maximum duty hours and subjects of collective bargaining. "21109. Pilot projects. "21110. Regulatory authority.".

2019), such as LSTM (Long Short-Term Memory). When the RNN tries to learn from the past words, the information from the earlier words can gradually get lost or "vanish" as it goes deeper into the network, this is the vanishing error problem that is solved by the LSTM (STAUDEMEYER; MORRIS, 2019). These models process words sequentially, updating their hidden states at each step. The final hidden state represents the word's contextual information, allowing the model to capture the context-dependent meaning of words.

The second approach revolutionized contextual embeddings, popularized by transformer models such as BERT (DEVLIN et al., 2018) and GPT (Generative Pre-Training Transformer) (FLORIDI; CHIRIATTI, 2020). These models employ self-attention mechanisms to capture global dependencies within a sequence of words, creating rich contextual representations for words.

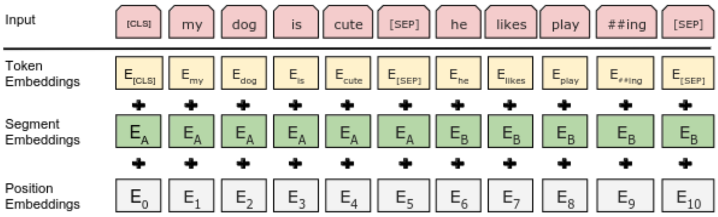
Contextual word embeddings offer several benefits that enhance language understanding and NLP applications (NASEEM et al., 2020). Firstly, they provide a more nuanced understanding of language by capturing fine-grained contextual information, such as polysemy. This enables better performance in various NLP tasks such as sentiment analysis, question answering, and named entity recognition. Secondly, contextual embeddings facilitate disambiguation of polysemous and homonymous words. For example, the word 'bad' in the sentences 'this whole crew has rocked thru bad weather' and 'giving up on (company name). #badservice' has different representations when using contextual embeddings. By assigning different representations based on specific contexts, these embeddings capture the distinct meanings, enhancing the precision and accuracy of language understanding. Furthermore, contextual embeddings handle rare and out-of-vocabulary words more effectively. By leveraging contextual cues, these embeddings can infer the meaning of such words, even if they were not encountered during training.

In this work, our proposed solution uses contextual word embeddings, more specifically, BERT.

2.4 BERT

BERT (DEVLIN et al., 2018) is a groundbreaking language representation model that has revolutionized NLP tasks. BERT's primary innovation lies in its ability to capture contextual information by considering both the left and right context of a word during training, unlike previous models that relied on unidirectional context. Therefore, during

Figure 2.3: BERT input representation



Source: (DEVLIN et al., 2018)

the training process, BERT takes into account the entire sentence or document to create a rich representation for each word, considering the surrounding words on both sides.

To utilize a pre-trained BERT model, it is essential to structure the input data following a specific format, as depicted in Figure 2.3. In this arrangement, the text should be segmented into tokens. For each data instance, the sentence's start should be indicated with a [CLS] tag. Subsequent sentences are demarcated with the [SEP] tag. Beyond the input sequence, Figure 2.3 introduces the following components:

- Token Embeddings: These are responsible for converting each word into consistent vector representations with 768 dimensions.
- Segment Embeddings: They discern input sequences based on separation using [CLS] and [SEP] tags.
- Position Embeddings: This layer assigns distinct representation vectors to the same token. It relies on information from the attention mechanism, which retains pertinent details during training steps. This assists in recognizing the various contexts a term has been utilized in, thereby aiding in selecting the optimal representation for the term within the analyzed context.

It is worth noting that BERT models vary in terms of architecture and training data. In this work we have used BERT multilingual uncased as the chosen model, since it has demonstrated good performance at zero-shot cross-lingual model transfer (PIRES; SCHLINGER; GARRETTE, 2019).

2.5 BM25

BM25 (ROBERTSON; ZARAGOZA et al., 2009) is a well-established information retrieval algorithm that ranks documents based on their similarity concerning a query.

It can be used in document ranking, content recommendation, question-answering system, legal document summarization. Furthermore, its usage transcends the conventional query-document relationship, finding application in tasks such as query expansion (AKLOUCHE; BOUNHAS; SLIMANI, 2019).

The core elements that define BM25's functionality include:

- Term Frequency Saturation: BM25 introduces a saturation term to mitigate the impact of excessively high term frequencies. This addresses the problem of term frequency dominance and ensures that document ranking is influenced by both term presence and frequency while preventing over-amplification of frequent terms.
- Inverse Document Frequency (IDF): Tuning BM25 refines the IDF component by introducing a freely adjustable parameter to control its impact. This adjustable parameter allows practitioners to fine-tune the algorithm to suit different retrieval scenarios, making it adaptable to varying collections and user preferences.
- Length Normalization: To account for document length discrepancies, BM25 applies length normalization to document scores. This normalization factor counteracts the bias towards shorter documents and ensures that document ranking is not skewed by length-related factors.

The combined use of BERT and BM25 is recurrent for document retrieval in the Competition on Legal Information Extraction/Entailment (COLIEE) (ASKARI et al., 2022; ROSA et al., 2021; ALTHAMMER et al., 2021), but its potential has not been fully examined for legal document summarization. The resulting summarization model can benefit from the strengths of both approaches to produce high-quality summaries and help to overcome some of the traditional methods' hurdles, such as the reliance on feature engineering and the difficulty in handling long documents.

2.6 Clustering

Different clustering approaches have been proposed, each of which uses a different inclusion principle (SAXENA et al., 2017). Popular clustering techniques include hierarchical, density-based, and partition-based techniques.

Hierarchical is either diviside (top-down approach, by breaking up a cluster containing all objects into smaller clusters) or agglomerative (bottom-up approach, starting with single object and then merging these atomic clusters into larger clusters) (SAXENA et al., 2017).

Density-based clustering defines clusters as sets of data objects that are spread across contiguous regions of high object density (KRIEGEL et al., 2011). These clusters are distinguished by areas of low density that separate them from one another.

Among the partition-based techniques, K-means is very popular. The main idea of K-means is to define k centroids, one for each cluster (BOOMIJA; PHIL, 2008). Ideally, the centroids should be far away from each other. The next step is to take each point belonging to a given data set and associate it to the cluster of the nearest centroid. This is done in a loop in order to minimize the squared error function (VELMURUGAN; SANTHANAM, 2011).

We chose K-means as the clustering approach, since it is easy to implement and yielded good results. However, one of the challenges of K-means clustering is to find the appropriate value for K. To do that, we used silhouette scores, which evaluate the cohesion and separation of data points within clusters (THINSUNGNOENA et al., 2015). Specifically, for a given entity i belonging to a cluster, its silhouette width is calculated as the difference between the average distance to its own cluster members and the minimum average distance to members of other clusters (KODINARIYA; MAKWANA et al., 2013). The resulting silhouette width ranges from -1 to 1. A value near 0 indicates potential suitability for another cluster, while nearing -1 implies misclassification, and values close to 1 signify effective clustering of the dataset. The scores of individual aspects can be aggregated to assess a cluster, as well as the clustering (i.e. set of clusters). By using the sklearn.metrics silhouette_score function, we varied the value of k, clustering with the highest silhouette score.

2.7 Evaluation metrics

2.7.1 Precision, Recall, and F1-Measure

F1-measure, precision, and recall are fundamental metrics used in classification tasks and information retrieval. These metrics are described in (KANAPALA; PAL; PAMULA, 2019) and are widely employed in text summarization, sentiment analysis, document classification tasks, among others. The F1-measure is a harmonic mean of precision and recall, providing a balanced evaluation of a system's performance. Precision measures the proportion of correctly identified positive instances among all instances identified as positive, while recall measures the proportion of correctly identified positive instances out of all actual positive instances. In other words, Recall reflects informedness (which, in our work is referred to as completeness) and Precision reflects markedness (POWERS, 2011) (which, in our work is described as conciseness).

- **Precision**: it measures the proportion of correctly generated summary content compared to the total content in the generated summary. It focuses on the relevancy and accuracy of the information included in the summary. A high precision score indicates that the generated summary contains mostly relevant and accurate information from the source text. In text summarization, precision helps to ensure that the summary captures the most important and meaningful content while minimizing irrelevant or misleading information. It is calculated with the following Formula 2.1.

$$P = \#relevant\ items\ retrieved/\#retrieved\ items \tag{2.1}$$

- **Recall**: it measures the proportion of relevant content from the source text that is correctly included in the generated summary. It focuses on the completeness and coverage of the summary. A high recall score indicates that the generated summary captures a significant amount of relevant information from the source text. In text summarization, recall helps to ensure that important details and key points are not missed or omitted in the summary. Its calculation is described by Equation 2.2.

$$R = \#relevant\ items\ retrieved/\#relevant\ items \tag{2.2}$$

- **F1-measure**: it combines precision and recall into a single metric to provide a balanced evaluation of the summary's quality. It is the harmonic mean of precision and recall and provides a comprehensive assessment of the trade-off between precision and recall. F1-measure is particularly useful when both precision and recall are equally important in the context of text summarization. It provides a single value that reflects the overall effectiveness of the generated summary, considering both the accuracy and completeness of the information included. It is calculated according to Equation 2.3.

$$F1 = (2 * (P * R))/(P + R) \tag{2.3}$$

2.7.2 ROUGE Evaluation Metrics

ROUGE is a widely employed metric for assessing the quality of text summarization systems. It measures the overlap between the generated summaries and the reference summaries to determine the quality of a summary by comparing it to other (ideal) summaries created by humans (LIN, 2004). ROUGE has become a standard evaluation measure, enabling objective comparisons between system-generated summaries and human-created summaries.

The underlying assumption of ROUGE is that a good summary should contain essential information present in the reference summaries. The metric calculates various n-gram-based statistics, such as ROUGE-N, which evaluates the overlap of n-grams between the generated and reference summaries.

ROUGE-1 measures the unigram overlap between the reference summary and the generated summary. It calculates the precision, recall, and F1-score by considering the percentage of unigrams (individual words) in the generated summary that are also present in the reference summary. The precision represents the ratio of common unigrams to the total number of unigrams in the generated summary, while recall represents the ratio of common unigrams to the total number of unigrams in the reference summary. The F1-score is the harmonic mean of precision and recall, providing a balanced measure of the two.

ROUGE-2 extends the concept of ROUGE-1 to bigrams. It focuses on the overlap of consecutive word pairs (bigrams) between the reference and generated summaries. Similar to ROUGE-1, ROUGE-2 computes precision, recall, and F1-score using the ratio of common bigrams to the total number of bigrams in the generated and reference summaries. By considering the arrangement and co-occurrence of words, ROUGE-2 provides a more comprehensive evaluation of the summarization quality.

ROUGE-L evaluates the longest common subsequence (LCS) between the reference and generated summaries. Instead of considering individual words or pairs of words, it captures the longest sequence of words that appears in the same order in both summaries. This metric is particularly useful in handling paraphrases and minor variations in wording. ROUGE-L calculates precision, recall, and F1-score based on the length of the LCS and the lengths of the reference and generated summaries. It emphasizes content overlap while being more tolerant of word variations and reordering.

For all these metrics, values closer to 1 indicate better performance.

In conclusion, ROUGE scores serve as an indicator of similarity based on shared words in the form of n-grams or word sequences. This gives insight into how well the automated summary captures relevant information.

3 RELATED WORK

This chapter describes the works in legal documents summarization, in order to present a better idea of the landscape that pertains this research.

3.1 Legal document summarization techniques

Legal document summarization has explored various techniques. In the field of legal document summarization, two primary approaches are widely used: extractive summarization and abstractive summarization.

Extractive summarization forms summaries by selecting and concatenating the most important spans (typically sentences) in a document (LIU, 2019). This method relies on NLP techniques to determine the most salient information by ranking sentences based on their importance and relevance to the main content. The resulting summary retains the exact wording and phrasing from the source document, making it a more objective representation of the original content.

CaseSummarizer (POLSLEY; JHUNJHUNWALA; HUANG, 2016) uses extractive summarization and combines standard summary methods based on word relevance (TF-IDF) with domain-specific knowledge to summarize legal documents with highlighted multicolor sentences based on their scores.

Graph-based ranking models explore the relationships and similarities between nodes representing the text to select the relevant portions of legal documents. In this sense, (JAIN; BORAH; BISWAS, 2023a) propose a graph-based ranking with TextRank along with extractive summarization, BERT and bayesian optimization.

Licari et al. (2023) approach is done in an extractive way by using BERT to extract sentences within the Italian legal domain and highlighting the top 5 scoring sentences with different colors.

KLSumm is based on the Kullback-Leibler (KL) divergence, a measure of how one probability distribution differs from another (JAIN; BORAH; BISWAS, 2022). In this algorithm, documents and candidate summaries are represented using probability distributions. The goal is to select in an extractive way sentences for the summary that minimize the KL divergence between the document and the candidate summary.

DCESumm (JAIN; BORAH; BISWAS, 2023b) uses a score sentencing system which bases the score of a sentence within the score of the cluster that sentence is located at, without mentioning specified summary presentation.

On the other hand, abstractive summarization, done by works such as (DALAL; SINGHAL; LALL, 2023; FEIJO; MOREIRA, 2019; MORO et al., 2023; AN et al., 2021), involves generating new sentences that capture the essential meaning of the legal document, rather than relying solely on existing content. This allows for the creation of more concise and readable summaries that may not mirror the exact wording of the source text.

A Graph-based ranking model that uses abstractive summarization is LexRank (DALAL; SINGHAL; LALL, 2023), using PEGASUS transformer model and presenting its summary in a textual manner.

Another way of summarizing is by using chunking and Transformers with contextual embeddings (FEIJO; MOREIRA, 2019). Chunking is a way of breaking down a large piece of text, like a document or a paragraph, into smaller, more manageable parts. This approach presents textual summaries with highlighted sentences in a single color, abstractively generating text summaries from legal decisions. The input source document is split into smaller chunks, which are then passed through a Transformer, which generates a summary. Then, each chunk-summary pair is submitted to a set of BERT models that output scores, keeping the highest scores summaries.

Most current summarization systems make summaries based only on the source document's content. However, even humans often need examples or references to fully grasp a document's meaning and write summaries in a certain style. Incorporating high-quality examples into summarization systems is a challenge. In this context, An et al. (2021) propose a solution that includes a dense Retriever and a Summarizer. Retrieved examples provide additional knowledge and guide how to write in a specific way and this is done with abstractive summarization.

There are reasons, however, not to use abstractive summarization in the legal documents context. Legal texts often contain precise language and critical details, making it crucial to preserve the original wording accurately, which extractive methods achieve by directly lifting relevant sentences. Secondly, extractive summarization ensures the retention of legal terminologies and technical jargon, which are paramount for maintaining the document's legal accuracy and integrity. Thirdly, in the legal domain, alterations to the phrasing can lead to significant changes in meaning (JAIN; BORAH; BISWAS, 2021), raising the risk of misrepresentation—an issue minimized by extractive methods. Therefore, when legal document summarization prioritizes fidelity to the original text and the preservation of legal nuances, extractive summarization can be deemed more favorable than abstractive techniques. Similarly, this caution may extend to other fields dealing with people's lives, such as medicine. Precision in language and the accurate representation of medical information are crucial in preserving the integrity of documents in these domains, making extractive summarization a prudent choice.

State-of-the-art has been achieved by using pre-trained models such as BERT (JAIN; BORAH; BISWAS, 2023b; LICARI et al., 2023; JAIN; BORAH; BISWAS, 2023a; FEIJO; MOREIRA, 2019; AN et al., 2021), which capture complex relationships between words and sentence. Yet, the aforementioned works do not combine methods such as BERT and BM25, lacking the chance to improve specially their precision scores. Jain, Borah e Biswas (2023a) conclude that due to the fixed length summary generation step in their work, some low scoring sentences are also getting included in the predictions, resulting in lower ROUGE scores.

3.2 Presentation of the summary

The focus in some works is the presentation of the generated legal summary in a highlighted way. Licari et al. (2023) use different colors to highlight the top 5 scoring sentences, and Polsley, Jhunjhunwala e Huang (2016) propose a heatmap to distinguish the importance of sentences. However, the relevance of a sentence may be a secondary aspect for legal workers, given that they generally seek the key arguments within a legal document. Hence, placing relevant sentences in a context is vital in the legal area. In this sense, clustering is important, since it enables the coloring of sentences within similar (or, perhaps, not different) contexts.

By creating and presenting multiple summaries for a single legal document, multiple colors subsidiary highlighting is a way to tackle two gaps in legal documents summarization: (1) the need for tailored summaries according to a user need and (2) to ensure fairness and reduce human bias in legal domain.

Concerning the first gap, depending upon the type of end user, the summary need might be different - for example, a judge might be more interested in finding out judicial decision summaries, whereas a lawyer might be more interested in finding the factual summary of a legal case document (JAIN; BORAH; BISWAS, 2021).

Regarding the second gap, one of the challenges associated with the automatic summarization of legal documents is related to its fairness. If there is only one summary, it might not cover all the aspects of the legal documents. Moreover, since reference summaries are human generated and abstractive in nature, it is very much prone to high bias. To deal with such problems, multiple reference summaries need to be considered for a single document, while building benchmark datasets (JAIN; BORAH; BISWAS, 2021).

It is important to highlight that the proposed concepts of intersectional and subsidiary highlighting aren't restricted solely to clustering; instead, they have relevance across a range of algorithms used in text summarization. This adaptability extends to scenarios where at least two distinct criteria are utilized to create a summary, such as distinguishing between precision-focused and recall-focused summaries. In this regard, it's theoretically feasible to incorporate subsidiary and intersectional highlighting into various methods like TF-IDF, graph-based approaches, probabilistic methods, TextRank, and more.

On the other hand, by showing the summaries in a textual way, the works with state-of-the-art results in applications that focus in legal documents summarization, such as neural networks (ANAND; WAGH, 2019), deep clustering (JAIN; BORAH; BISWAS, 2023b), or efficient memory-enhanced transformer-based architecture (MORO et al., 2023), do not tackle the limitations that arise from (1) not highlighting (therefore, omitting crucial information, which can have significant consequences in a judicial scenario), (2) in a subsidiary way (missing opportunities to showcase different arguments more effectively) (3) or in a contextual manner (it is important for a lawyer to be able to distinguish different kinds of arguments inside a legal document).

3.3 Readability assessment

The quality of generated summaries is typically assessed by comparing the generated summary against some reference summary using ROUGE (POLSLEY; JHUNJHUNWALA; HUANG, 2016; JAIN; BORAH; BISWAS, 2023b; DALAL; SINGHAL; LALL, 2023).

Moro et al. (2023) divide long documents into smaller parts (chunks), and the model remembers and compares these chunks afterwars. This way, it can understand the whole document without using too much memory. Through this manner, a textual summary is created by using chunking and BERT in an abstractive way with a readability assessment performed by legal workers.

In the context of ROUGE, recall refers to how much of the reference summary is captured in the system summary, precision measures how much of the system summary is relevant, and F1 combines recall and precision. Although having works with a readability assessment, such as (MORO et al., 2023), necessary assessments on legal text summarization remain unaddressed, such as properties of the readability of the summaries (e.g., the trade-off between conciseness and completeness) and the relationship between performance efficiency and reference summaries, typically used as the gold standard to evaluate the proposed summary systems (JAIN; BORAH; BISWAS, 2021).

3.4 Final considerations

Table 3.1 summarizes the described studies providing summarization solutions in the legal documents area.

Table 3.1: Works proposing summarization solutions for legal documents

Study	Summarization approach	Techniques	Embeddings	Readability assessment	Summary presentation
CaseSummarizer (POLSLEY; JHUNJHUNWALA; HUANG, 2016)	Extractive	TF-IDF + Domain specific	Non contextual		Highlighted (multicolor) colors given according to sentence scores
Bayesian Optimization based Score Fusion of Linguistic Approaches... (JAIN; BORAH; BISWAS, 2023a)	Extractive	BERT + Bayesian optimization	Contextual		Textual
Improving Kullback-Leibler based legal document summarization using enhanced text representation (JAIN; BORAH; BISWAS, 2022)	Extractive	BERT + Kullback-Leibler	Contextual		Textual
LexRank (DALAL; SINGHAL; LALL, 2023)	Abstractive	LexRank algorithm + PEGASUS transformer	Contextual		Textual
A sentence is known by the company it keeps (JAIN; BORAH; BISWAS, 2023b)	Extractive	Sentence and cluster scores (BERT)	Contextual		Textual
EMMA (MORO et al., 2023)	Abstractive	Chunking+BERT	Contextual	✓	Textual
RetrievalSum (AN et al., 2021)	Abstractive	BERT/BERT Retriever + Summarizer	Contextual	✓	Textual
Legal Holding Extraction from Italian Case Documents... (LICARI et al., 2023)	Extractive	Italian LEGAL-BERT	Contextual	✓	Highlighted (multicolor) colors given according to sentence scores
LegalSumm (FEIJO; MOREIRA, 2019)	Abstractive	Chunking + BERT	Contextual	✓	Highlighted (one color)
Our work	Extractive	Clustering+ Sentence and cluster scores (BERT) +BM25	Contextual	✓	Highlighted (multicolor) colors given in a subsidiary manner according to 3 different criteria

Our work contributes with a solution that leverages BERT and BM25 to produce legal document summaries, and with a method for presenting the generated summaries using highlighting that enables the examination of the trade-off between conciseness and completeness for readability of legal documents summaries.

4 FRAMEWORK FOR SUMMARIZATION OF LEGAL DOCUMENTS

This chapter presents the proposed framework for the summarization of legal documents. In the following sections, we first provide an overview of the framework, highlighting its main contributions. Then, each component is presented in detail. Finally, we illustrate how the proposed metrics could be used to summarize a legal document.

4.1 Overview

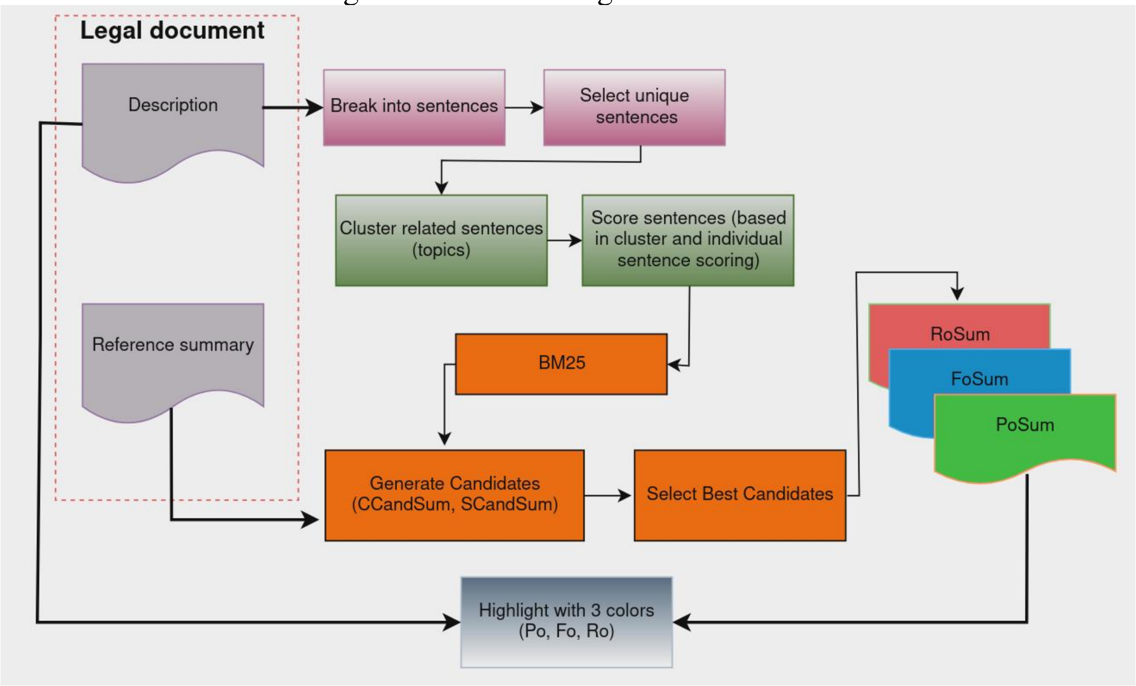
BB25HLegalSum presents a novel approach to summarizing legal documents. Consider a legal document *D* comprising a legal description (*desc*) and a reference summary (*refSum*). The primary objective is to select essential sentences from desc and combine them into a generated summary, referred to as *GSum*. The core premise is to identify the most crucial arguments in a way that helps legal professionals. Our premise is that, for legal workers, the most important aspect of legal document summarization is the extraction of the most relevant arguments and the ability to identify their importance within a context. Hence, the *refSum* may synthesize the document, but it does not necessarily provide all the useful information they need.

BB25HLegalSum is outlined in Figure 4.1, and it unfolds through four key phases:

1. Selection of a set of unique sentences from D.desc using BERT;
2. Clustering unique sentences, and scoring sentences and clusters;
3. Generation and selection of Candidate Summaries;
4. Presentation of the resulting *GSum* within the original document, with selected sentences highlighted using distinct colors to provide different perspectives on importance.

A significant concern in our work is understanding the trade-off between conciseness and completeness as a measure of the quality of the generated summaries. Hence, our method proposes and assesses three strategies to select the best-generated summary (GSum), given a set of possible candidates, according to the metrics used for the selection (ROUGE precision, f-measure and recall, respectively). We select the best candidate summaries, and ultimately the *GSum* for a document, according to three strategies, as represented by Rouge metrics: a) *precision-oriented summary* (***PoSum***), focused on

Figure 4.1: BB25HLegalSum overview


Legal document
Description
Reference summary
Break into sentences
Select unique sentences
Cluster related sentences (topics)
Score sentences (based in cluster and individual sentence scoring)
BM25
Generate Candidates (CCandSum, SCandSum)
Select Best Candidates
RoSum
FoSum
PoSum
Highlight with 3 colors (Po, Fo, Ro)

conciseness; b) *recall-oriented summary* (***RoSum***), focused on completeness; and c) (*f-measure-oriented summary* (***FoSum***), as a trade-off. Conciseness refers to conveying the message clearly and succinctly without including unnecessary details. Completeness relates to the inclusion of key information from the original text. A summary with good conciseness and completeness will be easy to read and understand, ensuring that produced summaries convey key information from the legal document to the target audience.

The remainder of the chapter provides details on our method.

4.2 Extracting BERT Unique Sentences

Given a legal document $D(desc, refSum)$, our objective is to break down $D.desc$ into sentences s_i (where $0 < i < D.desc.length$). Then we select a subset *uniqueSentences*, i.e. a set of unique sentences that meet a minimum length criterion.

We break the sentences separated by a semicolon, colon, quotation mark, a double hyphen, or a dot symbol using the regex ‘;|“|–|\.’. We also split sentences longer than 30 characters with the ‘(’, in order to get segments of text that usually relate to an item.

Then we produce two sets of sentence indices, $minSizeFilterIDX$ and $BERTFilterIDX$, where each index is within $\{a \mid 0 \leq a \leq D.desc.length\}$, ensuring that a corresponding sentence s_i is found within $D.desc$.

- *(a) Minimum Size Filter:* Our initial step addresses the minimum sentence size required for a sentence to be considered a candidate, guided by a *size_threshold*. The rationale is to eliminate overly short sentences, given that legal datasets often comprise lengthy *refSum* sentences. We determine *size_threshold* by measuring the shortest sentence in each *refSum* (shortestSentrefSum) across all documents and calculating its average (*shortestSentrefSumAvg*). In our experiments, we set $size\_threshold = 2 * shortestSentsrefSumAvg$. The list *minSizeFilterIDX* contains indices of sentences in *D.desc* that meet this minimum size requirement.

- *(b) BERT Filter:* The goal of the BERT filter is to identify and eliminate duplicated sentences. Initially, each sentence s_i is transformed into an equivalent BERT representation br_i using a pre-trained BERT model. To identify duplications, we assess the similarity of each br_i with other previously selected br_j representations. A *similarity_threshold* determines uniqueness, and sentences exceeding this threshold are regarded as duplicates and excluded. We set *similarity_threshold* at 0.9, achieving a balance between identifying repetitive and related sentences. The list *BERTFilterIDX* contains indices of non-duplicate sentences in *D.desc*, adhering to the *similarity_threshold*.

Finally, we compute the intersection of *minSizeFilterIDX* and *BERTFilterIDX* to create *uniqueSentences*, i.e. a set sentences $s_i \in D.desc$, where $i \in uniqueSentences$.

Table 4.1 shows unique sentences filtered by the BERT filter in Bill 1104 of Bill Sum US, and the respective similarity score regarding the most similar sentence within *D.desc*. This shows that the filtered sentences may present some similarity to the compared sentence but are not duplicates.

4.3 Scoring sentences and clusters

The goal of this stage is to cluster related sentences from *uniqueSentences* to group them based on subject or topic, and assign a relevance score to each sentence and cluster. These scores are used in the next phase to compose and select candidate summaries. The remaining of this section describes the steps performed in this phase.

Table 4.1: Sentences and their similarity with *refSum* - Bill 1104 (US test data)

Similarity score	**Candidate sentence**	**Most Similar Sentence**
0.7077	director" means the director of the office of personnel management	administrator" means the administrator of the small business administration
0.6277	program" means the federal entrepreneur-in-residence program established under section 3	entrepreneur-in-residence" means an individual appointed to a position under the program
0.6399	federal entrepreneur-in-residence program	federal entrepreneur-in-residence act of 2012"
0.6012	(b) may not appoint more than 10 entrepreneurs-in-residence during any year	(a) shall appoint entrepreneurs-in-residence under the program during each year
0.7503	in appointing entrepreneurs-in-residence, the director shall	the director shall select entrepreneurs-in- residence from among individuals who
0.7613	(b) may not serve as an entrepreneur-in-residence for more than a period of 2 years	(b) to the extent practicable, not appoint more than 2 entrepreneurs-in-residence to positions in the same agency during the same year
0.6391	(1) assist federal agencies in improving outreach to small business concerns and entrepreneurs	(2) strengthen coordination and interaction between the federal government and the private sector on issues relevant to entrepreneurs and small business concerns
0.7375	(3) provide recommendations to the head of the agency employing the entrepreneur-in-residence on methods to improve program efficiency at the agency or new initiatives, if any, that may be instituted at the agency	(2) provide recommendations to the head of the agency employing the entrepreneur-in-residence on inefficient or duplicative programs, if any, at the agency
0.6526	(5) facilitate in-service sessions with employees of the agency employing the entrepreneur-in-residence on issues of concern to entrepreneurs and small business concerns	(4) facilitate meetings and forums to educate small business concerns and entrepreneurs on programs or initiatives of the agency employing the entrepreneur-in-residence
0.6566	(6) provide technical assistance or mentorship to small business concerns and entrepreneurs in accessing programs at the agency employing the entrepreneur-in-residence	(4) facilitate meetings and forums to educate small business concerns and entrepreneurs on programs or initiatives of the agency employing the entrepreneur-in-residence
0.7330	if an entrepreneur-in-residence with a rate of pay equivalent to the rate of basic pay for a position at gs-13 or gs-14 satisfactorily completes 1 year of service in position under this section, the entrepreneur-in-residence may receive an increase in the rate of basic pay to be equal to the rate of basic pay for a position 1 grade higher on the general schedule than the initial rate of basic pay of the entrepreneur-in-residence	the rate of basic pay for an entrepreneur- in-residence shall be equivalent to the rate of basic pay for a position at gs-13, gs-14, or gs-15 of the general schedule, which shall be determined in accordance with regulations promulgated by the director

4.3.1 Cluster related sentences

The first step in this phase is to identify clusters of related, unique sentences. Recall that, due to the BERT filter, sentences in a cluster are more thematically connected than strictly similar. The purpose is to group them based on subject or topic, and then merge them later on to form candidate summaries. The rationale is to narrow the search space for relevant sentences to be included in the candidate summaries.

We utilize as input BERT representations of sentences from *uniqueSentences*, and and we cluster them using the K-means algorithm. To select the value of *k*, we vary k from 2 to 50, choosing the clustering associated with the highest Silhouette score.

Table 4.2 illustrates how bill 1104 in Billsum US was split into clusters.

Table 4.2: Clusters from bill number 1104 within US test dataset

Cluster	**Sentences**	**Cluster score**
Cluster 0	"entrepreneur-in-residence” means an individual appointed to a position under the program "the director, in consultation with the administrator, shall establish a federal entrepreneur-in-residence program under which the director, with the concurrence of the head of an agency, may appoint an entrepreneur-in-residence to a position in the excepted service in the agency to carry out the duties described in subsection (d)" (a) shall appoint entrepreneurs-in-residence under the program during each year (a) give priority to placing entrepreneurs-in- residence across the federal government at separate agencies to the extent practicable, not appoint more than 2 entrepreneurs-in-residence to positions in the same agency during the same year (a) shall be a full-time employee of the agency to which the entrepreneur-in-residence is appointed (2) provide recommendations to the head of the agency employing the entrepreneur-in-residence on inefficient or duplicative programs, if any, at the agency	0.3262
Cluster 1	(1) provide for better outreach by the federal government to the private sector (2) strengthen coordination and interaction between the federal government and the private sector on issues relevant to entrepreneurs and small business concerns (3) make federal programs simpler, quicker, more efficient, and more responsive to the needs of small business concerns and entrepreneurs	0.1616
Cluster 2	the director shall select entrepreneurs-in- residence from among individuals who an entrepreneur-in-residence shall report directly to the head of the agency employing the entrepreneur-in-residence 1 - the director may not appoint an entrepreneur-in- residence under this section after september 30, 2016	0.1264
Cluster 3	administrator” means the administrator of the small business administration agency” means an executive agency, as defined in section 105 of title 5, united states code small business concern” has the meaning given that term under section 3 of the small business act	0.1255
Cluster 4	(b) have demonstrated success in working with small business concerns and entrepreneurs (c) have successfully developed, invented, or created a product and brought the product to the marketplace (4) facilitate meetings and forums to educate small business concerns and entrepreneurs on programs or initiatives of the agency employing the entrepreneur-in-residence	0.1533
Cluster 5	the rate of basic pay for an entrepreneur- in-residence shall be equivalent to the rate of basic pay for a position at gs-13, gs-14, or gs-15 of the general schedule, which shall be determined in accordance with regulations promulgated by the director	0.145
Cluster 6	federal entrepreneur-in-residence act of 2012”	0.021

4.3.2 Scoring Clusters and Sentences

This second step involves assigning relevance scores to clusters and sentences. As a scoring system, we have chosen ROUGE metrics to score all sentences and clusters. Following the methodology described in (JAIN; BORAH; BISWAS, 2023b), cluster scores stem from averaging sentence relevance scores relative to a reference summary *refSum*.

First, we calculate an individual sentence score (ISC_i) for all sentences $s_i \in$

uniqueSentences. We have chosen to calculate and average the F1-measure of Rouge-1, Rouge-2, and Rouge-L with regard to *refSum*, to achieve the best trade-off of recall and precision when comparing unigrams, bigrams, and longest common subsequences, respectively. To calculate individual sentence scores ISC_i, we compare each sentence s_i with refSum using the aforementioned F1 ROUGE scores, and then we average the three F1 metrics.

Next, we use the individual scores ISC of the sentences within a cluster to calculate the score of the respective cluster (CSC_j). Given a $Cluster_j$, we compute the respective CSC_j by averaging the scores ISC_l of all sentences s_l belonging to $Cluster_j$.

Finally, we calculate a final sentence score ($FinalSC_i$) for all sentences $s_i \in uniqueSentences$ by combining the initial individual score and the respective cluster score. Given a sentence $s_l \in Cluster_j$ with score ISC_l and CSC_j, respectively, the $FinalSC_l = (1+CSC_j)/2) * ISC_l$.

Table 4.3 shows the *Cluster score*, *Sentence*, *Individual sentence score* and *Final sentence score* for bill 1104 of BillSum US dataset for the top 20 final scored sentences. The sentences are presented in decreasing order of final score. As we can see, the third sentence has a higher overall score than the fourth sentence since the individual score is more important than the cluster score when defining the final score for each sentence.

Table 4.3: Output scores from bill number 1104 within US dataset for top-20 sentences

Cluster score	**Sentence**	**Individual sentence score**	**Final sentence score**
0.3262	entrepreneur-in-residence” means an individual appointed to a position under the program	0.0422	0.028
0.3262	the director, in consultation with the administrator, shall establish a federal entrepreneur-in-residence program under which the director, with the concurrence of the head of an agency, may appoint an entrepreneur-in-residence to a position in the excepted service in the agency to carry out the duties described in subsection (d)	0.1579	0.1047
0.3262	(a) shall appoint entrepreneurs-in-residence under the program during each year	0.0413	0.0274
0.3262	(a) give priority to placing entrepreneurs-in- residence across the federal government at separate agencies	0.0556	0.0369
0.3262	(b) to the extent practicable, not appoint more than 2 entrepreneurs-in-residence to positions in the same agency during the same year	0.0784	0.052
0.3262	(a) shall be a full-time employee of the agency to which the entrepreneur-in-residence is appointed	0.0574	0.0381
0.3262	(2) provide recommendations to the head of the agency employing the entrepreneur-in-residence on inefficient or duplicative programs, if any, at the agency	0.0825	0.0547
0.1616	(1) provide for better outreach by the federal government to the private sector	0.0525	0.0305
0.1616	(2) strengthen coordination and interaction between the federal government and the private sector on issues relevant to entrepreneurs and small business concerns	0.0847	0.0492
0.1616	(3) make federal programs simpler, quicker, more efficient, and more responsive to the needs of small business concerns and entrepreneurs	0.0789	0.0458
0.1264	the director shall select entrepreneurs-in- residence from among individuals who	0.038	0.0214
0.1264	an entrepreneur-in-residence shall report directly to the head of the agency employing the entrepreneur-in-residence	0.052	0.0293
0.1264	the director may not appoint an entrepreneur-in- residence under this section after september 30, 2016	0.063	0.0355
0.1255	administrator” means the administrator of the small business administration	0.027	0.0152
0.1255	agency” means an executive agency, as defined in section 105 of title 5, united states code	0.0586	0.033
0.1255	small business concern” has the meaning given that term under section 3 of the small business act	0.0644	0.0363
0.1533	(b) have demonstrated success in working with small business concerns and entrepreneurs	0.0507	0.0292
0.1533	(c) have successfully developed, invented, or created a product and brought the product to the marketplace	0.0626	0.0361
0.1533	(4) facilitate meetings and forums to educate small business concerns and entrepreneurs on programs or initiatives of the agency employing the entrepreneur-in-residence	0.0869	0.0501
0.145	the rate of basic pay for an entrepreneur- in-residence shall be equivalent to the rate of basic pay for a position at gs-13, gs-14, or gs-15 of the general schedule, which shall be determined in accordance with regulations promulgated by the director	0.145	0.083

4.4 Generation and Selection of Candidate Summaries

In this phase, we take as input all sentence and cluster scores, as well as the clustering resulting from the previous step (Section 4.3). The goal is to iteratively generate a set of candidate summaries using the best-ranked sentences. To compose the ranking, we combine a BM25 filter to identify the most relevant sentences with regard to the whole document, the sentences that belong to the best-scored clusters, and the best individually-ranked sentences. The created candidate summaries are compared against the reference summary using Rouge-1 scores for precision, recall, and F1. We select among all the candidate summaries the ones with the highest precision (PoSum), recall (RoSum) and F1 (FoSum).

- *(a) BM25 Ranking Filter:* We initially pass all *uniqueSentences* through the BM25 algorithm. BM25 functions as a bag-of-words retrieval mechanism, ranking documents based on query term occurrences. The query terms encompass all tokens extracted from *D.desc*, and sentences s_i are ranked according to their relevance. The objective is to select sentences that best represent the overall document, enhancing the quality of the generated summary. The top-n best-ranked sentences are chosen. *BM25FilterIDX* holds indices of the top-n most relevant sentences according to BM25 ranking. We defined the value of $top-n = 10\%$ according to the conclusions of the experiments done in (JAIN; BORAH; BISWAS, 2023b), aiming to keep a balance between reference summary and desc average ratios. In our experiments, we identified the following ratios: 15% in BillSum and 7,51% for RulingBR. Hence, we adopted 10% as the threshold of top-ranked document sentences within *uniqueSentences*.

- *(b) Create candidate summaries from clusters - cCandSum*: in this step, we create candidate summaries cCandSum that are the concatenation of the sentences included in the top-scored clusters, and which are also included in *BM25FilterIDX*. Let $C = \{C_j | 0 <= j < n\}$ be the set of best-ranked clusters, where the cluster scores $CSC_i > CSC_k$, for $i < k < n$. Experimentally, we have defined the top-5 clusters, i.e. $n = 5$. The first $cCandSum_0$ contains all the sentence indices within C_0 that are also in *BM25FilterIDX*. Then, $cCandSum_1$ contains all the sentence indexes in $cCandSum_0$, and the sentence indices within C_1 that are also in *BM25FilterIDX*. This goes on until the fifth cluster ($cCandSum_4$), or there are no other clusters to concatenate.

- *(c) Create variations of cCandSum by adding relevant sentences - sCandSum*: in this step, we generate more candidate summaries that include not only the sentences of the best-scored clusters but also the best individually scored sentences, given they are also present in the *BM25FilterIDX*, i.e. they are among the most relevant sentences of the document. Let $SS = \{s_k | 0 <= k < m\}$ be the set of best individually ranked sentences, where the individual sentence scores $FinalSC_a > FinalSC_b$, for $a < b < m$. The code iterates over all $cCandSum_l$ $(0 <= l <= 4)$. Initially, $sCandSum_{l,0} = cCandSum_l \cup s_0$. Then, we iterate through all remaining sentences $s_i \in SS$ $(1 <= i < m)$ creating $sCandSum_{l,i} = sCandSum_{l,i-1} \cup s_i$. Experimentally, we have defined the sentences with the top-30 highest FinalSC, i.e. $m = 30$.

- *(d) Selection of the best candidate summary*: in this final step, we compare all the generated candidate summaries with the reference summary *D.refSum* using ROUGE1 precision, recall, and F1. Let CCS be the set of all $cCandSum_l$ and SCS be the set of all $sCandSum_{l,i}$, and $CandSum = CCS \cup SCS$. The algorithm records the $candSum_j \in CandSum$ of which the set of sentences yields the most favorable score, based on different ROUGE-1 metrics. Ultimately, the selected $candSum_j$, known as *GSum*, is chosen as the final representative summary based on a specific criterion (there are three best candidate summaries - one for recall, one for precision and one for f-measure). The *GSum* that achieved the best result in each score is referred to as *recall-oriented summary* (***RoSum***), *precision-oriented summary* (***PoSum***) and (*f-measure-oriented summary* (***FoSum***), as a trade-off.

4.5 Highlighting and distinct showcasing summaries in the legal document

To be useful, it is important that the generated summaries are readable. A summary with good conciseness and completeness will be easy to read and understand, ensuring that produced summaries convey key information from the legal document to the target audience. Conciseness refers to conveying the message clearly and succinctly without including unnecessary details. Completeness relates to the inclusion of key information from the original text.

We propose to present the generated summaries as highlights with distinct colors in the original text. Highlighting text improves the reader's knowledge and understanding of the topic being explored (ROY et al., 2021) and it allows the reader to fully grasp not only the relevant words but their context, which can be inspected whenever necessary.

We opted to present the three types of summaries within a single document, in a subsidiary manner, using distinct colors, one for each criterion-focused summary, i.e. one color for PoSum, another distinct color for FoSum, and a third color for RoSum. As displayed in Figure 4.3, in our implementation we adopted green for PoSum, blue for FoSum, and red for RoSum, but is crucial to note that we did not conduct an experiment to determine the optimal colors. Consequently, the adopted color pattern is purely illustrative of the presentation properties.

The highlighting pattern allows the reader to understand the different nuances for each highlighted color while condensing the three generated summaries into a single text. We chose to highlight with three colors in a subsidiary way (*subsidiary highlighting*) instead of highlighting the colors of the intersections (*intersectional highlighting*). This deliberate choice enables us to emphasize the differences between PoSum, FoSum, and RoSum, particularly within the areas where their content overlaps. Compared to related work (POLSLEY; JHUNJHUNWALA; HUANG, 2016; LICARI et al., 2023), we provide the context for the relevant sentences and highlight them according to different points of view (precision vs. coverage).

This choice also enables us to conduct a qualitative assessment of summary readability, evaluating the assessment of our proposition on legal workers. In Section 5.5 we qualitatively assess the summaries generated according to each strategy in terms of conciseness and completeness.

Our method relies on the premise that PoSums are shorter than (or equal to) the FoSums, which in turn are shorter than (or equal to) RoSums. Given the PoSum, FoSum and RoSum generated for a given document D, we start by highlighting every selected sentence that is in the PoSum. Then we highlight sentences that appear in the FoSum that were not included in the PoSum. Finally, we highlight all sentences that are shown in the RoSum and which have not been highlighted yet.

Figure 4.2 displays an example of a highlighted bill using the subsidiary highlighting solution. It is important to mention that, in this particular case, no red color is shown, since the FoSum and RoSum summaries have the same content. Hence, the summaries with the best f1-measure (FoSum) and recall scores (RoSum) have the same content or that RoSum does not contain any new sentence in comparison to PoSum and FoSum.

On the other hand, by doing intersectional highlighting, it is possible to show the

Figure 4.2: Bill 1104: Reference summary and highlighted bill according to the three strategies in a subsidiary manner.

Reference summary	**Precision focused (PoSum color), F-measure focused (PoSum color + FoSum color) and Recall focused (PoSum color + FoSum color + RoSum color) summaries**
Federal Entrepreneur-in-Residence Act of 2012 - Directs the Director of the Office of Personnel Management (OPM) to establish an entrepreneur-in-residence program to appoint in-house entrepreneurs who have demonstrated success in working with small business concerns and entrepreneurs to: (1) assist federal agencies in improving outreach to small business concerns and entrepreneurs, (2) provide recommendations on inefficient or duplicative agency programs and on methods to improve agency efficiency, (3) facilitate meetings and forums to educate small business concerns and entrepreneurs on agency programs and initiatives, and (4) provide technical assistance or mentorship. Limits to 10 the number of entrepreneurs-in-residence that the Director may appoint in any year. Terminates such program after FY2016.	SECTION 1. SHORT TITLE. This Act may be cited as the “Federal Entrepreneur-in-Residence Act of 2012”. SEC. 2. DEFINITIONS. In this Act– (1) the term “Administrator” means the Administrator of the Small Business Administration; (2) the term “agency” means an Executive agency, as defined in section 105 of title 5, United States Code; (3) the term “Director” means the Director of the Office of Personnel Management; (4) the term “entrepreneur-in-residence” means an individual appointed to a position under the program; (5) the term “program” means the Federal entrepreneur-in- residence program established under section 3(a); and (6) the term “small business concern” has the meaning given that term under section 3 of the Small Business Act (15 U.S.C. 632). SEC. 3. FEDERAL ENTREPRENEUR-IN-RESIDENCE PROGRAM. (a) Program Established.–The Director, in consultation with the Administrator, shall establish a Federal entrepreneur-in-residence program under which the Director, with the concurrence of the head of an agency, may appoint an entrepreneur-in-residence to a position in the excepted service in the agency to carry out the duties described in subsection (d). (b) Mission of Program.–The mission of the program shall be to– (1) provide for better outreach by the Federal Government to the private sector; (2) strengthen coordination and interaction between the Federal Government and the private sector on issues relevant to entrepreneurs and small business concerns; and (3) make Federal programs simpler, quicker, more efficient, and more responsive to the needs of small business concerns and entrepreneurs. (c) Appointments.– (1) In general.–The Director– (A) shall appoint entrepreneurs-in-residence under the program during each year; and (B) may not appoint more than 10 entrepreneurs-in- residence during any year. (2) Selection.–The Director shall select entrepreneurs-in- residence from among individuals who– (A) are successful in their field; (B) have demonstrated success in working with small business concerns and entrepreneurs; or (C) have successfully developed, invented, or created a product and brought the product to the marketplace. (3) Placement.–In appointing entrepreneurs-in-residence, the Director shall– (A) give priority to placing entrepreneurs-in- residence across the Federal Government at separate agencies; and (B) to the extent practicable, not appoint more than 2 entrepreneurs-in-residence to positions in the same agency during the same year. (4) Terms of appointment.–An entrepreneur-in-residence– (A) shall be a full-time employee of the agency to which the entrepreneur-in-residence is appointed; and (B) may not serve as an entrepreneur-in-residence for more than a period of 2 years. (d) Duties.–An entrepreneur-in-residence shall– (1) assist Federal agencies in improving outreach to small business concerns and entrepreneurs; (2) provide recommendations to the head of the agency employing the entrepreneur-in-residence on inefficient or duplicative programs, if any, at the agency; (3) provide recommendations to the head of the agency employing the entrepreneur-in-residence on methods to improve program efficiency at the agency or new initiatives, if any, that may be instituted at the agency; (4) facilitate meetings and forums to educate small business concerns and entrepreneurs on programs or initiatives of the agency employing the entrepreneur-in-residence; (5) facilitate in-service sessions with employees of the agency employing the entrepreneur-in-residence on issues of concern to entrepreneurs and small business concerns; and (6) provide technical assistance or mentorship to small business concerns and entrepreneurs in accessing programs at the agency employing the entrepreneur-in-residence. (e) Compensation.– (1) In general.–The rate of basic pay for an entrepreneur-in-residence shall be equivalent to the rate of basic pay for a position at GS-13, GS-14, or GS-15 of the General Schedule, which shall be determined in accordance with regulations promulgated by the Director. (2) Promotion.–If an entrepreneur-in-residence with a rate of pay equivalent to the rate of basic pay for a position at GS-13 or GS-14 satisfactorily completes 1 year of service in position under this section, the entrepreneur-in-residence may receive an increase in the rate of basic pay to be equal to the rate of basic pay for a position 1 grade higher on the General Schedule than the initial rate of basic pay of the entrepreneur-in-residence. (f) Reporting.–An entrepreneur-in-residence shall report directly to the head of the agency employing the entrepreneur-in-residence. (g) Termination.–The Director may not appoint an entrepreneur-in- residence under this section after September 30, 2016.

overlapping and differences between each oriented summary. For example, by using the same coloring, it is possible to show where the colors overlap with the scheme shown in Figure 4.4.

This work has not addressed accessibility issues for the proposed highlighting presentation. For instance, if the reader is colorblind, we might select distinct colors suitable for such a condition, or propose additional ways of differentiating the text for each summarization approach, such as underlining, bold or italic.

Figure 4.3: Subsidiary highlighting in colors scheme

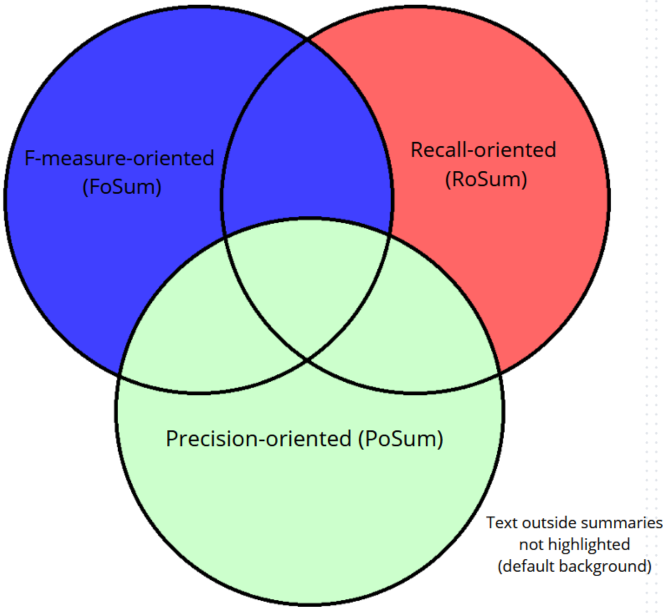
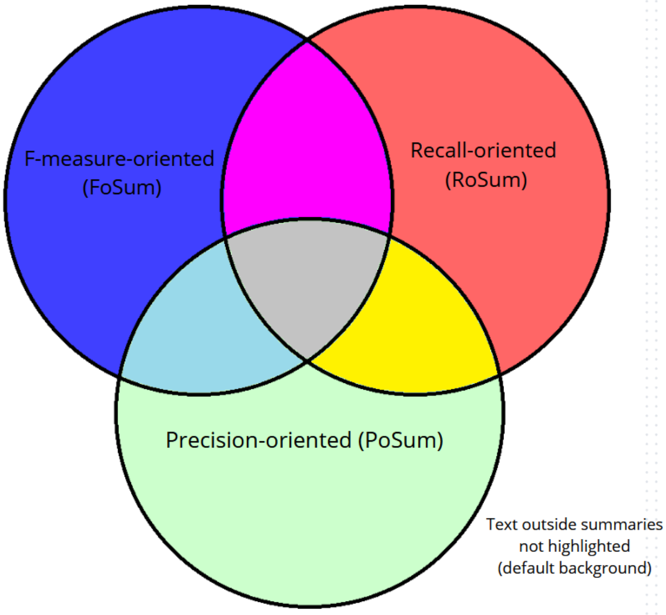


Figure 4.4: Intersectional highlighting in colors scheme



5 EXPERIMENTS

5.1 Research questions

There are three experiments, where each one addresses a specific research question:

- Experiment 1: "*How does the performance of BB25HLegalSum compare to baseline methods for legal document summarization?*"
- Experiment 2: "*How does the length of the reference summary impact the recall and precision of the generated summary?*"
- Experiment 3: "*Which type of summary is more suitable in the legal documents context concerning its readability: focused on precision, on recall, or f-measure?*"

5.2 Datasets and baselines

We used BillSum and RulingBR as datasets used for experiments. BillSum is a dataset of legislative bills, where each bill is composed of a title, a text, and a summary (KORNILOVA; EIDELMAN, 2019). According to (KORNILOVA; EIDELMAN, 2019), it is the first dataset for summarization of US Congressional and California state bills. The data is decomposed in into training and test data. Since our method is unsupervised, we used only the test datasets (US and CA) in our experiments. US test data contains 3269 bills, and CA test data has 1238 bills.

RulingBR is the largest Brazilian dataset containing STF decisions as the main source (FEIJÓ; MOREIRA, 2018). It is composed of 10574 rulings, each of them characterized by a summary, a report, a vote and a judgment section. As D.desc, we have used the concatenation of the report, vote and judgment to make the final summaries, in order to compare with the ementa as reference summary.

For performance comparison (Experiment 1), we have selected the following baselines according to the results reported for the above data sets:

Billsum US and CA test data: We considered as baselines the best results as reported in the literature for legal document extractive summarization in BillSum. We considered the best three results compiled in (JAIN; BORAH; BISWAS, 2021), namely from LSTM with Word2vec, LexRank and TextRank, as well as DCESumm, reported in (JAIN; BORAH; BISWAS, 2023b).

RulingBR test data: we used the results reported in (FEIJÓ, 2021), namely LegalSumm and Feijo and Moreira (2019).

To our knowledge, these are the best legal document extractive results reported in billSum literature and legal document summarization results for rulingBR dataset.

5.3 Experiment 1

5.3.1 Method

The goal of this experiment is to assess the performance of BB25HLegalSum in comparison with legal document summarization approaches. For all the documents in the datasets mentioned in Section 5.2, we produced system summaries according to the three strategies (RoSum, PoSum and FoSum). As an embedding model, we chose distilBERT multilingual version[1], since it has demonstrated good performance at zero-shot cross-lingual model transfer (PIRES; SCHLINGER; GARRETTE, 2019). We have used BM25 version from the Gensim library[2].

Then, we measured the ROUGE-1, ROUGE-2 and ROUGE-L in terms of precision, recall and F1 scores against the respective reference summaries, and compared them against the baselines.

5.3.2 Results - BillSum

Tables 5.1 and 5.2 present the average scores for each strategy, referred to as *BB25HLegalSum PoSum* (precision), *BB25HLegalSum RoSum* (recall) and *BB25HLegalSum FoSum* (F1).

It is possible to see that BB25HLegalSum RoSum consistently achieved higher scores across all ROUGE metrics in both BillSum datasets. In the same datasets, BB25HLegalSum FoSum and PoSum outperformed all the baselines in terms of F1 for all ROUGE criteria.

In terms of recall, BB25HLegalSum FoSum exhibited better results than the base-

[1] <https://huggingface.co/distilbert-base-multilingual-cased>

[2] <https://radimrehurek.com/gensim_3.8.3/summarization/bm25.html>

lines for the CA test data (Table 5.2), but it was superior to all baselines only for Rouge-2 recall (Table 5.1). Comparatively, BB25HLegalSum PoSum presented the worst performance, outperforming all the baselines only on the CA test data regarding Rouge-2 recall.

It is important to note a potential source of bias in our methodology. Unlike some other approaches in the field of legal document summarization, our method directly compares sentences with a reference summary to assess performance. While this direct comparison provides better results for measuring the readability of the generated summary, readers should approach the presented results with an awareness of this methodological choice and its potential impact on the results.

5.3.3 Results - RulingBR

Table 5.3 displays the results for the RulingBR dataset. In terms of F1, BB25HLegalSum RoSum and FoSum outperformed both baselines. BB25HLegalSum PoSum outperformed both baselines only for Rouge-2 and Rouge-L F1. In terms of recall, BB25HLegalSum RoSum and FoSum also outperformed the two baselines. Comparatively, BB25HLegalSum PoSum presented the worst results.

If we consider the scores of our systems for the RulingBR test data, it sounds reasonable to say that LegalSumm results seem more precision-oriented, while the Feijo and Moreira (2019) results seem more f-measure-oriented. In this sense, although we have not always surpassed all results, some are comparable (recall in ROUGE-2 scores were 0.2475 in BB25HLegalSum PoSum in comparison to LegalSumm 0.25; f-measure ROUGE-1 scores were 0.4263 in BB25HLegalSum PoSum while 0.44 in LegalSumm). Therefore, the performance seems fitting even when it did not outperform certain criteria.

Again, we note that our results might have some bias, as we use the reference summary to generate the summary and assess its performance, unlike the baselines. Readers should approach the presented results with an awareness of this methodological choice

Table 5.1: Performance on US test data

Methods	**Rouge-1**			**Rouge-2**			**Rouge-L**		
	F	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
LSTM-with-w2v	0.3615	N/A	0.6539	0.2086	N/A	0.3720	0.3664	N/A	0.5358
Lexrank	0.3704	N/A	0.5415	0.1811	N/A	0.2604	0.3365	N/A	0.4230
Textrank	0.3269	N/A	0.6295	0.1793	N/A	0.3423	0.3383	N/A	0.5037
DCESumm	0.4200	N/A	N/A	0.2428	N/A	N/A	0.3887	N/A	N/A
BB25HLegalSum FoSum	**0.5748**	**0.5701**	0.5999	**0.3699**	**0.3681**	**0.3854**	**0.4933**	**0.5092**	0.4910
BB25HLegalSum PoSum	**0.5088**	**0.7110**	0.4373	**0.3439**	**0.4932**	0.2942	**0.4664**	**0.6377**	0.3974
BB25HLegalSum RoSum	**0.5283**	**0.4406**	**0.7095**	**0.3446**	**0.2894**	**0.4572**	**0.4693**	**0.4212**	**0.5529**

Table 5.2: Performance on CA test data

Methods	**Rouge-1**			**Rouge-2**			**Rouge-L**		
	F	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
LSTM-with-w2v	0.4073	N/A	0.4638	0.1883	N/A	0.2093	0.3312	N/A	0.3588
Lexrank	0.4144	N/A	0.4529	0.1936	N/A	0.2083	0.3406	N/A	0.3531
Textrank	0.4069	N/A	0.5055	0.2015	N/A	0.2461	0.3457	N/A	0.3848
DCESumm	0.4366	N/A	N/A	0.2389	N/A	N/A	0.3915	N/A	N/A
BB25HLegalSum FoSum	**0.5443**	**0.5848**	**0.5313**	**0.3282**	**0.3576**	**0.3176**	**0.4458**	**0.4732**	**0.4345**
BB25HLegalSum PoSum	**0.5077**	**0.7177**	0.4352	**0.3188**	**0.4832**	**0.2576**	**0.4445**	**0.5940**	0.3755
BB25HLegalSum RoSum	**0.5261**	**0.5155**	**0.5804**	**0.3130**	**0.3117**	**0.3409**	**0.4295**	**0.4252**	**0.4551**

Table 5.3: Performance on RulingBR test data

Methods	**Rouge-1**			**Rouge-2**			**Rouge-L**		
	F	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**
LegalSumm	0.44	0.64	0.38	0.29	0.42	0.25	0.36	0.52	0.31
Feijo and Moreira (2019)	0.4427	0.4938	0.4624	0.2650	0.2836	0.2826	0.3527	0.3852	0.3736
BB25HLegalSum FoSum	**0.5118**	0.5215	**0.5327**	**0.3380**	0.3504	**0.3463**	**0.4700**	0.4867	**0.4759**
BB25HLegalSum PoSum	0.4263	**0.7164**	0.3437	**0.3114**	**0.5494**	0.2475	**0.4291**	**0.6788**	0.3477
BB25HLegalSum RoSum	**0.4527**	0.3766	**0.6641**	**0.2983**	0.2524	**0.4213**	**0.4285**	0.3716	**0.5668**

and its potential impact on the superior results.

5.4 Experiment 2

5.4.1 Method

This experiment assesses the performance of the method according to the length of the reference summaries. We have performed it using US test BillSum and RulingBR datasets. As an embedding model, we chose again the distilBERT multilingual version and the BM25 version from Gensim.

First, we divided the reference summaries into length intervals in terms of the number of characters. We defined the following intervals: [0..500],[501..1000],[1001..1500], [1501 .. 2000], and an interval greater than 2001 characters. Then, we calculated the different scores for each interval for each summarization strategy (PoSum, FoSum, RoSum). The results of our evaluation provide insights into the effectiveness of different summarization techniques for different lengths of reference summaries.

5.4.2 Results and discussion - BillSum

The results for the BillSum US dataset are presented in Figures 5.1, 5.2 and 5.3 for PoSum, FoSum, and RoSum summaries, respectively. All tables provide ROUGE-1, ROUGE-2 and ROUGE-L precision, f-measure, and recall values according to the reference summary length intervals considered.

Considering the PoSums (Figure 5.1), we observe that the ROUGE-2 and ROUGE-L recall and f1-measure scores increased with the length of the reference summaries. Other values do not present any significant change in scores, since when comparing minimum and maximum values for each ROUGE-1 criteria (recall, precision and f-measure) in all US dataset result tables (5.1, 5.2 and 5.3), the difference does not surpass 0,02.

Figure 5.1: PoSum scores (US test Data)

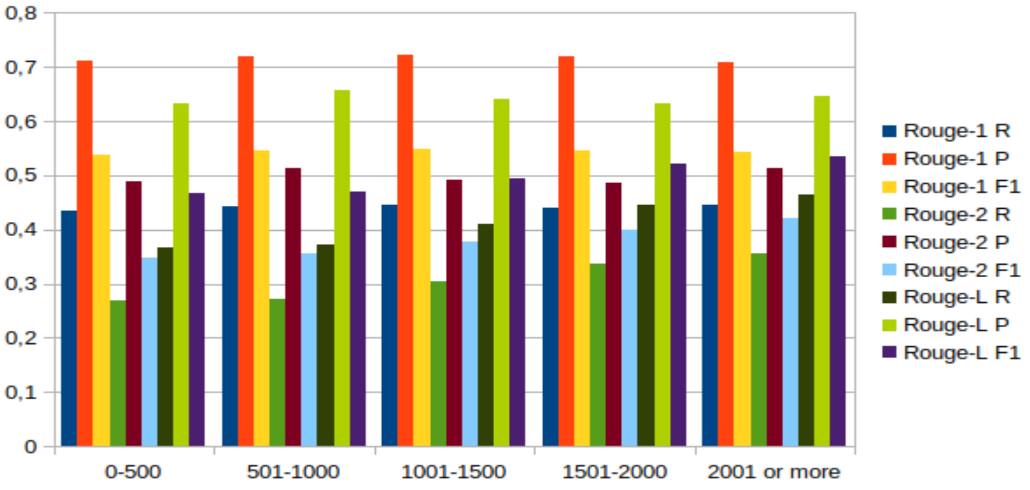


Figure 5.2: FoSum scores (US test Data)

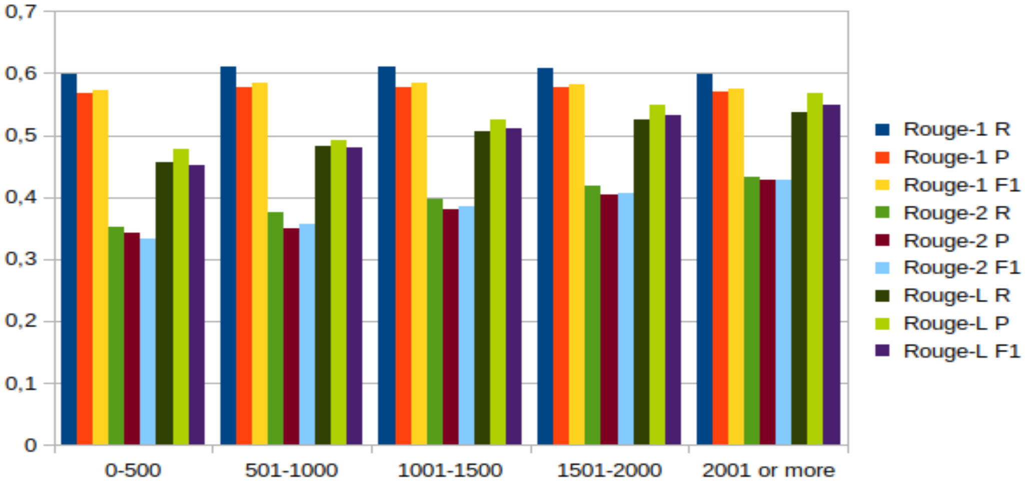
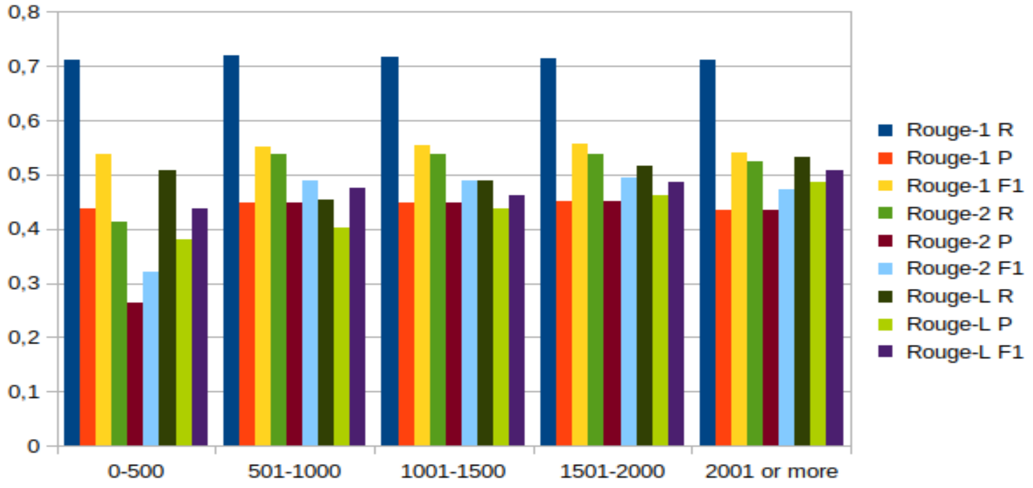


Figure 5.3: RoSum scores (US test Data)



As seen in Figure 5.2, the ROUGE-2 and ROUGE-L scores of FoSums increased with the length of the reference summaries. Other values (ROUGE-1 related variables) present a balanced score.

Finally, considering the RoSum strategy displayed in Figure 5.3, BB25HLegalSum had a positive spike in Rouge-2 (R,P,F) and Rouge-L (P,F) values, when comparing the interval shorter than 500 characters scores with the 501-1000 scores, with the exception of Rouge-L (R), with values falling from the 500 to the 501-1000 ranges, but increasing later on.

Regardless of the summarization strategy (whether focused on precision, f-measure, or recall), practically no score (with a few exceptions, such as ROUGE-2 F1 scores dropping when comparing 1501-2000 with 2001 or more scores, as seen in Figure 5.3) has decreased with lengthier reference summaries and overall the scores have increased, even if slightly in most cases.

In short, the length of the reference summary does not impact the ROUGE-1 recall (RoSum) and precision (PoSum) of the summaries generated using BB25HLegalSum. On the other hand, it slightly impacts ROUGE-2 and ROUGE-L in a positive manner. This result is good since it is expected for ROUGE-2 and ROUGE-L values to increase when the reference summary is lengthier due to the nature of how these ROUGE metrics are calculated. ROUGE-2 calculates the overlap of bigrams between the generated summary and the reference summary. ROUGE-L measures the longest common subsequence between the generated and reference summaries. When the reference summary is lengthier, it will likely contain more bigrams and longer common subsequences that match the generated summary.

One important variable to examine in this experiment is ROUGE-1. As we can see, it does not get impacted by longer reference summaries. That is a good result since when dealing with clusters, we would expect that it could severely impact the ROUGE-1 scores, in the sense of performing better when the reference summary is larger than average. This shows that the BM25 filter in this dataset is performing well by cutting unnecessary sentences inside each cluster, raising the overall precision levels.

5.4.3 Results and discussion - RulingBR

The results for the RulingBR dataset are presented in Figures 5.4, 5.5 and 5.6 for PoSum, FoSum, and RoSum summaries, respectively. The tables provide ROUGE-1, ROUGE-2 and ROUGE-L precision, f-measure, and recall values according to the reference summary length intervals considered..

Figure 5.4: PoSum scores (RulingBR)

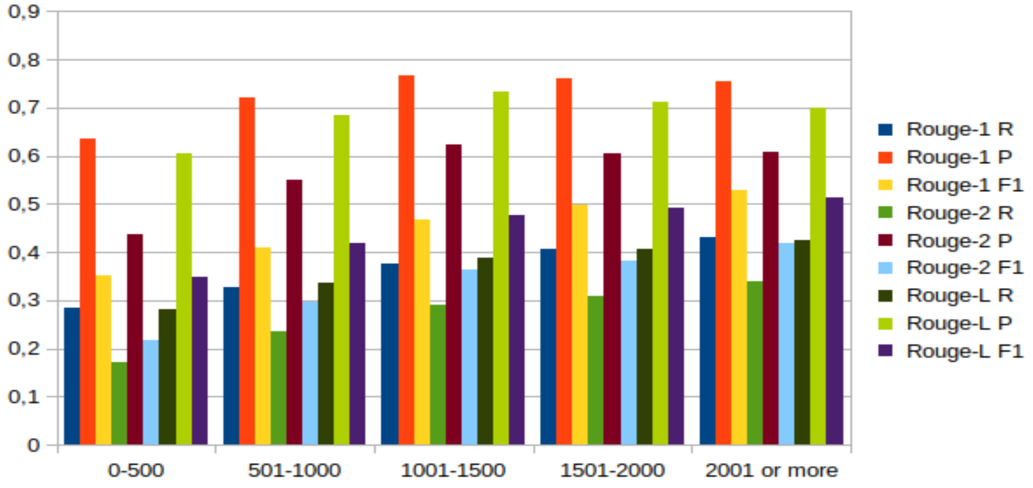


Figure 5.5: FoSum scores (RulingBR)

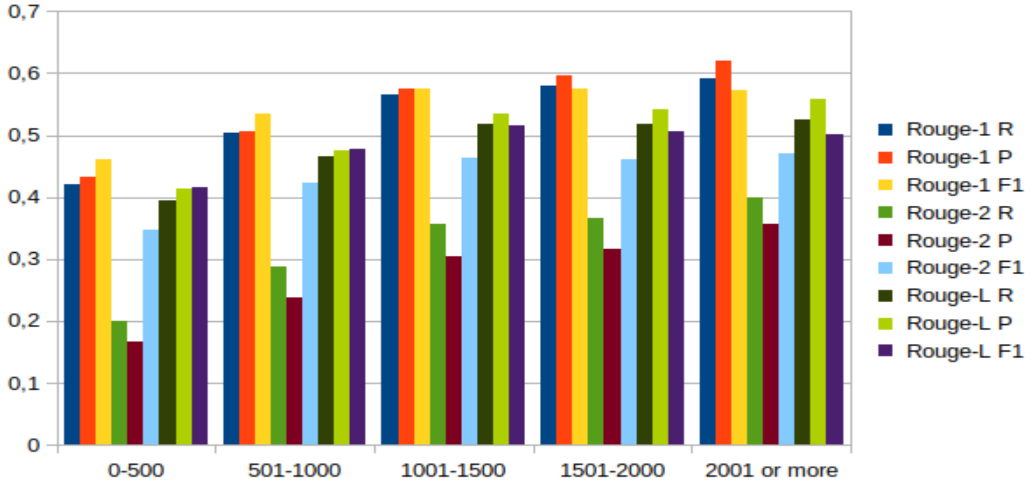
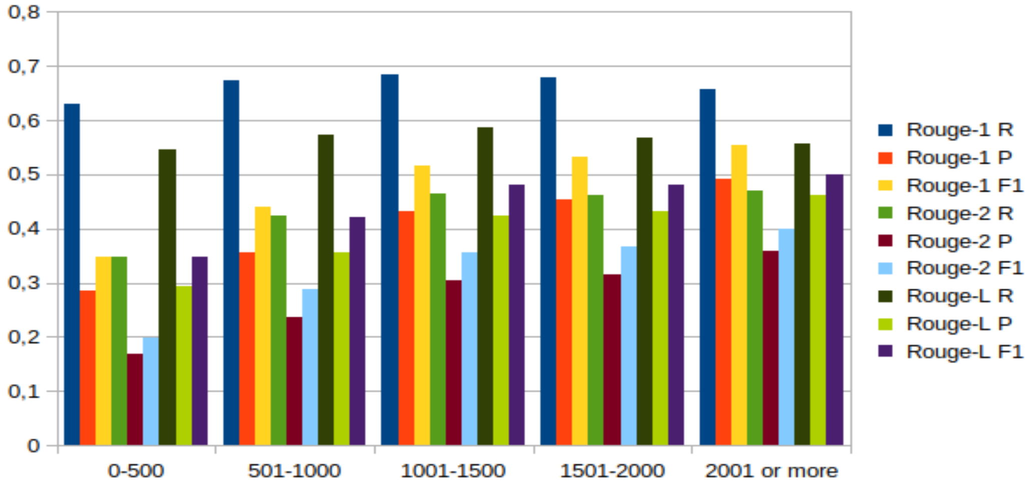


Figure 5.6: RoSum scores (RulingBR)



As depicted in Figure 5.4, using the PoSum strategy, BB25HLegalSum precision scores slighted dropped when comparing the 1001-1500 scores with the 1501-2000 and 2001 or more scores. All other scores increased according to refSum length.

Figure 5.5 shows that the f-measure scores increase until 1001-1500 and from there onwards, the f-measure results remain unchanged. Other scores increased according to refSum length.

As Figure 5.6 shows, using the RoSum strategy, BB25HLegalSum f-measure and precision values increased according to the refSum length, while recall scores dropped when comparing the 1001-1500 scores with the 1501-2000 and 2001 or more scores.

In general, we can see that the length of the reference summary does impact the scores, especially until the 1001-1500 refSum characters range. We can see that, when comparing all 0-500 intervals with the 2001 or more interval results, that ROUGE-1, ROUGE-2 and ROUGE-L are impacted positively with longer reference summary lengths. This shows room for improvement in this dataset, in the sense that the BM25 filter, in this dataset, could have a better performance by cutting unnecessary sentences inside each cluster, raising the overall precision levels.

5.5 Experiment 3

5.5.1 Method

This experiment targets the last research question "*Which type of summary is more suitable in the legal documents context concerning its readability: focused on precision,* *on recall, or f-measure?*". In order to assess these types of summaries and how they are presented, we have selected specific bills from a set of US test data and RullingBR datasets, and asked three jurists to assess the quality of the summaries produced by BB25HLegalSum for creating accurate and useful legal document summaries.

First, the same 3 jurists were given 50 three-colored highlighted bills from the BillSum US dataset, in the same formatting as Figure 5.7), as well as their reference summary for comparison. The same assessment was performed later using 8 highlighted rulings from rulingBR, using the same formatting as Figure 5.7) to read.

The highlights were produced using the sentences of the respective PoSum, RoSum and FoSum summaries. We chose to present the three types of summaries within a single document, using three different colors, one for each criterion-focused summary: green for PoSum, blue for FoSum, and red for RoSum. Upon the reading, they were asked to answer a number of questions.

The questions were:

- (1) Regarding the reference summary, do the three colors' highlights outline the main arguments?

- (2) Regarding the highlights in GREEN, do the highlights in BLUE or RED seem to bring new relevant information?

- (3) Based on the highlights alone, can you understand the context, only the main arguments, or both?

- (4) Among the three forms of highlighting, which method of highlighting do you believe is the most suitable for jurists and why? Consider the following options: (a) emphasis only in GREEN; (b) highlight in GREEN + BLUE; (c) griffin in GREEN + BLUE + RED. Write your observations in a few lines.

The documents were selected according to the following criteria:

- BillSum US: We selected 50 bills and we compared the respective PoSum and FoSum summaries, seeking the ones in which the trade-offs between completeness and conciseness were more clear. More specifically, we selected bills for which the PoSum and FoSum summaries differed in content, meaning they did not correspond to the same generated summary. Figure 5.7 displays an example of a document assessed in this experiment. We have presented to the jurists the same format table.

The first column shows the reference summary, while the second column displays the bill's text highlighted with different colors.

- RulingBR: we selected rulings based on the criteria that some kind of material judging was done and that the ruling would not be preliminarily discarded. In this sense, most rulings have the keyword "princípio", which is indicative that there was a more thorough analysis on behalf of the Supreme Federal Court. We have selected 8 random rulings out of the resulting rulings. Figure 5.8 shows an example of a highlighted document, and the respective reference summary.

5.5.2 Results and Discussion - BillSum

All participants answered *yes* to the first and second questions. While answering question 1, one participant said that in some cases, the colors did not bring important sentences that could be added. This brings room for future improvement in the summarization solution.

Regarding question 2, one of the jurists made the comment that the highlights help to understand the context better. For example, the PoSum exposes the topic, while the RoSum complements it with more details. Another jurist said that it is possible to understand the context and main arguments of the case. This jurist pointed out that there is a tendency in Brazilian Law to draft very long-winded decisions, but the main arguments of the case can often be summarized in a few sentences, and that the griffins expressed that clearly.

Regarding the third question, all participants agreed on the possibility of inferring context and the main arguments from the highlights alone. One participant pointed out that, since jurists sometimes prepare very long decisions, there is difficulty in choosing what to put in the "summary" (handwritten summary). There is a fine line between summarizing with the main arguments, which results in the exclusion of the relevant arguments in favor of producing a smaller summary. With the proposed summarization, the colors highlight not only what appears on the reference summary but also relevant information that was not included there. The proposed method eliminates the fear of only reading the reference summary and running the risk of missing some important information that was left out.

Also concerning the third question, another legal expert pointed out that the impor-

Figure 5.7: Bill 1678: Reference summary and highlighted bill according to the three strategies.

Reference summary	**Precision focused (PoSum color), F-measure focused (PoSum color + FoSum color) and Recall focused (PoSum color + FoSum color + RoSum color) summaries**
Combating Climate Change Through Individual Action Act of 2008 - Amends the Internal Revenue Code to allow tax credits for: (1) 30 of carbon sequestration and soil conservation expenditures made by taxpayers engaged in the business of farming. (2) 10 of qualifying planting expenditures, including expenditures for the purchase and planting of any tree, plant, shrub, or bush, and the purchase and installation of a vegetated roof system. (3) the conversion of cropland to pasture for grazing purposes or to grassland or rangeland, and (4) certain types of reforestation and afforestation of land.	SECTION 1. SHORT TITLE. This Act may be cited as the "Combating Climate Change Through Individual Action Act of 2008". SEC. 2. FINDINGS. Congress finds the following: (1) Agricultural, grassland, and forestry practices play an essential role in capturing atmospheric carbon and sequestering it as soil organic matter. (2) Released carbon can be captured through improved grassland management, tree planting, forest preservation, and enhanced agronomic and irrigation practices. (3) Promoting increased natural carbon sinks could have a significant impact on the world's projected carbon emissions from the burning of fossil fuels. (4) Certain agricultural and forestry practices can reduce greenhouse gases: (A) avoiding emissions by maintaining existing carbon storage in trees and soils; (B) increasing carbon storage by, e.g., tree planting, conversion from conventional to conservation tillage practices on agricultural lands; (5) The large potentials exist through known cropping and land management practices such as adoption of no-till, reduced fallow and use of cover crops, and conservation set-asides with perennial grasses and trees. SEC. 3. CARBON SEQUESTRATION AND SOIL CONSERVATION CREDIT. (a) In General.–Subpart D of part IV of subchapter A of chapter 1 of the Internal Revenue Code of 1986 (relating to business related credits) is amended by adding at the end the following new section: "SEC. 45O. CARBON SEQUESTRATION AND SOIL CONSERVATION. "(a) In General.–For purposes of section 38, in the case of a taxpayer engaged in the business of farming, the credit determined under this section for the taxable year is an amount equal to 30 percent of the qualified carbon sequestration and soil conservation expenditures for the taxable year which are paid or incurred with respect to the land used in such farming. "(b) Limitation.–The credit allowed with respect to a taxpayer under this section for a taxable year shall not exceed an amount equal to $10,000, reduced by the sum of the credits allowed with respect to the taxpayer under subsection (a) for all preceding taxable years. "(c) Qualified Carbon Sequestration and Soil Conservation Expenses.–For purposes of this section– " (...) –The amendments made by this section shall apply to expenditures paid or incurred after December 31, 2008. SEC. 4. QUALIFYING PLANTING EXPENDITURE CREDIT. (a) In General.–Subpart B of part IV of subchapter A of chapter 1 of the Internal Revenue Code of 1986 (relating to other credits) is amended by adding at the end the following new section: "SEC. 30D. QUALIFIED PLANTING EXPENDITURE CREDIT. "(a) Allowance of Credit.–There shall be allowed as a credit against the tax imposed by this chapter for the taxable year an amount equal to 10 percent of the qualified planting expenditures of the taxpayer for the taxable year. "(b) Limitations.–The amount taken into account under subsection (a) for any taxable year shall not exceed– "(1) in the case of expenditures paid or incurred by the taxpayer with respect to an area which is included under section 121 as part of the taxpayer's principal residence, $5,000, "(2) in the case of expenditures paid or incurred by the taxpayer in the course of, or with respect to, a trade or business carried on by the taxpayer, $50,000, and "(3) in any other case, zero. "(c) Qualified Planting Expenditures.–For purposes of this section– "(1) In general.–The term 'qualifying planting expenditures' means expenditures paid or incurred– "(A) for the purchase and planting of any tree, plant, shrub, or bush which meets the requirements of paragraph (2), and "(B) for the purchase and installation of a vegetated roof system. Such term shall not include expenditures relating to any property which is held by the taxpayer for use in a trade or business or for the production of income, or which is property described in section 1221(a)(1) in the hands of the taxpayer. "(2) Trees, plants, shrubs, or bushes.–A tree, plant, shrub, or bush satisfies the requirements of the paragraph if such tree, plant, shrub, or bush is certified, in accordance with guidance prescribed by the Secretary (after consultation with the Administrator of the Environmental Protection Agency and the Secretary of Agriculture), to be quick-growing, appropriate for the region in which it is planted, and effective in capturing carbon. "(3) Vegetated roof system.–The term 'vegetated roof system' means a system by which vegetation growing in a substrate is integrated with the roof (or portion thereof) of a building owned by the taxpayer. "(d) Application With Other Credits.–The credit allowed under subsection (a) for any taxable year shall not exceed the excess (if any) of– "(1) the regular tax liability (as defined in section 26(b)) reduced by the sum of the credits allowable under subpart A and sections 27, 30, 30B, and 30C, over "(2) the tentative minimum tax for the taxable year. "(e) Definition and Special Rules.–For purposes of this section– ' (...) –Not later than 180 days after the date of the enactment of this Act, the Secretary of the Treasury, in consultation with the Department of Agriculture, shall establish an appropriate tax credit, with respect to land located in the United States, for– (1) the conversion of cropland to pasture for grazing purposes or to grassland or rangeland, and (2) reforestation and afforestation of land– (...) –The Secretary shall provide– (A) an appropriate basis adjustment for property with respect to which such credit is allowed, and (B) rules disallowing such deductions and other credits as may be appropriate to avoid allowing additional tax benefits for the same conservation method or expenses. (c) Effective Date.–The credit established by the Secretary shall apply to taxable years beginning after December 31, 2008. SEC. 6. CARBON SEQUESTRATION CREDIT REPORT. (a) In General.–In the case of any substantial change in the carbon sequestration market (including the enactment into law of a carbon cap and trade program), the Secretary of the Treasury shall, in consultation with any appropriate Federal officers, study such change and any effect of such change on the efficiency of, and need for, the credits allowed under section 5 of this Act and sections 45O and 30D of the Internal Revenue Code of 1986. (b) Report.–As soon as practicable after sufficient opportunity to observe the effect of such change in the carbon sequestration market, the Secretary shall submit a report to Congress containing the results of the study conducted under subsection (a) and any recommendations of the Secretary for modifying such credits based on such results.

tance of specific information can vary depending on the intended audience. For instance, consider a bill that addresses funding for veterans. When the audience is legislators, details regarding how the fund will be established, its composition, and its utilization are critical factors. However, when the audience consists of regular users, the primary con-

Figure 5.8: Ruling 224: Reference summary and highlighted ruling according to the three strategies.

Reference summary	**Precision focused (PoSum color), F-measure focused (PoSum color + FoSum color) and Recall focused (PoSum color + FoSum color + RoSum color) summaries**
PENAL E PROCESSUAL PENAL. AGRAVO REGIMENTAL EM HABEAS CORPUS. HC SUBSTITUTIVO DE RECURSO ORDINÁRIO CONSTITUCIONAL. COMPETÊNCIA DO SUPREMO TRIBUNAL FEDERAL PARA JULGAR HABEAS CORPUS: CF. ART. 102, I, "D" E "I". ROL TAXATIVO. MATÉRIA DE DIREITO ESTRITO. INTERPRETAÇÃO EXTENSIVA: PARADOXO. ORGANICIDADE DO DIREITO. FURTO (ART. 155, CAPUT, DO CP). REINCIDÊNCIA NA PRÁTICA CRIMINOSA. PRINCÍPIO DA INSIGNIFICÂNCIA. INAPLICABILIDADE. FURTO FAMÉLICO. ESTADO DE NECESSIDADE X INEXIGIBILIDADE DE CONDUTA DIVERSA. AGRAVO REGIMENTAL EM HABEAS CORPUS A QUE SE NEGA PROVIMENTO. 1. O princípio da insignificância incide quando presentes, cumulativamente, as seguintes condições objetivas: (a) mínima ofensividade da conduta do agente, (b) nenhuma periculosidade social da ação, (c) grau reduzido de reprovabilidade do comportamento, e (d) inexpressividade da lesão jurídica provocada. 2. A aplicação do princípio da insignificância deve, contudo, ser precedida de criteriosa análise de cada caso, a fim de evitar que sua adoção indiscriminada constitua verdadeiro incentivo à prática de pequenos delitos patrimoniais. 3. O valor da res furtiva não pode ser o único parâmetro a ser avaliado, devendo ser analisadas as circunstâncias do fato para decidir-se sobre seu efetivo enquadramento na hipótese de crime de bagatela, bem assim o reflexo da conduta no âmbito da sociedade. 4. In casu, O paciente foi condenado pela prática do crime de furto (art. 155, caput, do Código Penal) por ter subtraído 4 (quatro) galinhas caipiras, avaliadas em R$ 40,00 (quarenta reais). As instâncias precedentes deixaram de aplicar o princípio da insignificância em razão de ser o paciente contumaz na prática do crime de furto. 5. Trata-se de condenado reincidente na prática de delitos contra o patrimônio. Destarte, o reconhecimento da atipicidade da conduta do recorrente, pela adoção do princípio da insignificância, poderia, por via transversa, imprimir nas consciências a ideia de estar sendo avalizada a prática de delitos e de desvios de conduta. 6. O furto famélico subsiste com o princípio da insignificância, posto não integrarem binômio inseparável. É possível que o reincidente cometa o delito famélico que induz ao tratamento penal benéfico. 7. In casu, o paciente é conhecido - consta na denúncia - por "Fernando Gatuno", alcunha sugestiva de que se dedica à prática de crimes contra o patrimônio; aliás, conforme comprovado por sua extensa ficha criminal, sendo certo que a quantidade de galinhas furtadas (quatro), é apta a indicar que o fim visado pode não ser somente o de saciar a fome à falta de outro meio para conseguir alimentos. 8. Agravo regimental em habeas corpus a que se nega provimento.	Vistos, relatados e discutidos estes autos, acordam os Ministros da Primeira Turma do Supremo Tribunal Federal, sob a Presidência do Senhor Ministro Luiz Fux, na conformidade da ata de julgamento e das notas taquigráficas, por maioria de votos, em negar provimento ao agravo regimental, nos termos do voto do Relator, vencidos o Senhor Ministro Marco Aurélio e a Senhora Ministra Rosa Weber. (...) 1. O princípio da insignificância incide quando presentes, cumulativamente, as seguintes condições objetivas: (a) mínima ofensividade da conduta do agente, (b) nenhuma periculosidade social da ação, (c) grau reduzido de reprovabilidade do comportamento, e (d) inexpressividade da lesão jurídica provocada. 2. A aplicação do princípio da insignificância deve, contudo, ser precedida de criteriosa análise de cada caso, a fim de evitar que sua adoção indiscriminada constitua verdadeiro incentivo à prática de pequenos delitos patrimoniais. 3. O valor da res furtiva não pode ser o único parâmetro a ser avaliado, devendo ser analisadas as circunstâncias do fato para decidir-se sobre seu efetivo enquadramento na hipótese de crime de bagatela, bem assim o reflexo da conduta no âmbito da sociedade. 4. In casu, O paciente foi condenado pela prática do crime de furto. As instâncias precedentes deixaram de aplicar o princípio da insignificância em razão de ser o paciente contumaz na prática do crime de furto. 5. Trata-se de condenado reincidente na prática de delitos contra o patrimônio. Destarte, o reconhecimento da atipicidade da conduta do recorrente, pela adoção do princípio da insignificância, poderia, por via transversa, imprimir nas consciências a ideia de estar sendo avalizada a prática de delitos e de desvios de conduta. 6. O furto famélico subsiste com o princípio da insignificância, posto não integrarem binômio inseparável. É possível que o reincidente cometa o delito famélico que induz ao tratamento penal benéfico. 7. In casu, o paciente é conhecido - consta na denúncia por 'Fernando Gatuno', alcunha sugestiva de que se dedica à prática de crimes contra o patrimônio; aliás, conforme comprovado por sua extensa ficha criminal, sendo certo que a quantidade de galinhas furtadas (quatro), é é apta a indicar que o fim visado pode não ser somente o de saciar a fome à falta de outro meio para conseguir alimentos. 8. Habeas corpus a que se nega seguimento. Colhe-se dos autos que o paciente foi denunciado como incurso nas sanções do artigo 155, caput, c/c o artigo 61, inciso I, ambos do Código Penal, por ter subtraído 4 (quatro) galinhas caipiras do quintal de uma residência. O valor total dos bens subtraídos foi avaliado em R$ 40,00 (quarenta reais). Concluída a instrução criminal, o paciente foi condenado a 1 (um) ano de reclusão, em regime inicial semiaberto, e ao pagamento de 10 (dez) dias-multa. A defesa interpôs apelação. (...) Requer, ao final, o provimento do recurso a fim de que seja concedida a ordem de habeas corpus no sentido de determinar a aplicação do princípio da insignificância e, por conseguinte, absolver o paciente da prática do crime de furto. É o relatório. (...) Em que pese haver entendimento de que somente devem ser considerados critérios objetivos para o reconhecimento dessa causa supralegal de extinção da tipicidade, a prudência recomenda que se leve em conta a obstinação do agente na prática delituosa, a fim de evitar que a impunidade o estimule a continuar trilhando a senda criminosa. In casu, o paciente foi condenado a 1 (um) ano de reclusão, em regime inicial semiaberto, pela prática do crime de furto (artigo 155 do CP), por ter subtraído 4 (quatro) galinhas caipiras do quintal de uma residência. O valor total dos bens subtraídos foi avaliado em R$ 40,00 (quarenta reais). Verifica-se, que, as, instâncias, precedentes, deixaram, de, aplicar, o, princípio, da, insignificância, em, razão, de, ser, o, paciente, contumaz, na, prática, do, crime, de, furto. (...) Destarte, o reconhecimento da atipicidade da conduta do recorrente, pela adoção do princípio da insignificância, poderia, por via transversa, imprimir nas consciências a ideia de estar sendo avalizada a prática de delitos e de desvios de conduta. Verifica-se que o julgamento apontado como paradigma pelo recorrente – HC 108.872, Segunda Turma, Relator o Ministro Gilmar Mendes – foi proferido em 06.09.11. Todavia, ambas, as, Turmas, deste, Supremo, Tribunal, em, julgados, mais, recentes, têm, sedimentado, o, entendimento, no, sentido, de, que, a, reincidência, na, prática, criminosa, obsta, a, aplicação, do, princípio, da, insignificância, sob, pena, de, incentivar-se, a, prática, de, pequenos, delitos. (...) 1. Para a incidência do princípio da insignificância, devem ser relevados o valor do objeto do crime e os aspectos objetivos do fato, tais como, a mínima ofensividade da conduta do agente, a ausência de periculosidade social da ação, o reduzido grau de reprovabilidade do comportamento e a inexpressividade da lesão jurídica causada. 2. Nas circunstâncias do caso, não se pode aplicar o princípio em razão da reincidência dos Pacientes. 3. O valor do bem furtado (R$ 350,00, trezentos e cinquenta reais) corresponde a mais de 50% do valor do salário mínimo nacional, à época do crime (R$ 465,00, quatrocentos e sessenta e cinco reais, Lei n. 11.944/ 2009). 4. Ordem denegada" – Sem grifos no original. (HC 113.196, Primeira Turma, Relatora a Ministra Cármen Lúcia, DJe de 1o.10.12) (...) O SENHOR MINISTRO MARCO AURÉLIO - Um ano de reclusão e pagamento de dez dias-multa. O SENHOR MINISTRO LUÍS ROBERTO BARROSO - Portanto, regime aberto. Eu acompanho Vossa Excelência, Presidente. 24/09/2013 PRIMEIRA TURMA AG.REG. NO HABEAS CORPUS 115.850 MINAS GERAIS VOTO O SENHOR MINISTRO DIAS TOFFOLI: Senhor Presidente, sendo reincidente, eu vou me manter coerente com o que tenho despachado monocraticamente. Acompanho Vossa Excelência.

cern typically revolves around how the fund will be utilized.

Considering the fourth question for this specific dataset, most participants agreed, in all cases, that the highlighting method that is most fitting to jurists would be the method with the three colors together (PoSum color + FoSum color + RoSum color). One jurist said that the red color tends to bring unnecessary details for a quick reading of the case, which would be more relevant if there is an interest in a more in-depth reading by highlighting secondary arguments and sometimes opposing the final decision as they highlight the sequence of discussions to reach the final decision.

Thus, from our observation, the PoSum and FoSum summaries are shorter because they usually lack key arguments. In a legal document context, having a higher recall as a suitable criterion is important because failing to identify a relevant piece of information can have serious consequences, such as missing a key piece of context or failing to make a critical argument. Another important remark is that highlighting with multiple colors allows the reader to select pieces of information in an easier, faster, and more intuitive manner.

5.5.3 Results and Discussion - RulingBR

Considering the documents from RulingBR, all participants also answered *yes* to the first and second questions. One of the jurists observed that the highlighted segments aid in achieving a more comprehensive grasp of the context.

Regarding the third question, all participants agreed that the colored highlights bring new information compared to the reference summary, enhancing the elucidation of the rulings. A participant pointed out that the precision-focused summaries (i.e. PoSum highlights only) concentrate solely on the "ultimate conclusion" without encompassing the underlying rationale or grounds.

Considering the fourth question, two participants agreed, in all cases, that the highlighting method that seems most fitting for this specific dataset would be the method with the two colors together (GREEN + BLUE), which corresponds to the full contents of the FoSum. They pointed out that the GREEN + BLUE + RED method might cause confusion in the reader since it brings jurisprudence and arguments contrary to the final decision (but related to the ruling in question). One jurist differed from this opinion, who described that there are arguments that are both againts and in favor of the final decision in the RED color. Hence, to this jurist, the highlighting method with all colors together seems more fitting for a better understanding of the rulings, but if one needs to understand only the main arguments, the GREEN + BLUE is the more fitting method.

We believe that the legal workers concluded that the FoSum is the most appropriate type of summary (in the detriment of RoSum, as in the previous assessment) because the rulings in the RulingBR dataset are more complex. Often they present contrasting points of view, and are longer, more prolix, and wordy in comparison to the bills in the BillSum dataset (which are commonly written towards one line of argumentation and in a less lengthy manner).

In any case, the conclusion that can be made from these two assessments is that the PoSum (precision oriented summary method), which focuses on conciseness and neglects completeness, is not the recommended choice for the legal document context, because missing key pieces of context or failing to make a critical argument would be unwelcome. Depending on how the dataset is organized, the most fitting methods might fall either in FoSum (as concluded in this dataset), or in RoSum (as shown in billSum), which present a higher coverage/completeness of arguments.

In summary, multiple summaries can be derived from a single document, and a tool could offer the capability to choose colors strategically, facilitating the reader in achieving their specific goals.

6 CONCLUSIONS AND FUTURE WORK

This dissertation makes two main contributions. Firstly, it presents a method that leverages BERT and BM25 to generate legal document summaries. This method outperforms baselines (ANAND; WAGH, 2019; MIHALCEA; TARAU, 2004; ERKAN; RADEV, 2004; FEIJÓ, 2021) in two different legal datasets (BillSum and RulingBR). The second is a presentation method for the generated summaries using different colors that highlight in their original context the importance of sentences according to distinct points of view (precision vs. coverage). This presentation enabled to assess the utility and trade-offs of consiness and completeness in generated summaries. Legal workers also positively assessed this presentation. Part of this work was summarized in a paper published in an international conference (RANLP - Qualis A3) (BONALUME; BECKER, 2023).

Concerning the advantages of the method are the fact that we can generate different summaries from a single document and a tool could allow the selection of colors in a way to help the reader reach his objectives. We also propose two different ways of highlighting: subsidiary and intersectional.

Our method uses BERT, which is a black box model. In recent years, the utilization of black box models has led to improved summarization capabilities compared to earlier technologies. However, these models introduce limitations, such as the complexity of model interpretation and the substantial resources and time consumption required for their processing prior to generating summaries.

Our work has not addressed the accessibility of the proposed summary presentation. The choice of colors needs to consider the constraints of users with special accessibility needs (e.g. colorblind). Another limitation is the fact that we did not delve into the differences between intersectional and subsidiary highlighting and when they should be preferred and why.

For future work, some ideas would be resource-efficient processing (in order to find ways to mitigate the computational demands of transformer or similar models for more efficient summarization processing) and the increase in other domain-specific summaries, as well as other experiments regarding the presentation of summarized texts.

REFERENCES

AKLOUCHE, B.; BOUNHAS, I.; SLIMANI, Y. Bm25 beyond query-document similarity. In: SPRINGER. **International Symposium on String Processing and Information Retrieval**. [S.l.], 2019. p. 65–79.

ALLAHYARI, M. et al. Text summarization techniques: a brief survey. **arXiv preprint arXiv:1707.02268**, 2017.

ALTHAMMER, S. et al. Dossier@ coliee 2021: Leveraging dense retrieval and summarization-based re-ranking for case law retrieval. **arXiv preprint arXiv:2108.03937**, 2021.

AN, C. et al. Retrievalsum: A retrieval enhanced framework for abstractive summarization. **arXiv preprint arXiv:2109.07943**, 2021.

ANAND, D.; WAGH, R. Effective deep learning approaches for summarization of legal texts. **Journal of King Saud University-Computer and Information Sciences**, Elsevier, 2019.

ASKARI, A. et al. Leibi@ coliee 2022: Aggregating tuned lexical models with a cluster-driven bert-based model for case law retrieval. **arXiv preprint arXiv:2205.13351**, 2022.

BONALUME, L.; BECKER, K. Bb25hlegalsum: Leveraging bm25 and bert-based clustering for the summarization of legal documents. In: **Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2023)**. [S.l.: s.n.], 2023. p. 255–263.

BOOMIJA, M. D.; PHIL, M. Comparison of partition based clustering algorithms. **Journal of Computer Applications**, v. 1, n. 4, p. 18–21, 2008.

DALAL, S.; SINGHAL, A.; LALL, B. Lexrank and pegasus transformer for summarization of legal documents. In: SPRINGER. **Machine Intelligence Techniques for Data Analysis and Signal Processing: Proceedings of the 4th International Conference MISP 2022, Volume 1**. [S.l.], 2023. p. 569–577.

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

EL-KASSAS, W. S. et al. Automatic text summarization: A comprehensive survey. **Expert systems with applications**, Elsevier, v. 165, p. 113679, 2021.

ERKAN, G.; RADEV, D. R. Lexrank: Graph-based lexical centrality as salience in text summarization. **Journal of artificial intelligence research**, v. 22, p. 457–479, 2004.

ETHAYARAJH, K. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. **arXiv preprint arXiv:1909.00512**, 2019.

FEIJO, D.; MOREIRA, V. Summarizing legal rulings: Comparative experiments. In: **Proceedings of the international conference on recent advances in natural language processing (RANLP 2019)**. [S.l.: s.n.], 2019. p. 313–322.

FEIJÓ, D. d. V. Summarizing legal rulings. 2021.

FEIJÓ, D. de V.; MOREIRA, V. P. Rulingbr: A summarization dataset for legal texts. In: SPRINGER. **Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13**. [S.l.], 2018. p. 255–264.

FLORIDI, L.; CHIRIATTI, M. Gpt-3: Its nature, scope, limits, and consequences. **Minds and Machines**, Springer, v. 30, p. 681–694, 2020.

FURNITUREWALA, S. et al. **Legal text classification and summarization using transformers and joint text features**. [S.l.]: FIRE, 2021.

HUANG, Y. et al. Legal public opinion news abstractive summarization by incorporating topic information. **International Journal of Machine Learning and Cybernetics**, Springer, v. 11, p. 2039–2050, 2020.

JAIN, D.; BORAH, M. D.; BISWAS, A. Summarization of legal documents: Where are we now and the way forward. **Computer Science Review**, Elsevier, v. 40, p. 100388, 2021.

JAIN, D.; BORAH, M. D.; BISWAS, A. Improving kullback-leibler based legal document summarization using enhanced text representation. In: IEEE. **2022 IEEE Silchar Subsection Conference (SILCON)**. [S.l.], 2022. p. 1–5.

JAIN, D.; BORAH, M. D.; BISWAS, A. Bayesian optimization based score fusion of linguistic approaches for improving legal document summarization. **Knowledge-Based Systems**, Elsevier, v. 264, p. 110336, 2023.

JAIN, D.; BORAH, M. D.; BISWAS, A. A sentence is known by the company it keeps: Improving legal document summarization using deep clustering. **Artificial Intelligence and Law**, Springer, p. 1–36, 2023.

KANAPALA, A.; PAL, S.; PAMULA, R. Text summarization from legal documents: a survey. **Artificial Intelligence Review**, Springer, v. 51, p. 371–402, 2019.

KODINARIYA, T. M.; MAKWANA, P. R. et al. Review on determining number of cluster in k-means clustering. **International Journal**, v. 1, n. 6, p. 90–95, 2013.

KORNILOVA, A.; EIDELMAN, V. Billsum: A corpus for automatic summarization of us legislation. **arXiv preprint arXiv:1910.00523**, 2019.

KRIEGEL, H.-P. et al. Density-based clustering. **Wiley interdisciplinary reviews: data mining and knowledge discovery**, Wiley Online Library, v. 1, n. 3, p. 231–240, 2011.

LICARI, D. et al. Legal holding extraction from italian case documents using italian-legal-bert text summarization. 2023.

LIN, C.-Y. Rouge: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. [S.l.: s.n.], 2004. p. 74–81.

LIU, Y. Fine-tune bert for extractive summarization. **arXiv preprint arXiv:1903.10318**, 2019.

MERCHANT, K.; PANDE, Y. Nlp based latent semantic analysis for legal text summarization. In: IEEE. **2018 international conference on advances in computing, communications and informatics (ICACCI)**. [S.l.], 2018. p. 1803–1807.

MIHALCEA, R.; TARAU, P. Textrank: Bringing order into text. In: **Proceedings of the 2004 conference on empirical methods in natural language processing**. [S.l.: s.n.], 2004. p. 404–411.

MORO, G. et al. Efficient memory-enhanced transformer for long-document summarization in low-resource regimes. **Sensors**, MDPI, v. 23, n. 7, p. 3542, 2023.

NASEEM, U. et al. Transformer based deep intelligent contextual embedding for twitter sentiment analysis. **Future Generation Computer Systems**, Elsevier, v. 113, p. 58–69, 2020.

NENKOVA, A.; MCKEOWN, K. et al. Automatic summarization. **Foundations and Trends® in Information Retrieval**, Now Publishers, Inc., v. 5, n. 2–3, p. 103–233, 2011.

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

PIRES, T.; SCHLINGER, E.; GARRETTE, D. How multilingual is multilingual bert? **arXiv preprint arXiv:1906.01502**, 2019.

POLSLEY, S.; JHUNJHUNWALA, P.; HUANG, R. Casesummarizer: A system for automated summarization of legal texts. In: **Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations**. [S.l.: s.n.], 2016. p. 258–262.

POWERS, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. **Journal of Machine Learning Technologies**, v. 2, n. 1, p. 37–63, 2011.

ROBERTSON, S.; ZARAGOZA, H. et al. The probabilistic relevance framework: Bm25 and beyond. **Foundations and Trends® in Information Retrieval**, Now Publishers, Inc., v. 3, n. 4, p. 333–389, 2009.

ROSA, G. M. et al. Yes, bm25 is a strong baseline for legal case retrieval. **arXiv preprint arXiv:2105.05686**, 2021.

ROY, N. et al. Note the highlight: incorporating active reading tools in a search as learning environment. In: **Proceedings of the 2021 conference on human information interaction and retrieval**. [S.l.: s.n.], 2021. p. 229–238.

SAXENA, A. et al. A review of clustering techniques and developments. **Neurocomputing**, Elsevier, v. 267, p. 664–681, 2017.

STAUDEMEYER, R. C.; MORRIS, E. R. Understanding lstm–a tutorial into long short-term memory recurrent neural networks. **arXiv preprint arXiv:1909.09586**, 2019.

THINSUNGNOENA, T. et al. The clustering validity with silhouette and sum of squared errors. **learning**, v. 3, n. 7, 2015.

VELMURUGAN, T.; SANTHANAM, T. A survey of partition based clustering algorithms in data mining: An experimental approach. **Information Technology Journal**, Asian Network for Scientific Information, 308-Lasani Town Sargodha Rd . . . , v. 10, n. 3, p. 478–484, 2011.

APPENDIX A — SUMÁRIO EXPANDIDO

Os processos legais em curso são uma preocupação comum que impacta os sistemas legais em todo o mundo. A quantidade de casos não resolvidos pode variar substancialmente com base no tamanho da população, na estrutura legal e no acúmulo de casos pendentes. Esse cenário motiva a pesquisa de técnicas computacionais que possam acelerar a análise judicial, selecionar casos semelhantes para julgamento em lotes ou identificar padrões que possam levar a decisões mais assertivas.

A implementação de um sistema automatizado para destacar informações-chave em documentos legais pode aliviar significativamente o ônus sobre os profissionais do direito, tornando suas tarefas de leitura mais agradáveis e menos árduas, podendo aumentar, consequentemente, a eficiência no processo de análise judicial. A sumarização automática de documentos legais para sintetizar sua essência é crucial nesse contexto.

Outro fator relevante é de que os algoritmos de sumarização de texto devem levar em consideração o público-alvo, o objetivo do sumário, bem como o gênero e o layout do texto original. Diante disso, a sumarização de texto encontra utilidade em diversas aplicações, como agregação de notícias, gerenciamento de documentos e sumarização de documentos legais.

A maioria dos trabalhos no domínio jurídico utiliza a sumarização extrativa para a criação de sumários, um conceito elucidado em (ANAND; WAGH, 2019) como "a geração de um sumário contendo um subconjunto de sentenças do texto original após a identificação das sentenças importantes". Foram exploradas várias técnicas para a sumarização extrativa de textos legais, incluindo relevância de palavras (POLSLEY; JHUNJHUNWALA; HUANG, 2016), modelos de classificação baseados em grafo (DALAL; SINGHAL; LALL, 2023; JAIN; BORAH; BISWAS, 2023a), modelos estatísticos (JAIN; BORAH; BISWAS, 2022; MERCHANT; PANDE, 2018) e *deep learning* (ANAND; WAGH, 2019). Mais recentemente, o modelo Bidirectional Encoder Representations from Transformers (BERT) (DEVLIN et al., 2018) tem sido utilizado na área jurídica (FURNITUREWALA et al., 2021), inspirado nos resultados de ponta alcançados na sumarização geral de texto extrativo (LIU, 2019).

Uma estratégia alternativa na área de documentos legais é o Best Match 25 (BM25), uma função de classificação comumente utilizada em recuperação de informações para determinar a similaridade de um documento em relação a uma consulta de pesquisa (ROBERTSON; ZARAGOZA et al., 2009). O uso combinado de BERT e

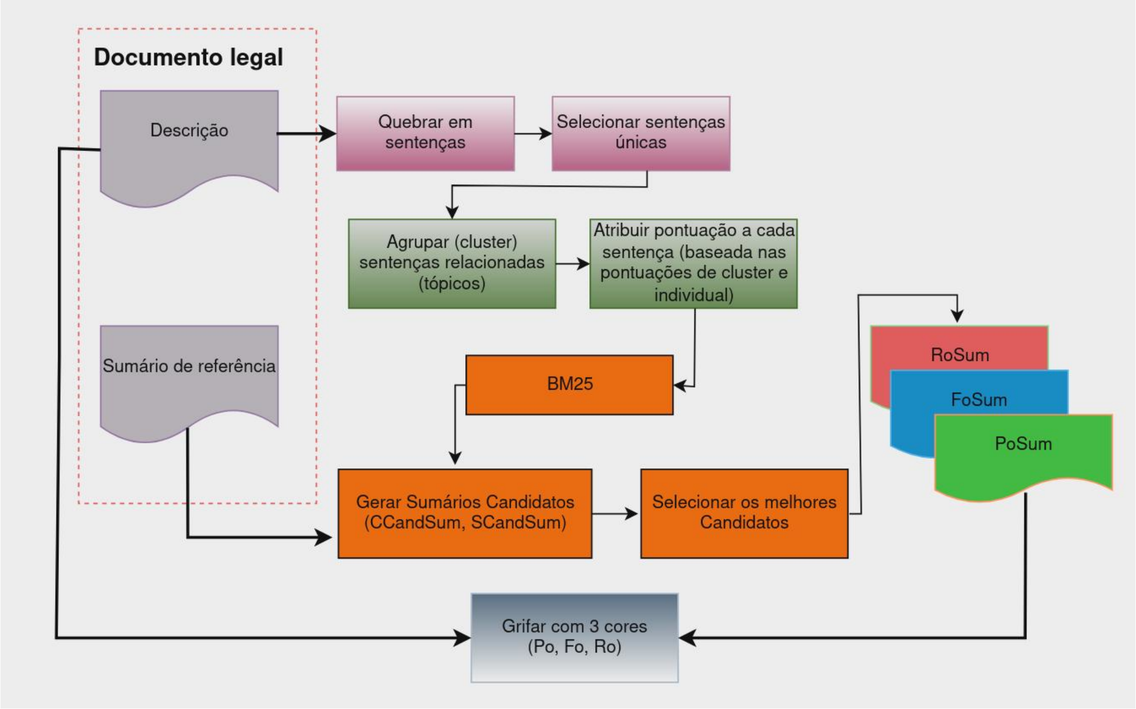
BM25 é recorrente para recuperação de informações em documentos legais (ASKARI et al., 2022; ALTHAMMER et al., 2021). No entanto, ainda está em estágios preliminares para a tarefa de sumarização de documentos legais. Cabe ressaltar que as potencialidades dessas técnicas podem ser combinadas para gerar sumários de alta qualidade e auxiliar na superação de desafios associados a métodos tradicionais, como *feature engineering* e documentos extensos. Isso porque o BERT captura relações intricadas entre palavras e frases de modo contextual, enquanto o BM25 opera como um eficaz algoritmo de recuperação de informações e atribuição de similaridade para sumarização de documentos.

De acordo com Jain, Borah e Biswas (2021), é necessário que haja mais análises sobre a legibilidade dos sumários gerados e como apresentá-los. Na área jurídica, a apresentação de sumários é abordada através de destaques (LICARI et al., 2023) e *heatmaps* (POLSLEY; JHUNJHUNWALA; HUANG, 2016), representando a relevância das sentenças dentro do documento original. Todavia, a relevância de uma sentença pode ser um aspecto secundário para os profissionais do direito, que podem estar mais interessados nos principais argumentos dentro de seu contexto.

Nesta dissertação, propomos o BB25HLegalSum (BERT + BM25 + Sumarização de Documentos Jurídicos com Destaques), um método para a sumarização extrativa de documentos jurídicos, esquematizado na Figura A.1. Esse método utiliza BERT e BM25 para identificar sentenças relevantes em um documento jurídico e combina clusters de sentenças para gerar sumários candidatos, que são selecionados usando métricas em comparação com um sumário de referência. Essa abordagem é importante na sumarização de documentos jurídicos, pois um sumário destacado, quando comparado a um sumário de referência estabelecido, facilita não apenas uma rápida avaliação dos pontos-chave do documento, mas também permite uma melhor compreensão dos argumentos subjacentes em torno do sumário de referência.

Desse modo, geramos sumários usando três estratégias para identificar as melhores partes de um documento, focando na precisão das sentenças selecionadas, na completude do texto (recall) e em um equilíbrio entre esses dois critérios. Outra característica distintiva do método é a apresentação do sumário gerado. Propomos uma abordagem de destaque que, usando cores diferentes (ou outras formas de diferenciação por questões de acessibilidade), representa as sentenças contidas nos sumários gerados de acordo com cada estratégia. Dessa forma, o usuário pode identificar e distinguir, em seu contexto original, as sentenças relevantes do documento de acordo com diferentes pontos de vista que enfatizam precisão, completude ou ambos. Resultados preliminares são discutidos

Figure A.1: Visão geral do método BB25HLegalSum


Documento legal
Descrição
Sumário de referência
Quebrar em sentenças
Selecionar sentenças únicas
Agrupar (cluster) sentenças relacionadas (tópicos)
Atribuir pontuação a cada sentença (baseada nas pontuações de cluster e individual)
BM25
Gerar Sumários Candidatos (CCandSum, SCandSum)
Selecionar os melhores Candidatos
RoSum
FoSum
PoSum
Grifar com 3 cores (Po, Fo, Ro)

em (BONALUME; BECKER, 2023).

Nossas avaliações revelaram resultados encorajadores. O nosso método supera resultados com trabalhos de referência em dois conjuntos de dados jurídicos, a saber BillSum (JAIN; BORAH; BISWAS, 2021) e RulingBR (FEIJÓ, 2021). Além disso, observamos que o comprimento do sumário de referência impacta o recall e a precisão dos sumários gerados, com a abordagem proposta apresentando melhores resultados para sumários de referência mais extensos. Por fim, um experimento qualitativo mostrou que, em um contexto de documentos jurídicos, a completude, em comparação com a precisão, é o critério mais importante para resumir, pois é mais crucial evitar a omissão de informações relevantes. Assim, o destaque com cores distintas permite identificar diferentes tipos de informações capturadas por cada estratégia.

As principais contribuições desta dissertação são:

1. um método que utiliza BERT e BM25 para gerar sumários de documentos jurídicos. Comparamos os resultados (ANAND; WAGH, 2019; MIHALCEA; TARAU, 2004; ERKAN; RADEV, 2004; FEIJÓ, 2021) em dois conjuntos de dados jurídicos diferentes (BillSum e RulingBR);

2. um método de apresentação para os sumários gerados usando cores diferentes, destacando em seu contexto original a importância das sentenças de acordo com

diferentes pontos de vista (precisão versus completude). Profissionais jurídicos avaliaram positivamente essa apresentação.

Obtivemos respostas encorajadoras para nossas perguntas de pesquisa e evidências de generalização para nossas descobertas. Os principais *insights* foram:

(a) o desempenho dos sumários gerados foi alto em relação ao recall e precisão;

(b) ao combinar BERT e BM25, o tamanho do sumário de referência impacta a performance em termos de recall e precisão;

(c) os diferentes modos de sumarizar (focado em precisão ou em recall, por exemplo) delineiam diferentes pontos de vista argumentativos;

(d) enquanto que os sumários focados em precisão trazem os argumentos centrais, os sumários focados em f-measure e em recall também trazem argumentos importantes que se relacionam com os principais argumentos;

(e) na avaliação feita por profissionais do direito, os sumários focados em f-measure e em recall foram avaliados como sendo mais relevantes do que os sumários focados em precisão.