UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Cinthia Mikaela de Souza

Proposta de uma abordagem para sumarização extrativa de textos científicos longos

Belo Horizonte
2022

Cinthia Mikaela de Souza

Proposta de uma abordagem para sumarização extrativa de textos científicos longos

Versão Final

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Renato Vimieiro
Coorientadora: Profa. Magali R. G. Meireles

Belo Horizonte
2022

Souza, Cinthia Mikaela de

S729p Proposta de uma abordagem para sumarização extrativa de textos científicos longos [recurso eletrônico] / Cinthia Mikaela de Souza — 2022.
1 recurso online (73 f. il, color.)

Orientador: Renato Vimieiro.
Coorientadora: Magali Rezende Gouvêa Meireles.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Departamento de Ciência da Computação, Instituto de Ciências Exatas
Referências: f. 57-61.

1. Computação – Teses. 2. Sumarização automática de textos – Teses. 3. Aprendizado de máquina multivisão– Teses. 4. Classificação– Teses. I Vimieiro, Renato. II. Meireles, Magali Rezende Gouvêa III. Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Computação. III.Título.

CDU 519.6*82.7(043)

Ficha Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

PROPOSTA DE UMA ABORDAGEM PARA SUMARIZAÇÃO EXTRATIVA DE TEXTOS CIENTÍFICOS LONGOS

CINTHIA MIKAELA DE SOUZA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores(a):

Prof. Renato Vimieiro - Orientador
Departamento de Ciência da Computação - UFMG

Profa. Magali Rezende Gouvêa Meireles - Coorientadora
Instituto de Ciências Exatas e Informática - PUC-MG

Prof. Rodrygo Luis Teodoro Santos
Departamento de Ciência da Computação - UFMG

Prof. Adriano Alonso Veloso
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 05 de dezembro de 2022.

Documento assinado eletronicamente por **Renato Vimieiro**, **Professor do Magistério Superior**, em 20/12/2022, às 13:39, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Rodrygo Luis Teodoro Santos**, **Professor do Magistério Superior**, em 24/12/2022, às 10:03, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Adriano Alonso Veloso**, **Professor do Magistério Superior**, em 24/12/2022, às 19:07, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Magali Rezende Gouvêa Meireles**, **Usuária Externa**, em 06/03/2023, às 17:33, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **1975467** e o código CRC **6AB45E35**.

Este trabalho é dedicado a todos que me apoiaram e ajudaram durante essa jornada. Sem vocês nada disso teria sido possível. A todos vocês, sou extremamente grata.

Resumo

A sumarização automática de textos é uma das soluções que permite aos usuários identificar as informações mais relevantes de um documento textual, consequentemente, reduzindo o tempo de busca pelas informações. O objetivo dessa técnica é condensar as informações de um texto em um resumo simples e descritivo, que dê ao leitor uma ideia geral do texto sem ter que ler todo o seu conteúdo. A maior parte da literatura em sumarização automática de texto se concentra em propor e aprimorar métodos de aprendizado profundo para tornar esses modelos aplicáveis no contexto de sumarização de textos longos. Infelizmente, esses modelos ainda possuem limitações no comprimento da sequência de entrada. Tal limitação pode levar a uma perda de informações que prejudica a qualidade dos resumos gerados. Por esta razão, propomos nessa dissertação uma nova abordagem de sumarização extrativa de textos longos. Temos duas hipóteses: (1) subdividir o problema de sumarização em problemas menores e resolvê-los, separadamente, e, posteriormente, combinar essas soluções pode trazer benefícios para a tarefa de sumarização de textos longos; (2) há outros atributos do texto que podem ser úteis na criação do resumo. Tendo isso em vista, nós modelamos o problema de sumarização de textos como um problema de classificação binária. Testamos diferentes algoritmos e mostramos que a sumarização multi-seção tem um desempenho superior à sumarização de seção única com um ganho de desempenho de, aproximadamente, 14% e 5% de BertScore para o conjunto de dados da Plos One e do ArXiv, respectivamente. Nós, também, avaliamos o desempenho do sumarizador proposto usando diferentes representações do texto e mostramos que a representação de visão única de atributos é a que obtém os melhores resultados. Isso mostra que, para a tarefa de sumarização extrativa de textos, os atributos selecionados para compor a visão de atributos permitem identificar melhor a importância das sentenças. Por fim, nós comparamos o método proposto com diferentes modelos do estado-da-arte em sumarização extrativa, abstrativa e híbrida e mostramos que a nossa abordagem supera esses modelos.

Palavras-chave: Sumarização extrativa de textos, Aprendizado Multi-visão, Classificação.

Abstract

Automatic text summarization is one of the solutions that allows users to identify the most relevant information in a textual document, consequently reducing the time to search for information. The objective of this technique is to condense the information of a text into a simple and descriptive summary, which gives the reader a general idea of the text without having to read all its content. Most of the literature in automatic text summarization focuses on proposing and improving Deep Learning methods in order to make these models applicable in the context of long text summarization. Unfortunately, these models still have limitations on the input sequence length. Such a limitation may lead to a loss of information that impairs the quality of the summaries generated. For this reason, we propose in this dissertation a new approach to extractive summarization of long texts. We have two hypotheses, the first is that subdividing the summarization problem into smaller problems and solving them separately, and later combining these solutions can be beneficial for the task of summarizing long texts. The second hypothesis is that there are other characteristics of the text that can be useful in the creation of the summary. With this in mind, we model the text summarization problem as a binary classification problem. We tested different algorithms and showed that multi-section summarization outperforms single-section summarization with a performance gain of approximately 14% and 5% of BertScore for the Plos One and ArXiv datasets, respectively. We also evaluated the performance of the proposed summarizer using different representations of the text and showed that the single-view representation of attributes is the one that gets the best results. This shows that, for the extractive text summarization task, the attributes selected to compose the attributes view allow to better identify the importance of the sentences. Finally, we compare the proposed method with different state-of-the-art models in extractive, abstractive and hybrid summarization and show that our approach outperforms these models.

Keywords: Extractive Text Summarization, Muti-view Learning, Classification.

Lista de Figuras

Lista de Tabelas

Lista de Quadros

Sumário

Capítulo 1

Introdução

Neste trabalho, propomos uma abordagem alternativa para sumarização de textos científicos longos. Investigamos se a combinação de resumos extraídos de várias seções pode melhorar a qualidade do resumo de todo o documento. Além disso, investigamos o desempenho de sumarizadores utilizando diferentes formas de representação do texto. Isso pode ser de particular interesse para métodos de sumarização que são incapazes de lidar com com textos longos devido a restrições de tamanho, principalmente aqueles com alta demanda de computação.

A sumarização automática de textos é uma das soluções que permite aos usuários identificar as informações mais relevantes de um documento textual, consequentemente, reduzindo o tempo de busca pelas informações [El-Kassas et al., 2021]. O objetivo dessa técnica é condensar as informações de um texto em um resumo simples e descritivo, que dê ao leitor uma ideia geral do texto sem ter que ler todo o seu conteúdo [Nazari e Mahdavi, 2019].

Com a evolução tecnológica e o crescimento da quantidade de textos nas bases de dados digitais, os resumos se tornaram ainda mais importantes. Devido a isso, houve um aumento do interesse da comunidade acadêmica em propor soluções para geração automática de resumos. Embora a tarefa de sumarização seja um processo inerente à mente humana, ela, ainda, é um desafio quando se trata de modelar computacionalmente essa tarefa.

Atualmente, existem diferentes abordagens de sumarização automática de texto, que incluem desde resumos gerados a partir das sentenças mais importantes do texto até modelos que geram resumos em linguagem natural. Cada uma dessas abordagens possui seus desafios. Para os modelos que geram resumos com as sentenças mais importantes do texto, um dos desafios consiste na representação e na definição da importância das sentenças. Por outro lado, quando falamos em modelos que geram resumos em linguagem natural, um desafio é lidar com o tamanho dos textos de entrada. Neste trabalho, exploramos esses dois desafios. Nós temos duas hipóteses. A primeira hipótese é de que segmentar os textos, de acordo com o padrão estrutural do resumo, e sumarizar cada seção, separadamente, pode produzir melhores resultados que a utilização de todo o texto como entrada. A segunda hipótese é de que há outras características dos textos, que

podem ser úteis na criação de um resumo. Tendo isso em vista, colocamos as seguintes questões de pesquisa:

1. Qual a melhor estratégia de treinamento de modelos de sumarização extrativa de textos longos, seção única ou multi-seção?

2. Qual a melhor forma de representar as sentenças de modo que permita identificar a sua importância para o texto?

O objetivo da primeira pergunta é investigar duas estratégias de treinamento de modelos de sumarização extrativa de textos longos. A primeira é a estratégia clássica. O sumarizador recebe como entrada todo o texto e retorna um resumo composto pelo conjunto de sentenças mais importantes do texto. A segunda estratégia, chamada aqui de estratégia multi-seção, segmenta o texto em diferentes seções e cria um sumarizador para cada seção. Após, os resumos são gerados concatenando os resumos de cada seção. A ideia aqui é ter um sumarizador especialista, em cada seção, e dividir o problema de gerar um resumo para um texto longo no problema de sumarizar textos mais curtos.

A segunda pergunta tem como objetivo entender, dentre as soluções de representações das sentenças testadas, qual apresenta o melhor desempenho. Como, neste trabalho, optamos por não usar técnicas de Aprendizado Profundo (AP) para resolver o problema, devido às restrições de tamanho da sequência de entrada dessas técnicas, uma das nossas preocupações é acerca da representação das sentenças do texto. Isto é, qual a melhor forma de representar as sentenças do texto de modo a capturar sua importância? Para responder essa pergunta, testamos diferentes representações do texto e avaliamos o impacto no desempenho dos classificadores.

1.1 Motivação

O avanço das técnicas de AP resultaram em melhorias na qualidade geral dos métodos de sumarização de texto. Os trabalhos mais recentes em sumarização de textos exploram, majoritariamente, modelos baseados em *transformers* em bases de dados de textos curtos. Quando o objetivo é sumarizar textos longos, a utilização dessas técnicas se mostra um desafio, devido ao elevado custo computacional e limites do tamanho da sequência de entrada. Para contornar esses problemas, trabalhos como os de Guo et al. [2021], Zaheer et al. [2020] e Beltagy et al. [2020] propõem adaptações que tornam esses modelos aplicáveis em sequências longas. Em geral, essas adaptações são realizadas no mecanismo de atenção e transformam o mecanismo de atenção global em um mecanismo

de atenção esparso. Outra alternativa usada na sumarização de textos longos é usar uma abordagem de divisão-e-conquista. De acordo com Gidiotis e Tsoumakas [2020] e Souza e Vimieiro [2021], uma abordagem que pode mitigar a perda de informações importantes é a utilização da estratégia de divisão e conquista. Nessa abordagem, os textos são divididos em texto menores e cada um desses textos menores são resumidos separadamente e, posteriormente, esses resumos são combinados.

Uma das aplicações da sumarização de textos longos é a sumarização de textos científicos. Uma pergunta comum neste contexto é: *Por que devemos nos preocupar em resumir esses documentos já que eles já possuem resumos?* Altmami e Menai [2020] destacam três motivos importantes para o estudo da sumarização de artigos científicos. O primeiro é porque o resumo disponibilizado pelos autores não, necessariamente, contém todas as informações importantes do conteúdo do artigo. Isso significa que o mesmo artigo pode ter diferentes bons resumos, abrangendo diferentes aspectos do mesmo. Em segundo lugar, o resumo escrito pelo autor representa seu ponto de vista, destacando, portanto, sua perspectiva sobre o que é importante em seu trabalho. Yang et al. [2016] afirmam que o ponto de vista do autor pode ser parcial e incompleto. Por fim, o terceiro motivo é que não existe um resumo único que atenda a todas as necessidades de todos os leitores, ou seja, não existe um resumo ideal. Cada leitor tem um histórico diferente e pode precisar de informações diferentes [Reeve et al., 2007].

Mesmo com o avanço das técnicas de sumarização e novas propostas para o problema de sumarização de textos longos, a maior parte da literatura se concentra em propor e aprimorar métodos de AP. No entanto, esses métodos possuem algumas limitações, principalmente, quando trata-se da sumarização de textos longos. Nossa hipótese, no entanto, é que há outras formas de explorar a estrutura e a representação dos textos que podem trazer resultados melhores.

1.2 Objetivos

Esta dissertação tem como objetivo propor uma abordagem para sumarização extrativa de textos científicos longos[1].

Destacam-se os seguintes objetivos específicos:

- Realizar uma pesquisa extensiva sobre métodos de sumarização de textos longos a fim de averiguar os diferentes métodos existentes na literatura.

[1]Implementação disponível em: https://github.com/CinthiaS/mv-text-summarizer

- Investigar se a combinação de resumos extraídos de múltiplas seções pode melhorar a qualidade do resumo de todo o documento.
- Explorar diferentes representações das sentenças de entrada na tarefa de sumarização de textos longos.

1.3 Principais Contribuições

As principais contribuições deste trabalho são:

- Proposta de uma abordagem para sumarização de textos científicos longos.
- Condução de experimentos em dois conjuntos de dados do mundo real.
- Disponibilização de código para reprodutibilidade dos experimentos.
- Proposta de uma arquitetura para fusão de representações.

1.4 Origens do Material

Esta dissertação foi elaborada a partir dos materiais apresentados a seguir.

- "A Long Text Summarization Approach to Scientific Articles", publicado em *Symposium in Information and Human Language Technology* (STIL) 2021.
- "A multi-view extractive text summarization approach for long scientific articles", publicado em *International Joint Conference on Neural Networks* (IJCNN) 2022.

1.5 Organização do Trabalho

Esta dissertação foi escrita em seis capítulos, incluindo a Introdução.

- No Capítulo 2, é apresentado o referencial teórico do trabalho. Neste capítulo são descritos o que é sumarização automática de textos e suas principais abordagens e as métricas utilizadas para avaliação da qualidade dos resumos. Além disso, descrevemos o que é aprendizado multi-visão e algumas estratégias para fusão de visões.

- O Capítulo 3 apresenta os trabalhos relacionados. Neste capítulo, apresentamos algumas soluções da literatura juntamente com seus prós e contras e discorremos sobre a necessidade de explorarmos a sumarização automática com enfoque para documentos longos.

- No Capítulo 4, é apresentada a abordagem proposta. Neste capítulo, são descritas todas as etapas para criação do modelo de sumarização de textos longos. Além disso, apresentamos a abordagem de fusão de visões proposta e discorremos sobre todas as etapas necessárias para desenvolver o sumarizador proposto.

- No Capítulo 5, é apresentada uma descrição dos conjuntos de dados utilizados nos experimentos seguido dos resultados experimentais. Ao todo foram realizados dois experimentos que foram subdivididos em experimentos menores. O primeiro experimento tem como objetivo responder a primeira pergunta de pesquisa e o segundo experimento tem como objetivo responder a segunda pergunta de pesquisa. Além disso, foi realizada uma análise comparativa dos resultados utilizando testes estatísticos.

- No Capítulo 6, é apresentada a conclusão do trabalho. Nesse capítulo, resumimos a conclusão do trabalho e apresentamos as respostas para cada pergunta de pesquisa apresentada na Introdução. Além disso, apresentamos algumas ideias para trabalhos futuros.

Capítulo 2

Referencial Teórico

Neste capítulo, definimos os principais conceitos do trabalho. Na Seção 2.1, definimos o conceito de sumarização juntamente com as métricas utilizadas para avaliação da qualidade dos resumos. Na Seção 2.2, apresentamos o conceito de aprendizagem multi-visão.

2.1 Sumarização Automática de Textos

A sumarização automática de textos tem como objetivo identificar as informações mais importantes de um texto e criar um resumo simples e descritivo. As técnicas de sumarização automática de textos facilitam o acesso à informação, permitindo representar os textos de modo compacto, propiciando ao leitor uma visão geral dos textos sem a necessidade de analisar todo seu conteúdo. Além de diminuir o tempo de leitura e de análise dos textos, a sumarização automática de textos facilita o processo de busca de informações e os resumos gerados por essas técnicas são menos tendenciosos em comparação com os resumos gerados por humanos [Mohd et al., 2020].

Os algoritmos de sumarização de textos podem ser divididos em extrativos, abstrativos e híbridos. A sumarização extrativa de textos (ETS[1]) seleciona as principais sentenças/tópicos do texto e, a partir dessas, gera um resumo. Basicamente, para criar resumos extrativos, os algoritmos ranqueiam as sentenças do texto, de acordo com a sua importância, e criam um resumo com as k sentenças mais importantes. Os algoritmos de sumarização abstrativa de textos (ATS[2]), por outro lado, criam uma representação semântica interna do texto e, posteriormente, expressam o conhecimento obtido em linguagem natural. Os algoritmos de sumarização híbrida de textos (HTS[3]) unem os algoritmos de ETS e ATS. A combinação dessas soluções pode variar. Existem arquiteturas que treinam um modelo abstrativo e usam os pesos do mecanismo de atenção da rede para ponderar as sentenças do texto [Trappey et al., 2009]. Há outros que criam um resumo

[1]Sigla em inglês de Extractive Text Summarization
[2]Sigla em inglês de Abstractive Text Summarization
[3]Sigla em inglês de Hybrid Text Summarization

extrativo do texto e, posteriormente, submetem-no a um modelo abstrativo [Gidiotis e Tsoumakas, 2020]. A ideia por trás dessa abordagem é utilizar o sumarizador extrativo para selecionar os principais tópicos do texto e, posteriormente, usar um modelo abstrativo para gerar um texto fluído a partir dos tópicos encontrados. Neste trabalho, nos concentramos na sumarização extrativa de textos.

Cada uma das abordagens de sumarização supracitadas possuem suas vantagens e desvantagens. As técnicas ATS, por exemplo, têm como vantagem a geração de resumos com conteúdo mais flexível, que condensam melhor as informações importantes do texto. No entanto, as técnicas ATS ainda apresentam alguns desafios que tornam as técnicas ETS mais atrativas em alguns cenários, por exemplo, na sumarização jurídica e científica. Uma das vantagens das técnicas ETS é que elas garantem o controle sobre o conteúdo contido nesses resumos [Dong et al., 2021]. Além disso, na sumarização de textos longos, o uso de técnicas ATS pode ser inviável, devido ao alto custo computacional e limitações do comprimento da sequência.

2.1.1 Métricas de Avaliação

Para avaliar o desempenho dos algoritmos de sumarização automática de texto a métrica mais utilizada atualmente é o conjunto de métricas Recall-Oriented Understudy for Gisting Evaluation[4] (ROUGE). As métricas ROUGE possuem uma limitação, elas não consideram a semântica das palavras, considerando apenas o casamento exato entre as palavras do resumo referência e do resumo candidato. Tendo isso em vista, neste trabalho, nós utilizamos, também, a métrica BERTScore. Nas próximas subseções são apresentadas as duas métricas utilizadas.

ROUGE

As métricas ROUGE são amplamente utilizadas na literatura e determinam a semelhança entre um resumo gerado por um modelo computacional e um resumo gerado por humanos. A métrica ROUGE é calculada a partir das métricas precisão, revocação e F1-escore. A métrica revocação, definida na Equação 2.1, indica o quanto do resumo de referência foi capturado pelo resumo candidato. A métrica precisão, definida na Equação 2.2, indica a quantidade de palavras relevantes capturada pelo resumo candidato. A métrica F1-escore, definida na Equação 2.3, combina os valores de precisão e revocação indicando a qualidade geral do modelo.

[4]Disponível em: https://github.com/Diego999/py-rouge/tree/master/rouge

$$Revocação = \frac{Número\ de\ palavras\ sobrepostas}{Total\ de\ palavras\ do\ resumo\ de\ referência}. \tag{2.1}$$

$$Precisão = \frac{Número\ de\ palavras\ sobrepostas}{Total\ de\ palavras\ do\ resumo\ candidato}. \tag{2.2}$$

$$F1 - escore = \frac{2 \times Precisão \times Revocação}{Precisão + Revocação}. \tag{2.3}$$

Neste trabalho, utilizamos as métricas ROUGE-1 (R1), ROUGE-2 (R2) e ROUGE-L (RL). As métricas R1 e R2 pertencem ao conjunto de métricas ROUGE-N. A métrica ROUGE-N avalia a sobreposição de N-gramas de palavras entre os resumos candidatos e os resumos referência [Sanchez-Gomez et al., 2018]. ROUGE-N é calculado de acordo com as Equações 2.1, 2.2 e 2.3, variando o tamanho do n-grama.

Outra métrica desse conjunto é o ROUGE-L, que avalia a sobreposição entre a maior subsequência comum (LCS[5]) compartilhada por duas sentenças [Sanchez-Gomez et al., 2018]. Dado que R é um resumo de referência e C é um resumo candidato. Esta métrica pressupõe que, quanto maior o valor LCS de dois resumos R e C, mais semelhantes eles são. Portanto, ROUGE-L será 1.0 quando as duas sequências forem iguais e 0.0 quando $LCS(R, C)$ for zero, indicando que não há uma sequência comum entre R e C. Para calcular esse valor, são utilizadas as Equações 2.4, 2.5, 2.6 e 2.7, propostas por Lin [2004]. As Equações 2.4, 2.5 e 2.7 representam, respectivamente, as métricas revocação, precisão e F1-escore considerando o LCS entre R e C. Onde m representa o tamanho do resumo R e n representa o tamanho do resumo C. Uma vantagem dessa métrica é que ela não requer um número fixo de n-gramas, incluindo automaticamente a maior sequência de palavras em comum [Lin, 2004].

$$R_{lcs} = \frac{LCS(R, C)}{m}, \tag{2.4}$$

$$P_{lcs} = \frac{LCS(R, C)}{n}, \tag{2.5}$$

$$\beta = \frac{P_{lcs}}{R_{lcs}}, \tag{2.6}$$

$$ROUGE{-}L = F_{lcs} = \frac{(1 + \beta^2) \times R_{lcs} \times P_{lcs}}{R_{lcs} + \beta^2 \times P_{lcs}}. \tag{2.7}$$

BERTScore

BERTScore (BS) é uma métrica recentemente proposta para avaliação da qualidade de modelos de geração de texto. De modo análoga a outras métricas de avaliação de modelos de geração de texto, como ROUGE e BLEU, BS computa a similaridade entre dois

[5]Sigla em inglês para *Longest Common Substring*

textos. No nosso cenário, nós utilizamos a métrica BS para avaliar a similaridade entre um resumo referência e um resumo candidato. Sendo assim, dado um resumo referência $x = \{x_1, ..., x_i\}$, composto por i *tokens* e um resumo candidato $\hat{x} = \{\hat{x}_1, ..., \hat{x}_j\}$, composto por j *tokens*, BS cria uma representação de *embeddings* para cada *token* do texto e computa a similaridade dos *tokens* usando a similaridade do cosseno entre suas representações vetoriais. BS utiliza uma abordagem de casamento guloso, que busca maximizar o escore de similaridade entre dois termos. Dado uma *token* referência x_i e uma *token* candidata $\hat{x}_j$, a similaridade entre esses vetores é o produto interno entre os dois vetores $x_i^T \cdot \hat{x}_j$. Esse cálculo de similaridade considera cada *token* isoladamente. Contudo, os *embeddings* de cada *token* consideram as informações do restante do texto, isto é, a informação contextual de cada *token*. O escore atribuído pela métrica BS é o F1-escore computado utilizando a precisão e a revocação do casamento entre as *tokens*. Assim, dado um resumo referência x e um resumo candidato $\hat{x}$, a revocação, a precisão e o F1-escore entre esses resumos são dados pelas Equações 2.8, 2.9 e 2.10, respectivamente.

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} max\ x_i^T \hat{x}_j \tag{2.8}$$

$$P_{BERT} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} max\ x_i^T \hat{x}_j \tag{2.9}$$

$$F_{BERT} = 2\frac{P_{BERT} \cdot R_{BERT}}{P_{BERT} + R_{BERT}} \tag{2.10}$$

2.2 Aprendizado Multi-Visão

O aprendizado multi-visão tem como objetivo ajustar uma função capaz de modelar cada aspecto (chamado de visão) de um dado para otimizá-lo, melhorando o desempenho de generalização do modelo [Zhao et al., 2017; Gonçalves e Vimieiro, 2021]. Baltrušaitis et al. [2018] divide a aprendizagem multi-visão em cinco desafios: representação, tradução, alinhamento, fusão e co-aprendizagem. O desafio da representação é descobrir como os dados devem ser representados para explorar plenamente sua complementação e redundância. A tradução consiste em mapear dados de uma visão para outra. O desafio do alinhamento é identificar as relações diretas entre essas visões. O desafio da fusão é encontrar a melhor maneira de mesclar as informações das visualizações. Por fim, a co-aprendizagem é o desafio de transferir conhecimento entre visões.

De acordo com Xu et al. [2013], o desempenho bem sucedido de algoritmos de aprendizagem multi-visão pode ser explicado pelos princípios de consenso e complemen-

tariedade. O princípio do consenso visa maximizar a concordância entre diferentes visões, enquanto o princípio complementar afirma que informações de diferentes visões podem se complementar, garantindo uma representação mais abrangente dos dados. Vale ressaltar que, no aprendizado multi-visão, o objetivo não é apenas encontrar diferentes visões dos dados que se complementam, mas também garantir que essas visões representem os dados suficientemente bem [Xu et al., 2013; Li et al., 2021].

Dentre os desafios citados, destaca-se o de fusão das visões. Baltrušaitis et al. [2018] dividiram os métodos de fusão em duas grandes classes, os métodos agnósticos de modelo e os métodos baseados em modelo. Os métodos baseados em modelo são implementados utilizando algoritmos de aprendizagem de máquina, enquanto que os métodos agnósticos de modelo são os que não possuem dependência em relação ao modelo. Os métodos agnósticos de modelo possuem três diferentes alternativas de fusão: fusão precoce, fusão tardia e fusão híbrida. Na fusão precoce, as visões são concatenadas diretamente, gerando uma única visão. Uma das vantagens desse tipo de fusão é que elas requerem o treinamento de um único modelo e são facilmente implementáveis. Contudo, em cenários onde os dados possuem alta dimensionalidade e um pequeno conjunto de treinamento, é comum que essa abordagem sofra *overfitting* [Li et al., 2021]. Na fusão tardia, cada visão é utilizada para treinar um modelo e, posteriormente, esses modelos são combinados. Esse tipo de fusão é caracterizado por possuir uma maior flexibilidade, pois cada visão pode ser treinada com diferentes modelos. Por fim, o modelo de fusão híbrido combina a fusão precoce e a fusão tardia, tentando combinar as vantagens de ambos os métodos.

Capítulo 3

Trabalhos Relacionados

Neste capítulo, são apresentados alguns dos trabalhos recentes da área de sumarização de textos. Esses trabalhos foram escolhidos com o intuito de apresentar uma perspectiva de como a tarefa de sumarização de textos é explorada atualmente.

Como mencionado anteriormente, os algoritmos de sumarização podem ser divididos em três abordagens principais, são elas: ETS, ATS e HTS. Quando falamos em ETS, o principal desafio é definir o escore de cada sentença do texto, isto é, a importância de cada sentença para o texto. Uma solução para esse problema é utilizar as palavras-chave, dando um peso maior para as sentenças que possuem um maior número de palavras-chave [Wang et al., 2011]. Outra solução são os modelos baseados em grafos como, LexRank [Erkan e Radev, 2004] e TextRank [Mihalcea e Tarau, 2004]. Além disso, há trabalhos que modelam o problema de sumarização como um problema de classificação, onde o objetivo é classificar se uma sentença deve ou não pertencer ao resumo. As estratégias que utilizam palavras-chave e os algoritmos LexRank e TextRank são abordagens simples, comumente utilizadas como base de comparação. Atualmente, o estado-da-arte em ETS utiliza arquiteturas baseadas em *transformers* como BERT [Liu e Lapata, 2019], arquiteturas com *Bidirectional Gated Recurrent Units* e *Long Short Term Memory Minus* [Xiao e Carenini, 2019] e soluções sofisticadas baseadas em grafos [Dong et al., 2021].

Quando falamos de ATS, verificamos que, atualmente, essa tarefa é explorada, majoritariamente, utilizando modelos de AP. Esses modelos utilizam, em geral, uma arquitetura *Sequence to Sequence* com *Recurrent Neural Networks* ou, mais recentemente, com *transformers* [Pang et al., 2022]. As soluções baseadas em *transformers* são o estado-da-arte em sumarização de textos, porém, elas possuem algumas limitações. Uma dessas limitações é o tamanho da sequência de entrada, que não pode ultrapassar 512 *tokens*. Quando se trata de sumarização de textos longos essa limitação ocasiona uma perda significativa de informação e, consequentemente, na perda de desempenho do sumarizador. Diante disso e com intuito de mitigar esse problema, trabalhos como os de Phang et al. [2022], Guo et al. [2021], Zaheer et al. [2020] e Beltagy et al. [2020] propõem modelos baseados em *transformers* para sequências longas. Nesses trabalhos, em geral, é realizada uma mudança no mecanismo de atenção. Originalmente, os modelos *transformers* foram propostos utilizando um mecanismo de atenção global com custo quadrático. Nos

trabalhos de Phang et al. [2022], Guo et al. [2021], Zaheer et al. [2020] e Beltagy et al. [2020] são propostos mecanismos de atenção com um custo inferior, tornando possível a sua utilização para sequências longas. Basicamente, essa mudança no mecanismo de atenção tem como objetivo criar um mecanismo de atenção esparso. Embora essas abordagens consigam ter como entrada sequências mais longas que a abordagem tradicional, elas ainda possuem limitações do tamanho da sequência. Outros modelos populares para sumarização de textos são os modelos GPT-3 [Brown et al., 2020] e BLOOM [Scao et al., 2022], pertencentes a classe dos *Large language models*. Por serem modelos de bilhões de parâmetros treinados em grandes bases de dados, em geral, eles apresentam um desempenho superior aos modelos menores presentes na literatura. Contudo, esses modelos também possuem limitações do tamanho da sequência de entrada, mostrando-se inviáveis quando trata-se da sumarização de textos longos. Além disso, um outro desafio é ter infraestrutura para trabalhar com esses modelos.

Outra alternativa aplicada na sumarização de textos longos é usar uma abordagem de divisão-e-conquista. Nessa abordagem, a sumarização de textos é aplicada em trechos/seções dos documentos e o resumo final é a combinação dos resumos de cada seção. Souza e Vimieiro [2021] mostram que a sumarização aplicada em apenas uma seção do texto possui desempenho inferior a aquela que utiliza todas as seções do texto como entrada. Sendo assim, truncar as informações do texto pode gerar resumos de pior qualidade. Gidiotis e Tsoumakas [2020] mostram que quebrar o problema de sumarização em subproblemas menores e, depois, combinar seus resumos pode produzir melhores resultados do que as abordagens tradicionais. Os autores realizaram experimentos com diferentes modelos de sumarização de última geração e mostraram que, usando a abordagem de dividir e conquistar, é possível obter melhores resultados. No contexto da sumarização de textos científicos, uma limitação dessa abordagem é que ela não considera a estrutura dos resumos para realizar a segmentação dos textos.

Há também, na literatura, trabalhos que exploram diferentes formas de representação dos textos para a realização da sumarização. Uma dessas formas de representação é a chamada representação multi-visão (ver Subseção 3.2). Trabalhos como os de Zhang et al. [2016] e Chen e Yang [2020] mostram como o aprendizado multi-visão pode ser utilizado no contexto da sumarização de textos e destacam as vantagens da utilização dessa estratégia. Dentre essas vantagens têm-se uma descrição mais abrangente do dado. Zhang et al. [2016] apresentam em seu trabalho um modelo de sumarização extrativa de multi-documentos que utiliza conceitos de aprendizado multi-visão. De acordo com os autores, a *Convolutional Neural Network* padrão não é capaz de capturar todas as informações contidas em uma sentença, por isso a utilização de uma entrada que considere múltiplas visões/representações de um conjunto de documentos é capaz de proporcionar melhores resultados. O objetivo é que as informações obtidas a partir de múltiplas visões dos documentos possam ser utilizadas para descrever os dados de modo abrangente e ma-

ximizar a concordância entre os diferentes modelos treinados com essas visões. Dentre as representações utilizadas pelos autores tem-se a representação de *embeddings* e atributos de posição da sentença. Chen e Yang [2020] mostram em seu trabalho como o aprendizado multi-visão pode ser aplicado no contexto de sumarização de diálogos. De acordo com os autores, o diálogo pode ter diferentes visões. Cada uma dessas visões representa um aspecto distinto do texto. Assim sendo, segmentar o diálogo em diferentes visões permite que o modelo se concentre em cada aspecto específico do diálogo. Embora esse trabalho tenha um enfoque na sumarização de textos curtos, ele nos traz uma perspectiva diferente de como tratar o problema de sumarização.

Com base nos trabalhos apresentados, verifica-se que há diferentes abordagens para resolver o problema de sumarização. Como mencionado anteriormente, atualmente, as soluções mais recentes utilizam AP. No entanto, essas soluções mostram-se inviáveis no contexto da sumarização de textos longos, devido ao elevado custo computacional dessas soluções e às limitações do tamanho de entrada desses algoritmos. A estratégia de divisão e conquista, por outro lado, mostra-se mais eficiente. Contudo, nessa abordagem há ainda a necessidade de algoritmos para realizar o ranqueamento das sentenças. Sendo assim, outro desafio encontrado é a escolha dos algoritmos para ranqueamento das sentenças e a definição da forma de representação das sentenças. No contexto da sumarização de textos, há diferentes formas para representar as sentenças como *Term Frequency–inverse Document Frequency* (TF-IDF) [Mihalcea e Tarau, 2004; Erkan e Radev, 2004], representação de *embeddings* [Miller, 2019] e também a representação multi-visão [Zhang et al., 2016; Chen e Yang, 2020]. No entanto, ainda não está claro qual a melhor forma de representação das sentenças. Sendo assim, podemos concluir que ainda há espaço para estudos na tarefa de sumarização extrativa de textos.

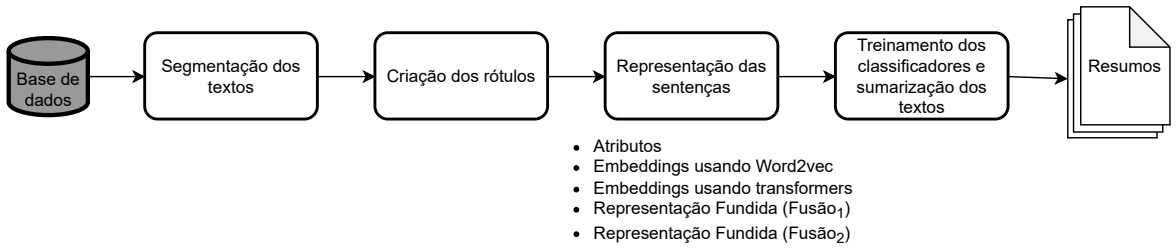
Capítulo 4

Metodologia

Nesta dissertação é proposta uma nova abordagem de sumarização extrativa de textos longos. Nós tratamos o problema de sumarização extrativa de textos como um problema de classificação binária e testamos diferentes algoritmos de classificação nessa tarefa. Nós testamos duas abordagens de treinamento dos classificadores, a primeira utiliza todo o texto como entrada, denominada aqui de abordagem de seção única, e a segunda utiliza divisão e conquista, denominada aqui de abordagem multi-seção.

Basicamente, nessa abordagem, segmentamos os textos de acordo com o padrão estrutural do resumo dos artigos, sumarizamos as seções separadamente, e, em seguida, combinamos seus resumos para obter um único resumo do texto. Após, combinamos os resumos de cada seção em um único resumo. Para representar os textos, criamos uma representação de atributos, denominada aqui de $Visão_1$ e duas representações de *embeddings*, denominada aqui de $Visão_2$. Além disso, combinamos essas duas representações em uma única representação utilizando dois métodos distintos. A ideia é que as visões sejam complementares e, assim, quando fundidas, gerem uma representação mais robusta das sentenças. Para fundir as representações dos textos, também chamadas aqui de visões do texto, nós propomos um método de fusão que é agnóstico ao modelo, isto é, é independente do modelo de classificação a ser utilizado.

Figura 4.1: Diagrama com as etapas da metodologia.



Fonte: Elaborada pela autora.

A abordagem de sumarização extrativa de textos longos proposta compreende quatro etapas principais, conforme representado na Figura 4.1. Na primeira etapa, segmentamos os artigos em diferentes seções. Posteriormente, na segunda etapa, realizamos a rotulação das sentenças do texto. Nessa etapa, o objetivo é atribuir rótulos {0, 1}, onde

0 diz que a sentença não deve pertencer ao resumo e 1 que a sentença deve pertencer ao resumo. Esses rótulos são definidos para cada sentença do texto e, posteriormente, são usados para treinamento dos classificadores. Na terceira etapa, são criadas as diferentes visões das sentenças do texto. Como mencionado anteriormente, criamos duas diferentes visões das sentenças do texto. Além das representações geradas nós criamos mais duas representações que foram geradas fundindo as duas primeiras representações criadas. Ainda na terceira etapa, é descrito o processo de fusão de visões. Nós propomos um método de fusão que utiliza *autoencoders*. Sendo assim, o método proposto é não supervisionado. A entrada para esse método são as duas visões criadas e o objetivo é criar uma representação latente fundida e compacta dessas duas visões. Para isso, o método é treinado para reconstruir as visões de entrada. Além do método proposto nós avaliamos o desempenho da tarefa de sumarização usando um método de fusão precoce. A última etapa compreende as tarefas de classificação e sumarização. Nesta etapa, os algoritmos de classificação são treinados utilizando as diferentes representações criadas e, em seguida, os algoritmos treinados são utilizados para a tarefa de sumarização. Descrevemos cada uma dessas etapas com mais detalhes nas próximas subseções.

4.1 Segmentação dos textos

Como, neste trabalho, utilizamos uma abordagem de sumarização de divisão-e-conquista, a primeira fase da metodologia consiste na segmentação dos textos dos artigos em seções. Neste trabalho, realizamos a segmentação dos textos dos artigos em quatro seções, chamadas aqui de S_1, S_2, S_3 e S_4. A seção S_1 é o resumo do artigo, usado aqui como resumo referência. As seções S_2, S_3, S_4 são as seções de introdução, materiais e métodos e resultados e conclusão, respectivamente. A segmentação dos artigos foi realizada considerando o padrão estrutural dos resumos dos artigos. A ideia é capturar, a partir de cada seção, uma introdução ao problema tratado, uma pequena descrição do método utilizado e as conclusões. Para segmentar os artigos, implementamos um algoritmo de segmentação para documentos XML e também usamos bases já segmentadas.

4.2 Criação dos rótulos

A segunda etapa consiste em rotular as sentenças do texto. Como os conjuntos de dados utilizados são não rotulados, é necessário atribuir um rótulo para cada sentença dos textos. Como tratamos o problema de sumarização como um problema de classificação binária, atribuímos a cada sentença um rótulo 0 ou 1 que indica se uma sentença deve ou não entrar para o resumo. Esta etapa é dividida em três fases. Na primeira fase, os textos são segmentados em sentenças[1]. Em seguida, as sentenças do texto e as sentenças do resumo do artigo são comparadas. Neste trabalho, usamos o resumo do artigo, como resumo de referência. Em seguida, comparamos cada sentença do texto com cada sentença do resumo e computamos sua pontuação ROUGE-1. Assim, por exemplo, se o resumo de referência tiver 5 sentenças e uma seção tiver 40 sentenças, terminaremos com uma lista de 5 pontuações para cada uma das 40 sentenças. Depois disso, usamos uma métrica chamada Máximo. A métrica Máximo define que a pontuação da sentença é a pontuação máxima da sentença relacionada com todas as outras sentenças do resumo de referência. Assim, considerando um resumo referência com 5 sentenças e uma seção com 40 sentenças, a métrica máximo atribuirá para cada uma das 40 sentenças o escore máximo de ROUGE-1 entre ela e as sentenças do resumo candidato. Por fim, elegemos as n sentenças com maior pontuação como relevantes para construção do resumo, isto é, atribuímos a elas o rótulo 1. As restantes são definidas com o rótulo 0.

4.3 Representação das sentenças

Uma representação de documento é denominada aqui de visão. Sempre que usamos uma única representação para ajustar o modelo de classificação, chamamos isso de processo de aprendizado de visão única. Por outro lado, chamamos de aprendizado multi-visão o processo no qual combinamos uma ou mais visões. Consideramos, neste trabalho, diferentes visões para representar sentenças do texto. Essas visões foram criadas para cada seção do documento separadamente. A primeira visão criada contém 13 atributos, incluindo atributos posicionais, atributos baseadas em ocorrências, atributos de centralidade e também valores de classificação gerados pelos algoritmos LexRank, TextRank e *Latent semantic analysis* (LSA). Para extrair os pesos desses algoritmos foi realizada

[1]Segmentador de sentenças disponível em: https://space.io/universe/project/python-sentence-boundary-disambiguation

uma adaptação no código da biblioteca Sumy[2]. A adaptação realizada tem como objetivo extrair não apenas os resumos dos textos mas, também, os pesos de cada sentença. A seguir, são descritos cada um dos atributos selecionados para compor a primeira visão. Esses atributos foram selecionados porque foram usadas anteriormente por Moratanch e Chitrakala [2017], Singh et al. [2016] e Fattah [2014]. A lista de atributos e suas descrições são resumidas na Quadro 4.1.

Como descrito anteriormente, ao todo foram extraídos 13 atributos do texto. O primeiro deles é a posição da sentença no texto. A ideia desse atributo é que sentenças que ocorrem no início e na conclusão são provavelmente importantes, pois a maioria dos documentos são estruturados hierarquicamente com informações importantes no início e no final dos parágrafos [Moratanch e Chitrakala, 2017]. O segundo atributo é uma avaliação das sentenças no texto a partir da quantidade de entidade nomeadas que elas possuem. Para esse atributo, a ideia é que sentenças que se referem a pessoas, objetos, localização geográfica e tempo, por exemplo, podem conter mais informações importantes [Singh et al., 2016]. O terceiro atributo é o Part-of-Speech. De acordo com Singh et al. [2016], algumas classes de palavras desempenham um papel mais importante nas sentenças. De acordo com o mesmo autor, os substantivos e verbos são informações básicas de uma sentença. O substantivo pode desempenhar o papel de sujeito, objeto e complemento em uma sentença e o verbo permite identificar as orações das sentenças. Sendo assim, quanto maior a quantidade de substantivos e verbos em uma sentença, mais informações ela traz. O quarto atributo é referente ao número de palavras-chave no texto. Palavras-chave são essenciais para identificar a importância da sentença. A sentença que apresenta as palavras-chave principais é mais provavelmente incluída no resumo final [Moratanch e Chitrakala, 2017]. Neste trabalho, as palavras-chave são avaliadas com diferentes tamanhos em termos de n-grama de palavras. A ideia é que uma sentença que tenha palavras-chave de 3-gramas seja mais importantes que as que possuem palavras-chave de 2-gramas e 1-grama. No entanto, como as sentenças são, em geral, pequenas, encontrar palavras-chave de 3-gramas pode ser uma tarefa difícil. Sendo assim, adicionamos quatro níveis de palavras-chave, são eles sem limitação de n-grama, 1-grama, 2-gramas e 3-gramas, totalizando quatro atributos. O sétimo atributo é o *Term Frequency - Inverse Sentence Frequency* (TF-ISF). O TF-ISF é uma medida estatística baseada no TF-IDF que calcula a importância de uma palavra para um conjunto de sentenças. Para essa medida, quando uma palavra ocorre com frequência em uma sentença, o valor de TF-ISF aumenta proporcionalmente. No entanto, se essa palavra possui uma alta frequência em todas as sentenças, então, o valor de TF-ISF dessa palavra diminui. O oitavo atributo é o tamanho da sentença. O tamanho da sentença é um atributo importante para o ranqueamento de sentenças. Sentenças curtas, em geral, não capturam todas as informações importantes do texto e sentenças muito longas tendem a conter informações muito específicas e/ou redundantes.

[2]Disponível em: https://pypi.org/project/sumy/ Versão: 0.11.0

O nono atributo é um atributo de centralidade. Esse atributo pondera as sentenças de acordo com a sua representatividade semântica em relação a todas as outras sentenças do texto. Inicialmente, todas as sentenças são representadas como vetores de *embeddings*. Após, todas as sentenças são agrupadas. Posteriormente, é calculada a distância do cosseno entre cada sentença do grupo e seu centroide. A hipótese aqui é de que, quanto mais próxima uma sentença é do centroide, melhor ela representa todas as outras sentenças do grupo. Cada grupo representa um tópico e *outliers* possuem pesos de representatividade 0. Por fim, os últimos atributos são os pesos dados pelos algoritmos LexRank, TexRank e LSA.

Quadro 4.1: Descrição dos atributos extraídos para compor a visão

Atributos	Descrição/Justificativa
Posição da sentença no texto	Sentenças que ocorrem no início e no fim do texto podem ser mais importantes.
Name Entity	Sentenças que se referem a pessoas, objeto, localização geográfica e tempo, por exemplo, podem conter informações mais importantes.
Part-Of-Speech	Avalia a importância da sentença considerando o número de substantivos, verbos e adjetivos.
Palavras-chave (sem limitação de n-grama)	Quanto maior o número de palavras-chave de uma sentença maior sua relevância.
Palavras-chave 1-grama	Sentenças que têm mais palavras-chave de 1-grama ganham maior peso.
Palavras-chave 2-grama	Sentenças que têm mais palavras-chave de 2-grama ganham maior peso.
Palavras-chave 3-gramas	Sentenças que têm mais palavras-chave de 3-gramas ganham maior peso.
TF-ISF	Sentenças com um alto valor de TF-ISF são mais importantes.
Tamanho da sentença	Sentenças que são muito grandes (muitas palavras) ou muito pequenas (poucas palavras) são penalizadas.
Centralidade	São gerados *embeddings* das palavras e esses *embeddings* são agrupados. Sentenças mais próximas do centróide são mais representativas, portanto, são mais importante.
LexRank	Peso atribuído pelo algoritmo LexRank.
TextRank	Peso atribuído pelo algoritmo TextRank.
LSA	Peso atribuído pelo algoritmo LSA.

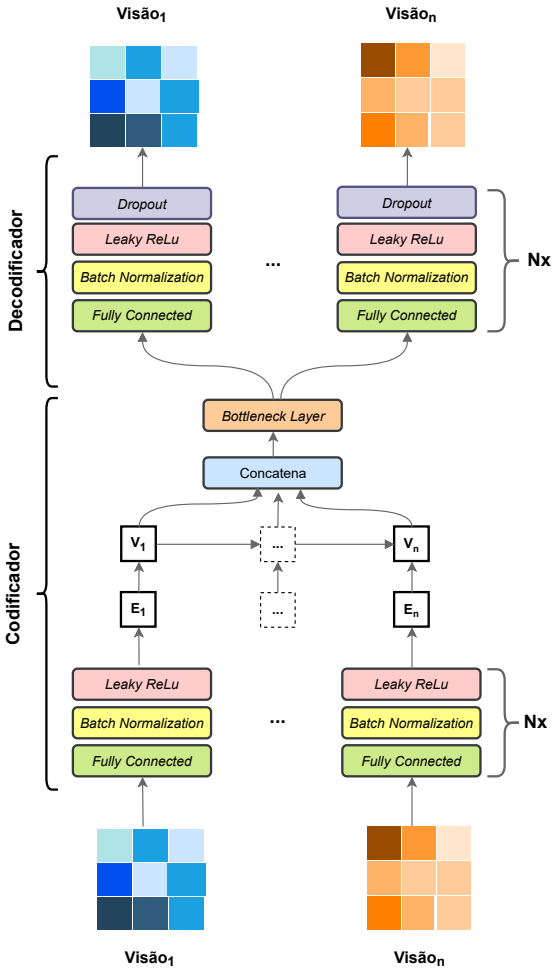
Fonte: Elaborado pela autora.

Nossa segunda visão compreende os *embeddings* que representam cada sentença. Criamos duas visões de *embeddings*, uma utilizando o word2vec[3] e a outra usando arquitetura *transformers*[4]. Para a primeira visão de *embeddings*, foram gerados, para cada sentença, vetores de *embeddings* de 300 dimensões. Para a segunda visão de *embeddings*,

[3]Dispnível em: https://spacy.io/universe/project/spacy-universal-sentence-encoder
[4]Disonível em: https://www.sbert.net/

foram gerados vetores de 738 dimensões. O objetivo é que a visão de *embeddings* seja capaz de capturar mais informações semânticas do texto.

Figura 4.2: Abordagem de fusão de visão proposta. A arquitetura é dividida em codificador e decodificador. A camada do codificador recebe n representações dos dados e mescla a entrada em uma nova representação. A entrada passa por N camadas densas com *Batch Normalization* e *Leaky Rectified Linear Unit* (Leaky ReLu). A saída deste bloco de camada densa são os estados codificados E_1 a E_n. Esses estados são usados para criar as representações V_1 a V_n. A representação fundida é uma concatenação de combinações dos estados ocultos da rede. A decodificação pode ser multi objetivo ou mono objetivo, ou seja, a reconstrução das representações durante o treinamento pode ser feita para uma ou para todas as entradas.



Fonte: Elaborada pela autora.

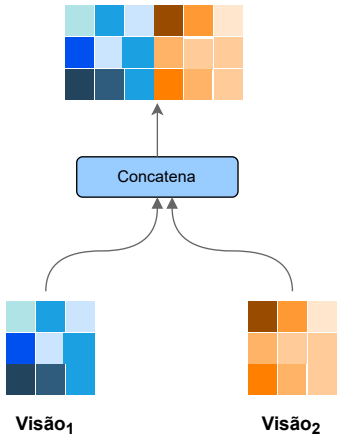
Após gerar as visões únicas das sentenças, uma nova representação é criada fundindo a primeira e a segunda visão. Para fundir as visões, propomos uma arquitetura de fusão, conforme descrito na Figura 4.2. A abordagem de fusão proposta é inspirada nas abordagens de fusão neural propostas por Guo et al. [2017], Chen e Yang [2020] e Silva Romualdo et al. [2021]. Em nosso trabalho, criamos uma arquitetura de fusão que é independente dos modelos de classificação e de sumarização. O objetivo é que a re-

presentação fundida possa ser usada por qualquer classificador. A arquitetura de fusão proposta utiliza *autoencoders*. A arquitetura é dividida em codificador e decodificador. O codificador pode receber n representações como entrada. No nosso caso, n é igual a 2, onde $Visão_1$ representa a matriz de atributos e $Visão_2$, a representação dos *embeddings*. Cada uma dessas representações passa por N camadas densas com *Batch Normalization* e função de ativação Leaky ReLu. A saída deste bloco são os estados codificados E_1 a E_n. Esses estados são usados para criar as representações V_1 a V_n, onde a representação V_1 é uma cópia de E_1 e a representação V_n é a concatenação de V_{n-1} e E_n. Posteriormente, as representações de V_1 a V_n são concatenadas em uma única representação. A representação concatenada passa então por uma Camada de *Bottleneck* cuja saída possui 64 dimensões. A saída da camada de *Bottleneck* é a entrada da arquitetura de decodificação. A arquitetura de decodificação é composta por N blocos de camadas densas com *Batch Normalization*, Leaky ReLu e *Dropout*. Resumindo, nesta arquitetura, o objetivo é usar o codificador para criar uma nova representação a partir das representações $Visão_1$ e $Visão_n$. Observe que as entradas $Visão_1$ e $Visão_n$ podem ser replicadas n vezes, gerando n representações codificadas, que, posteriormente, são mescladas em uma única representação. Além disso, podemos ter n representações diferentes como entrada, que serão mescladas em uma única representação, assim como a decodificação pode ser multi-objetivo ou um único objetivo.

Basicamente, os E_n são novas representações independentes de $Visão_1$ e $Visão_2$, que são as representações de atributos e de *embeddings*. Elas são independentes pois são criadas a partir de $Visão_1$ e $Visão_2$ e não possuem ligações entre si. V_n são também novas representações criadas a partir de E_n, porém, elas possuem ligações entre si. O objetivo dessas ligações é fazer com que V_2, por exemplo, carregue as informações de E_2 mais as informações de V_1, que foi gerada a partir de E_1. Assim, V_2 é construída com base nas representações de V_1 e de E_2.

Além da fusão utilizando a arquitetura apresentada na Figura 4.2, utilizamos também um método de fusão precoce. Nesse método, as visões do texto são fundidas concatenando-as. A Figura 4.3 representa o processo de fusão precoce. Basicamente, a entrada desse método são duas representações, $Visão_1$ e $Visão_2$, e o resultado é uma única visão formada pela concatenação das duas representações.

Figura 4.3: Abordagem de fusão precoce. O método de fusão é uma função de concatenação que recebe como entrada duas representações, $Visão_1$ e $Visão_2$ e retorna uma nova visão resultante da concatenação de $Visão_1$ e $Visão_2$.



Fonte: Elaborada pela autora.

4.4 Treinamento dos classificadores e sumarização dos textos

Esta etapa consiste em três fases: o treinamento do classificador, a geração de predições e a criação de resumos. Selecionamos aleatoriamente 20% dos documentos para compor o conjunto de testes e os 80% restantes para treinar um classificador. Devido à natureza de nossa tarefa, o conjunto de dados é extremamente desbalanceado. Para corrigir esse problema, balanceamos o conjunto de treinamento com subamostragem da classe majoritária.

Para realizar a sumarização, utilizamos uma abordagem de divisão-e-conquista. Dividimos o problema de criar um modelo para um documento no problema de criar três modelos de sumarização, um para cada seção do texto. Cada classificador recebe como entrada uma representação da respectiva seção. O treinamento é realizado usando os rótulos criados na primeira etapa. O objetivo desta fase é criar um classificador especialista para cada seção, que será usado na tarefa de sumarização.

Uma vez que os classificadores são treinados, a tarefa de sumarização é realizada. A entrada desta fase são as previsões do classificador treinado. Nesta fase, usamos exclusivamente o conjunto de teste. Em vez de usar o rótulo binário das sentenças previstas pelo classificador, usamos as probabilidades associadas à previsão do rótulo. O objetivo de usar essa abordagem é maximizar a confiança nas sentenças selecionadas, enquanto fixamos o tamanho do resumo. Portanto, para cada texto, obtemos a probabilidade de suas sentenças pertencerem ou não ao resumo candidato. Em nosso método, é necessário definir o número k de sentenças a serem extraídas.

Capítulo 5

Experimentos

Neste capítulo, são apresentados os resultados dos experimentos conduzidos em dois diferentes conjuntos de dados. Foram conduzidos diferentes experimentos a fim de responder às perguntas de pesquisa apresentadas no capítulo de Introdução e avaliar o desempenho do método proposto.

Inicialmente, na Seção 5.1, são apresentadas as descrições dos conjuntos de dados utilizados. Posteriormente, apresentamos os resultados dos experimentos conduzidos para responder cada pergunta de pesquisa. O primeiro experimento tem como objetivo identificar qual a melhor estratégia para treinamento de modelos de sumarização de textos longos. Para esse experimento, foram testadas duas estratégias, a primeira utiliza todo o texto como entrada. A segunda estratégia utiliza divisão-e-conquista. Ao invés de treinar um único classificador para cada texto nós treinamos três classificadores, sendo um para cada seção do texto e geramos um resumo com a combinação dos resumos de cada seção.

No segundo experimento, o objetivo é avaliar diferentes abordagens de representação de sentenças. Nós testamos cinco diferentes representações, sendo três representações de visão única e duas representações de múltiplas visões. Para simplificar, dividimos esse experimento em três. O primeiro tem como objetivo comparar as representações de visão única, no segundo comparamos as representações de múltiplas visões e no terceiro nós comparamos a melhor abordagem de visão única com a abordagem de múltiplas visões usando nossa abordagem de fusão. Para o primeiro experimento, foram criadas três representações de visão única. No segundo experimento, comparamos o desempenho da sumarização que utiliza como entrada as visões fundidas. Para fundir as representações, utilizamos dois métodos de fusão. O primeiro método é apresentado na Seção 4.3. Esse método utiliza como entrada a matriz de atributos ($Visão_1$) e a matriz de *embeddings* ($Visão_2$) e gera uma nova representação das sentenças do texto. Essa nova representação é chamada aqui de visão fundida, $Fusão_1$. O segundo método de fusão é um método de fusão precoce, denominado aqui como $Fusão_2$. Basicamente, a fusão das diferentes visões é realizada concatenando as visões. Por fim, no terceiro experimento, é realizada uma comparação entre a melhor abordagem de representação de visão única com a nossa abordagem de fusão.

Como mencionado anteriormente, tratamos o problema de sumarização como um

problema de classificação binária. Portanto, selecionamos seis algoritmos de classificação para nossos experimentos: *K Nearest Neighbor* (kNN), *Random Forest* (RF), *Adaptive Boost* (AB), *Cat Boost* (CB), *Gradient Boost* (GB) e uma Rede Neural *Multi Layer Perceptron* (MLP). Ajustamos os hiperparâmetros para todos os classificadores para os dois conjuntos de dados utilizados. A pesquisa de hiperparâmetros é realizada usando o algoritmo *Randomized Search* com *5-fold cross-validation*. Os hiperparâmetros testados e selecionados são apresentados nos Apêndices B e C, respectivamente. Para comparar o desempenho dos algoritmos na tarefa de sumarização, realizamos 30 testes com cada algoritmo para todos os experimentos. Realizamos uma comparação estatística dos resultados obtidos na tarefa de sumarização com cada um dos algoritmos de classificação. Para avaliar se as diferenças entre as médias dos modelos são estatisticamente significativas, os resultados obtidos com a métrica BS foram comparados estatisticamente. Todos os testes foram realizados com nível de significância de 0,05. Usamos o teste não paramétrico de Friedman para determinar se existem diferenças significativas entre os valores médios das populações. Usamos o teste post-hoc Nemenyi para inferir quais diferenças são significativas [Demšar, 2006]. Vale ressaltar que todos os resultados reportados neste capítulo foram obtidos utilizando a base de teste.

5.1 Conjunto de dados

Nós conduzimos os experimentos em dois conjuntos de dados públicos, o primeiro de artigos científicos disponibilizados pela Plos One[1] e coletado em Janeiro de 2022 e o segundo de artigos científicos do ArXiv[2].

A Plos One disponibiliza os artigos em formato XML. Para criar o conjunto de dados utilizado nos experimentos, foi necessário realizar a coleta, refatoração e segmentação dos artigos. A primeira fase da criação do conjunto de dados é a coleta dos dados em formato XML. Em alguns casos, os documentos XML possuem algumas inconsistências, por exemplo, não ter uma tag delimitando o texto das seções. Por consequência disso, antes do processo de segmentação, é necessário realizar uma refatoração dos documentos XML. O processo de refatoração tem como saída um novo arquivo XML corretamente tagueado que, posteriormente, passa pelo processo de segmentação. Após a segmentação dos artigos em seções é realizado um pré-processamento dos textos. Esse pré-processamento inclui a remoção de citações, a remoção dos títulos das seções, a conversão do texto em XML para o formato texto e a remoção de ruídos, sendo considerados ruídos caracteres

[1]Disponível em: http://api.plos.org/text-and-data-mining/ - Acessado em: Jan 31, 2022
[2]Disponível em: https://github.com/armancohan/long-summarization

especiais, como códigos unicode, quebras de linhas e espaços em branco em excesso. O objetivo desse pré-processamento é eliminar informações ruidosas e remover informações irrelevantes do texto, como os títulos das seções. Neste caso, os títulos das seções são considerados irrelevantes pois eles não poderiam ser utilizados como um das sentenças pertencentes ao resumo do artigo por esses serem pouco informativos.

A base do ArXiv utilizada foi coletada em 2018 e disponibilizada em formato JSON por Cohan et al. [2018]. Os atributos e seus respectivos tipos são apresentados no Quadro 5.2. Para segmentar a base nas seções de interesse, implementamos um algoritmo que percorre a lista de seções e atribui o texto de cada seção as seções Introdução, Materiais e Métodos e Resultados e Conclusão. A atribuição é realizada utilizando um dicionário de palavras-chave que associa termos a cada uma das seções de interesse. Originalmente, o conjunto de dados do ArXiv possui 203.037 registros de treino e 6.440 registros de teste. Para realizar os experimentos, utilizamos uma amostra desse conjunto de dados dando preferência para os artigos que possuíam todas as seções de interesse. Artigos sem resumo ou com apenas o resumo foram descartados.

Quadro 5.2: Conteúdo da base de dados do ArXiv.

Chave	Tipo
article_id	str
abstract_text	List[str]
article_text	List[str]
sections_names	List[str]
sections	List[List[str]]

Fonte: Elaborado pela autora.

Para criar o conjunto de dados final, todos os artigos foram segmentados em sentenças. Após a segmentação, as sentenças são rotuladas com 0 ou 1. Sendo que 0 significa que a sentença não deve pertencer ao resumo candidato e 1 que a sentença deve pertencer ao resumo candidato. Ao todo, para cada seção, foram rotuladas com 1 três sentenças. Assim, os classificadores são treinados para selecionar nove sentenças, três de cada seção. Após a rotulagem, os textos foram divididos em base de treino e teste. Devido à natureza da tarefa, as bases de treino e teste são extremamente desbalanceadas. Diante disso, nós realizamos o balanceamento da dobra de treino utilizando subamostragem. Ao todo, foram segmentados 7.368 artigos da base da Plos One sendo 6.392 para a base de treino e 976 para a base de teste. Para a base do ArXiv foram segmentados 40.346 artigos sendo 39.562 para a base de treino e 784 para a base de teste. Em alguns casos, podemos ver que a dobra de teste é maior que a dobra de treino. Isso ocorre pois a dobra de teste não é balanceada diferentemente da dobra de treino. Os resultados do conjunto de dados criado são apresentados na Tabela 5.1.

Tabela 5.1: Quantidade de sentenças na base de treino e teste de cada conjunto de dados

	Plos One		ArXiv	
Seção	Treino	Teste	Treino	Teste
Introdução	36.690	23.864	222.732	35.736
Materiais e Métodos	27.306	57.757	27.750	53.838
Resultados e Conclusão	32.478	208.107	181.536	74.701
Total	96.474	289.728	432.018	164.275

Fonte: Elaborada pela autora.

5.2 Comparação entre resumos gerados com a abordagem de seção única e com a abordagem multi-seção

Nesta seção, são apresentados os resultados dos experimentos conduzidos para responder à seguinte pergunta de pesquisa: *Qual a melhor estratégia de treinamento de modelos de sumarização extrativa de textos longos, seção única ou multi-seção?* Para responder a essa pergunta, realizamos dois experimentos. No primeiro experimento, os algoritmos de classificação são treinados utilizando como entrada todas as sentenças do texto. No segundo experimento, o texto é segmentado em seções e são criados classificadores para cada seção, separadamente, e, posteriormente, o resumo de cada seção é concatenado gerando um único resumo. Para o primeiro experimento, são extraídas nove sentenças para compor o resumo. Para o segundo experimento, são extraídas três sentenças de cada seção totalizando um resumo com nove sentenças. O número de sentenças extraídas foi baseado no tamanho médio dos resumos dos artigos. Para complementar a análise, utilizamos, também, os algoritmos LexRank e LSA para sumarizar os dados usando as duas abordagens. O objetivo da inclusão desses algoritmos é verificar se a diferença entre as abordagens é uma característica do nosso método ou se outros algoritmos apresentam o mesmo comportamento. Além disso, os algoritmos LexRank e TextRank são comumente utilizados como baselines não supervisionados. As Tabelas 5.2 e 5.3 apresentam os resultados obtidos. Em negrito são destacados os melhores resultados.

Tabela 5.2: Resultados obtidos usando as abordagens de seção única e multi-seção para o conjunto de dados da Plos One

Métrica	Abordagem	kNN	RF	AB	GB	CB	MLP	LexRank	LSA
R1 (%)	Seção única	12,99	12,6	12,64	12,76	12,93	12,91	18,3	8,37
	Multi-seção	44,48	42,76	45,89	**46,42**	45,53	44,67	31,30	26,89
R2 (%)	Seção única	2,94	2,9	2,82	2,92	3,02	2,95	3,04	1,74
	Multi-seção	14,21	14,44	15,88	**16,45**	15,82	15,29	6,81	6,84
RL (%)	Seção única	10,93	10,16	10,28	10,38	10,84	10,44	12,88	7,73
	Multi-seção	26,93	26,24	28,26	**28,62**	27,84	27,49	20,53	19,89
BS(%)	Seção única	64,88	52,92	57,63	59,35	58,63	60,14	66,47	61,91
	Multi-seção	79,85	79,39	80,28	**80,49**	80,19	79,98	76,51	75,92

Fonte: Elaborada pela autora.

Tabela 5.3: Resultados obtidos usando as abordagens seção única e multi-seção para o conjunto de dados do ArXiv

Métrica	Abordagem	kNN	RF	AB	GB	CB	MLP	LexRank	LSA
R1 (%)	Seção única	19,67	18,55	19,27	19,04	19,95	18,87	23,99	16,28
	Multi-seção	37,78	34,69	38,29	**38,67**	36,83	37,26	30,7	27,61
R2 (%)	Seção única	4,4	4,28	4,46	4,49	4,45	4,27	4,01	3,89
	Multi-seção	11,43	11,1	12,13	**12,42**	11,62	11,7	6,97	7,58
RL (%)	Seção única	15,34	14,1	14,61	14,45	15,5	14,52	18,07	3,89
	Multi-seção	24,48	22,86	24,73	**25,08**	23,8	24,3	21,24	21
BS(%)	Seção única	71,73	64,48	67,11	67,04	67,25	68,96	73,21	70,08
	Multi-seção	77,48	76,98	77,64	**77,76**	77,31	77,38	76,14	75,15

Fonte: Elaborada pela autora.

Como podemos ver, a abordagem multi-seção apresentou um desempenho superior ao da abordagem de seção única para ambos os conjuntos de dados. Houve um ganho de desempenho de, aproximadamente, 14% de BS para a base da Plos One e um ganho de, aproximadamente, 5% de BS para a base do ArXiv. Esses resultados foram obtidos subtraindo o resultado do sumarizador com melhor desempenho para a abordagem de multi-seção pelo sumarizador com melhor desempenho de BS para a abordagem de seção única. Os melhores resultados para o conjunto de dados da Plos One e do ArXiv foram obtidos com a abordagem multi-seção usando os algoritmos GB. Vale ressaltar que, na análise entre as estratégias, os resultados das métricas ROUGE também apresentaram um ganho de desempenho, mostrando assim que houve um aumento de n-gramas de palavras compartilhados entre os resumos de referências e candidatos com cada abordagem usando a estratégia de sumarização multi-seção. Além disso, constatamos que os sumarizadores LexRank e LSA também apresentam um melhor desempenho utilizando a abordagem multi-seção. Isso mostra que a abordagem multi-seção pode proporcionar ganhos de performance para outros algoritmos além da abordagem de sumarização aqui proposta. A fim de exemplificar a diferença da qualidade dos resumos obtidos com a abordagem de seção única e multi-seção é apresentado no Quadro 5.3 um exemplo de resumo gerado

pelo algoritmo GB usando as duas abordagens. Esse exemplo foi escolhido aleatoriamente dentre os resumos possíveis.

Quadro 5.3: Exemplo de resumo gerado usando a abordagem de seção única e a abordagem multi-seção para um artigo da Plos One

Referência	studies aimed at identifying body mass index (bmi) cutoffs representing increased diseased risk for asians are typically based on cross sectional studies. this study determines an optimal bmi cutoff for overweight that represents elevated incidence of hypertension in chinese adults with data from the china health and nutrition survey 20002004 prospective cohort. cumulative incidence was calculated by dividing new cases of hypertension over the study period by the total at risk population, aged 1865 years, in 2000. sex specific receiver operating characteristic (roc) curves were used to assess the sensitivity and specificity of bmi as a predictor of hypertension incidence. four year cumulative incidences of hypertension (13% and 19% for women and men, respectively) were significantly (p ¡ 0.005) related to the increase in bmi. the crude area under the curves (auc) were 0.62 (95% ci: 0.590.65) and 0.62 (95% ci: 0.580.65) for men and women, respectively; the age adjusted auc were 0.68 (95% ci: 0.650.70) and 0.71 (95% ci: 0.680.74) for men and women, respectively. a bmi of 23.5 kg/m2 for women and 22.5 kg/m2 for men provided highest sensitivity and specificity (60%). the finding was consistent in different age groups. a bmi level of 25 kg/m2 provided lower sensitivities (36% for women and 29% for men) with higher specificities (80% for women and 85% for men). our study supported the hypothesis that the bmi cutoff to define overweight should be lower in chinese than in western populations.
Seção única	Thus the inclusion of an older participant would bias the association between BMI and health outcome toward the null and lead to a higher BMI cutoff . In contrast the inclusion of persons with a lower risk of hypertension in the longitudinal sample would bias the estimate away from the null and increase the AUC values . Similar to other cardiovascular risk factors blood pressure might vary over time 18 and thus a hypertensive patient in one survey could become normotensive in the next survey . 295 . Effect of body mass index on all cause mortality and incidence of cardiovascular diseases report for meta analysis of prospective studies open optimal cut off points of body mass index in Chinese adults . a . 45 1 8 . ..
Multi-seção	Increased prevalence attributable death and economic burdens of overweight and non communicable diseases are emerging problems in China and other Asian countries 1 4 . Because these studies were based on cross sectional samples we are not certain that the exposure to a higher BMI had preceded the hypertension outcome 15 . We used an ROC curve analysis to determine an optimal BMI cutoff for overweight that represents elevated incidence of hypertension in Chinese adultsFor this analysis we used data from the CHNS conducted in 2000 and 2004 because these two surveys had the most comparable study sample questionnaires and protocol and equipment in measuring blood pressure weight height and waist circumference . Of the 5543 participants 4492 81 with normal blood pressure in 2000 were included in our longitudinal sample . Three measurements of systolic or diastolic blood pressure were averaged to reduce the effect of measurement errors .Based on these criteria the most reduced model had age as an effect measure modifier the association between BMI and hypertension was stronger among the younger participants and sex and drinking status as confounding factors . Area under the receiver operating characteristic curves AUC optimal body mass index BMI cutoff values sensitivities and specificities stratified by sex and age at baseline for the prediction of hypertension incidence . Effect of body mass index on all cause mortality and incidence of cardiovascular diseases report for meta analysis of prospective studies open optimal cut off points of body mass index in Chinese adults .

Fonte: Elaborado pela autora.

Tabela 5.4: Métricas para o exemplo da Tabela 5.5

Abordagem	R1 (%)	R2 (%)	RL (%)	BS (%)
Seção única	37,65	5,90	20,01	74,49
Multi-seção	45,51	15,23	28,70	77,45

Fonte: Elaborada pela autora.

Como podemos ver na Quadro 5.3 e na Tabela 5.4, o resumo criado pela abordagem multi-seção conseguiu capturar mais informações relevantes do que o resumo de

seção única. Esse é um comportamento bastante comum nos resumos de seção única gerados. Com base no exemplo e nos resultados apresentados nas Tabelas 5.2 e 5.3, podemos concluir que a abordagem de sumarização multi-seção é capaz de gerar resumos de maior qualidade que a abordagem de seção única. Vale ressaltar que, em geral, os resumos gerados por ambas as abordagens apresentaram problemas relacionados à coesão entre as sentenças do texto. No exemplo de multi-seção, por exemplo, podemos ver que as sentenças do resumo gerado não possuem ligações entre si. Esse é um problema característico de sumarizadores extrativos.

5.3 Comparação entre as diferentes formas de representação das sentenças

Nesta seção, são apresentados os resultados dos experimentos conduzidos para responder à seguinte pergunta de pesquisa: *Qual a melhor forma de representar as sentenças de modo que permita identificar a sua importância para o texto?* Para responder essa pergunta, nós avaliamos cinco diferentes representações de sentenças, sendo uma visão de atributos, duas visões de *embeddings* e duas visões fundidas, e comparamos os seus resultados na tarefa de sumarização extrativa de textos longos. Tendo em vista a conclusão do primeiro experimento, de que a abordagem multi-seção tem um desempenho superior ao da abordagem de seção única, os resultados dos experimentos apresentados, a seguir, utilizam apenas a melhor estratégia de treinamento, que é a estratégia multi-seção. Os experimentos realizados foram conduzidos em dois conjuntos de dados. As Tabelas 5.5 e 5.6 apresentam os resultados obtidos com cada uma das representações das sentenças utilizando os conjuntos de dados da Plos One e do ArXiv, respectivamente. Em negrito são destacados os melhores resultados com cada uma das métricas.

Tabela 5.5: Resultados obtidos utilizando os classificadores testados com cada uma das cinco representações para o conjunto de dados da Plos One

Métrica	Representação	kNN	RF	AB	GB	CB	MLP
R1 (%)	Atributos	44,48	42,76	45,89	**46,42**	45,53	44,67
	Word2vec	37,91	35,24	33,69	37,37	33,85	35
	Transformers	37,59	36,98	37,67	37,19	37,25	35,81
	Fusão 1	42,74	42,55	43,43	43,09	43,83	44,04
	Fusão 2	43,45	42,64	45,8	44,84	43,44	44,75
R2 (%)	Atributos	14,21	14,44	15,88	**16,45**	15,82	15,29
	Word2vec	9,75	9,12	8,9	10,04	8,7	9,01
	Transformers	9,57	9,62	9,82	9,84	9,79	9,24
	Fusão 1	13,15	13,92	14,31	14,04	14,47	14,68
	Fusão 2	13,33	13,78	15,66	15,08	14,69	15,31
RL (%)	Atributos	26,93	26,24	28,26	**28,62**	27,84	27,49
	Word2vec	23,51	22,76	22,4	23,65	22,19	22,62
	Transformers	23,33	23,26	23,47	23,37	23,49	23,40
	Fusão 1	26,1	25,89	26,38	26,24	26,8	26,95
	Fusão 2	26,31	26,01	28,1	27,44	26,74	27,51
BS(%)	Atributos	79,85	79,39	80,28	**80,49**	80,19	79,98
	Word2vec	78,04	77,34	77,39	78,00	77,16	77,31
	Transformers	77,89	76,02	77,93	77,92	77,90	77,54
	Fusão 1	79,27	79,36	79,52	79,48	79,67	79,74
	Fusão 2	79,47	79,2	80,26	79,94	79,59	79,98

Fonte: Elaborada pela autora.

Tabela 5.6: Resultados obtidos utilizando os classificadores testados com cada uma das cinco representações para o conjunto de dados do ArXiv

Métrica	Representação	kNN	RF	AB	GB	CB	MLP
R1 (%)	Atributos	37,78	34,69	38,29	**38,67**	36,83	37,26
	Word2vec	35,40	32,59	32,30	34,06	30,77	31,16
	Transformers	35,86	34,17	35,81	35,10	34,59	37,47
	Fusão 1	36,30	34,55	35,31	35,99	35,11	35,52
	Fusão 2	36,88	31,86	36,90	37,18	34,19	35,68
R2 (%)	Atributos	11,43	11,10	12,13	12,42	11,62	11,70
	Word2vec	9,33	7,87	8,21	8,48	7,24	7,64
	Transformers	9,72	8,63	9,53	9,00	9,09	11,10
	Fusão 1	11,45	11,50	11,78	12,07	11,62	11,74
	Fusão 2	12,58	11,61	13,16	**13,47**	12,34	12,40
RL (%)	Atributos	24,48	22,86	24,73	**25,08**	23,80	24,30
	Word2vec	22,85	21,43	22,00	22,55	20,99	21,00
	Transformers	23,06	22,40	23,16	22,78	22,39	23,92
	Fusão 1	23,52	22,83	23,21	23,57	23,12	23,35
	Fusão 2	23,88	21,66	24,02	24,25	22,69	23,46
BS(%)	Atributos	77,48	76,98	77,64	77,76	77,31	77,38
	Word2vec	76,87	75,68	76,06	76,44	75,45	75,36
	Transformers	77,01	76,44	76,96	76,77	76,56	77,61
	Fusão 1	77,5	77,19	77,4	77,59	77,27	77,42
	Fusão 2	78	77,3	78,12	**78,2**	77,7	77,83

Fonte: Elaborada pela autora.

Com base nos dados apresentados nas Tabelas 5.5 e 5.6, verifica-se que, para conjunto de dados da Plos One, a representação de atributos apresentou, para todas as métricas avaliadas, o melhor resultado. É importante notar também que as representações fundidas obtiveram, em todos os casos, resultados superiores aos das representações de *embeddings*. Além disso, podemos constatar que o desempenho dos sumarizadores usando as representações de *embeddings* foi, em geral, similar. Isso mostra que embora os *embeddings* gerados com *transformers* sejam considerados de melhor qualidade, na tarefa de sumarização, eles apresentam um resultado similar aos *embeddings* gerados usando word2vec. Quando comparamos o desempenho dos algoritmos testados, verificamos que, em geral, o algoritmo GB se destaca dos demais.

Para o conjunto de dados do ArXiv, os melhores resultados de R1 e RL foram obtidos utilizando a representação de atributos, enquanto que os melhores resultados de R2 e BS foram obtidos utilizando a representação fundida usando o método de fusão que concatena a visão de atributos com a visão de *embeddings*. Embora a visão fundida tenha obtido melhores resultados, verifica-se que a diferença entre os resultados utilizando a visão de atributos e os obtidos com a visão fundida apresentam um diferença inferior a 2%. Novamente, quando comparamos o desempenho das visões fundidas com o desempenho das visões de *embeddings*, concluímos que as visões fundidas apresentam o melhor

desempenho. Já quando comparamos a diferença de desempenho entre as duas representações de *embeddings* podemos ver que a diferença do desempenho dos algoritmos com cada uma das representações é pequena, isto é, inferior a 1%. Quando comparamos o desempenho dos algoritmos, novamente, verificamos que, em geral, o algoritmos GB se destaca dos demais.

Com intuito de identificar se as diferenças entre as médias dos resultados são estatisticamente significantes, conduzimos três experimentos comparativos. O primeiro experimento tem como objetivo avaliar, dentre as representações de visão única testadas, qual obtém o melhor resultado. Nesse experimento, comparamos apenas a visão de atributos com a visão criada usando word2vec. Essa decisão foi tomada devido a representação de *transformers* ter um arquitetura mais complexa e apresentar um resultado semelhante ao do word2vec. No segundo experimento, o objetivo é identificar qual dos métodos de fusão apresenta os melhores resultados. Por fim, o terceiro experimento avalia o desempenho entre a melhor visão única com a visão fundida com o nosso método. Para realizar as análises estatísticas, nós realizamos 30 execuções de cada um dos algoritmos com cada uma das representações e comparamos se os resultados obtidos com a métrica BS apresentam uma diferença estatisticamente significante. Para isso, primeiro realizamos o teste de Friedman, a fim de verificar a hipótese nula de que não há diferenças estatisticamente significantes nas performance dos métodos. Com base nos resultados apresentados na Tabela 5.7, verifica-se que, para ambos os conjunto de dados, a hipótese nula foi rejeitada. Posteriormente, nos casos onde a hipótese nula é rejeitada no teste de Friedman, aplica-se o pós-teste de Nemenyi, que serve para identificar o quão diferente, por meio da diferença crítica (CD), um algoritmo é do outro. Basicamente, a CD serve como um limiar para comparação par-a-par entre os algoritmos, agrupando os que estão dentro da faixa de CD e separando os que estão fora dessa faixa. No gráfico de diferença crítica, o eixo x representa o valor médio dos ranques dos algoritmos e a linha horizontal em negrito agrupa os algoritmos que não possuem um diferença estatisticamente significante entre eles. No topo do gráfico é apresentado o valor de CD que representa a diferença necessária para que duas comparações sejam considerada significativamente diferentes. A Tabela 5.8 apresenta os p-valores do teste post-hoc de Nemenyi para as comparações realizadas.

Tabela 5.7: P-valores do teste Friedman para as comparações realizadas

	Plos One	ArXiv
	p-valor	
$Visão_1$ x $Visão_2$	1.43e-8	1.87e-10
$Fusão_1$ x $Fusão_2$	4.33e-7	1.31e-08
$Visão_1$ x $Fusão_1$	5.65e-13	2.22e-09

Fonte: Elaborada pela autora.

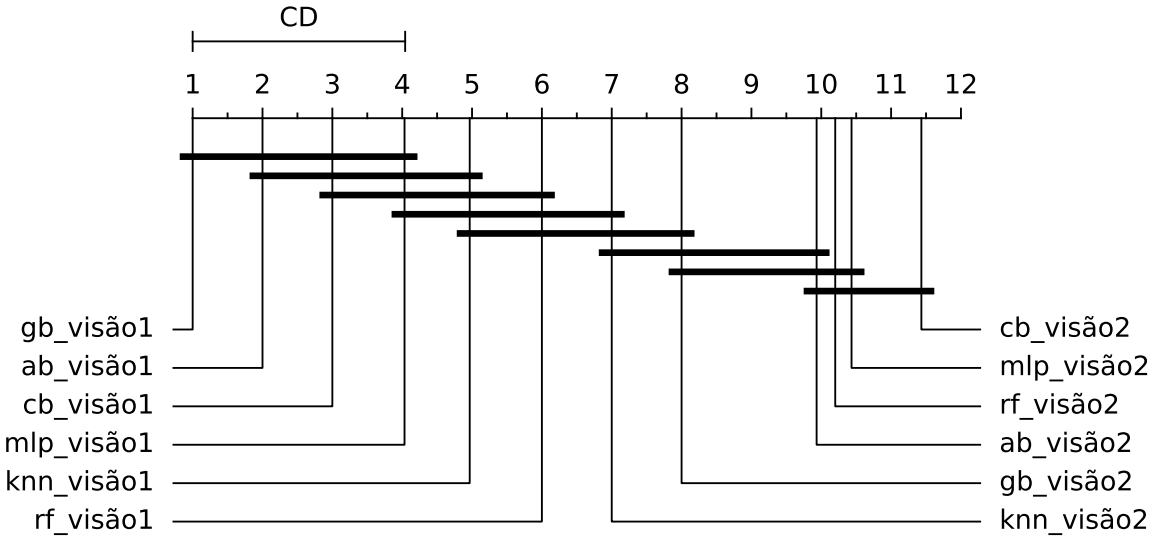
Tabela 5.8: P-valores do teste post-hoc de Nemenyi para as comparações realizadas

	Plos One	Arxiv
	p-valor	
$Visão_1$ x $Visão_2$	2.84e-62	2.57e-63
$Fusão_1$ x $Fusão_2$	4,35e-63	2.76e-16
$Visão_1$ x $Fusão_1$	1.90e-63	4.79e-46

Fonte: Elaborada pela autora.

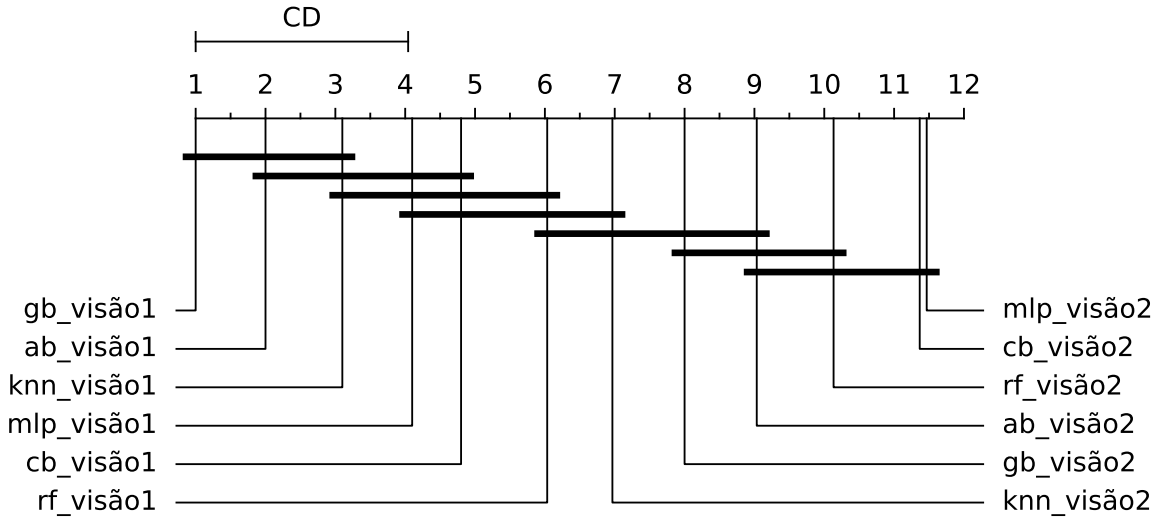
As Figuras 5.1 e 5.2 mostram os resultados obtidos com o primeiro experimento, que compara a visão de atributos ($Visão_1$) e a visão de *embeddings* ($Visão_2$) para a base da Plos One e do ArXiv, respectivamente.

Figura 5.1: Diagrama de diferença crítica para o conjunto de dados da Plos One. Comparação entre a média dos algoritmos utilizando como entrada a visão de atributos ($Visão_1$) e a visão de *embeddings* ($Visão_2$). Quanto menor a posição no rank melhor o desempenho do algoritmo. Há uma linha cruzando um ou mais algoritmos quando não há uma diferença estatisticamente significante entre eles.



Fonte: Elaborada pela autora.

Figura 5.2: Diagrama de diferença crítica para o conjunto de dados do ArXiv. Comparação entre a média dos algoritmos utilizando como entrada a visão de atributos ($Visão_1$) e a visão de *embeddings* ($Visão_2$). Quanto menor a posição no rank melhor o desempenho do algoritmo. Há uma linha cruzando um ou mais algoritmos quando não há uma diferença estatisticamente significante entre eles.

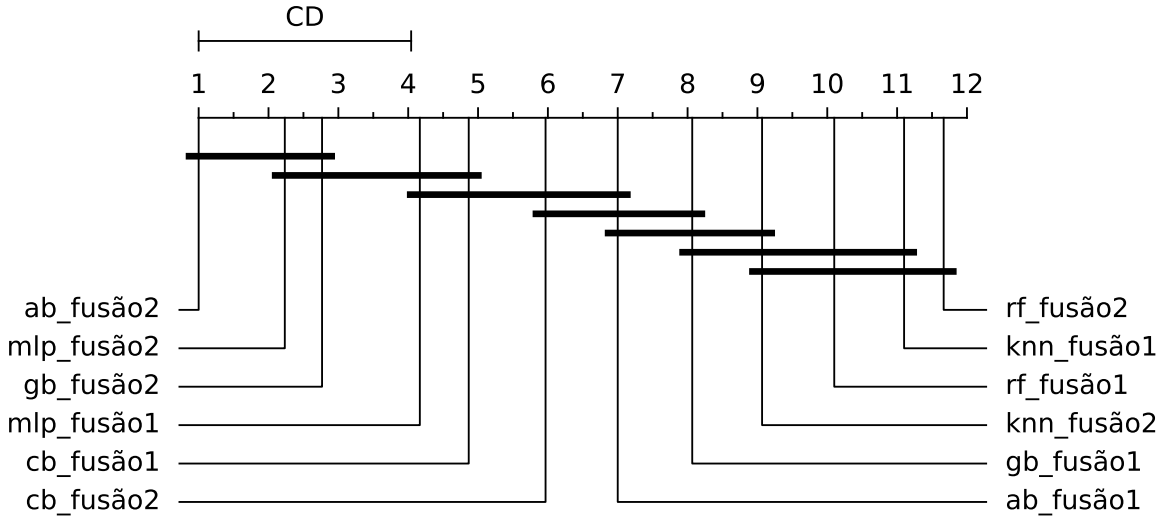


Fonte: Elaborada pela autora.

Com base no resultado da análise estatística apresentada nas Figuras 5.1 e 5.2, podemos concluir que o melhor resultado foi obtido com o algoritmo GB usando a visão de atributos. Analisando o diagrama para o conjunto de dados da Plos One, podemos ver que os algoritmos GB, AB, CB e a rede MLP, usando $Visão_1$ são agrupados, isso significa que eles estão na mesma faixa de CD. Com isso, podemos concluir que não existe uma diferença estatisticamente significante entre as médias desses algoritmos. Além disso, pelo diagrama, podemos ver que os algoritmos pior ranqueados são os que utilizando a $Visão_2$. Olhando para os resultados com a base do ArXiv, vemos que os algoritmos melhor ranqueados são o GB, AB e kNN usando, também, a $Visão_1$. Esses algoritmos, também, são agrupados no teste *post hoc* de Nemenyi. Isso significa que não há uma diferença estatisticamente significante entre eles. Novamente, podemos verificar que os algoritmos pior ranqueados são os que utilizam a $Visão_2$. Com base nesses dados, podemos concluir que, para ambos os conjuntos de dados a $Visão_1$ é a que apresenta o melhor desempenho. Se avaliarmos os resultados das Tabelas 5.5 e 5.6, podemos constatar que, em todos os casos e para ambos os conjuntos de dados, a representação de atributos obteve resultados melhores que a representação de *embeddings*. Isso mostra que, para a tarefa de sumarização extrativa de textos, os atributos selecionadas para compor a $Visão_1$ permitem identificar a importância das sentenças melhor que os *embeddings* ($Visão_2$). Isso é válido tanto para os *embeddings* criados com word2vec quanto para os criados usando *transformers*.

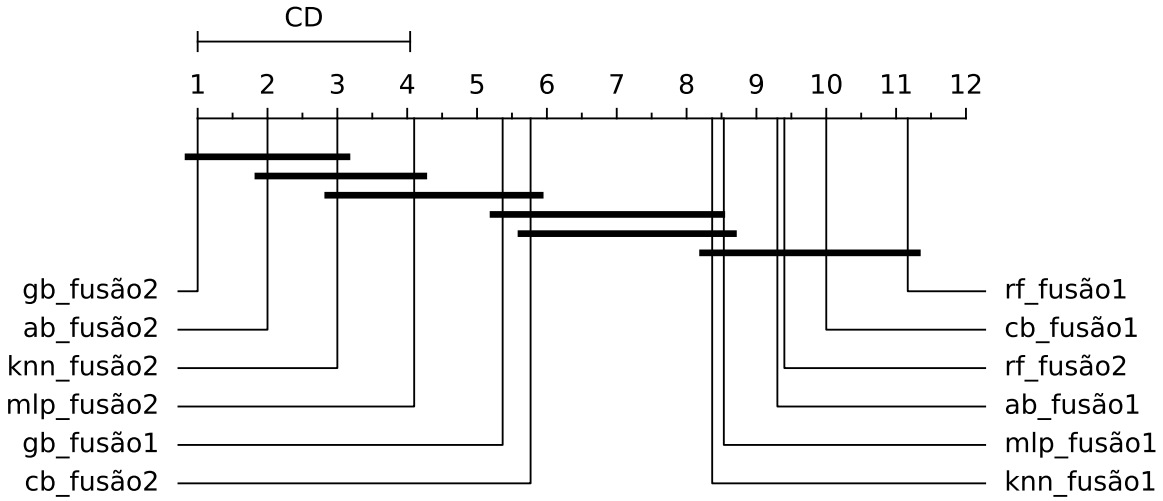
No segundo experimento, comparamos as abordagens de fusão. Ao analisar os resultados das Tabelas 5.5 e 5.6, verificamos que o método de fusão precoce, $Fusão_2$, apresentou o melhor desempenho para ambos os conjuntos de dados. Contudo, quando avaliamos o resultado da métrica BS, que considera mais que o casamento exato entre as palavras dos resumos referência e candidato, verificamos que ambos os métodos têm um desempenho semelhante. Para identificar se a diferença entre as médias é estatisticamente significante realizamos, novamente, o teste de post-hoc Nemenyi para inferir quais diferenças são significativas. As Figuras 5.3 e 5.4 mostram os resultados da comparação estatística da métrica BS dos algoritmos treinados com as representações fundidas por $Fusão_1$ e $Fusão_2$ para a base da Plos One e do ArXiv, respectivamente.

Figura 5.3: Diagrama de diferença crítica para o conjunto de dados da Plos One. Comparação entre a média dos algoritmos utilizando como entrada a visão fundida com o nosso método de fusão ($Fusão_1$), e com o método de fusão precoce ($Fusão_2$). Quanto menor a posição no rank melhor o desempenho do algoritmo. Há uma linha cruzando um ou mais algoritmos quando não há uma diferença estatisticamente significante entre eles.



Fonte: Elaborada pela autora.

Figura 5.4: Diagrama de diferença crítica para o conjunto de dados do ArXiv. Comparação entre a média dos algoritmos utilizando como entrada a visão fundida com o nosso método de fusão ($Fusão_1$), e com o método de fusão precoce ($Fusão_2$). Quanto menor a posição no rank melhor o desempenho do algoritmo. Há uma linha cruzando um ou mais algoritmos quando não há uma diferença estatisticamente significante entre eles.



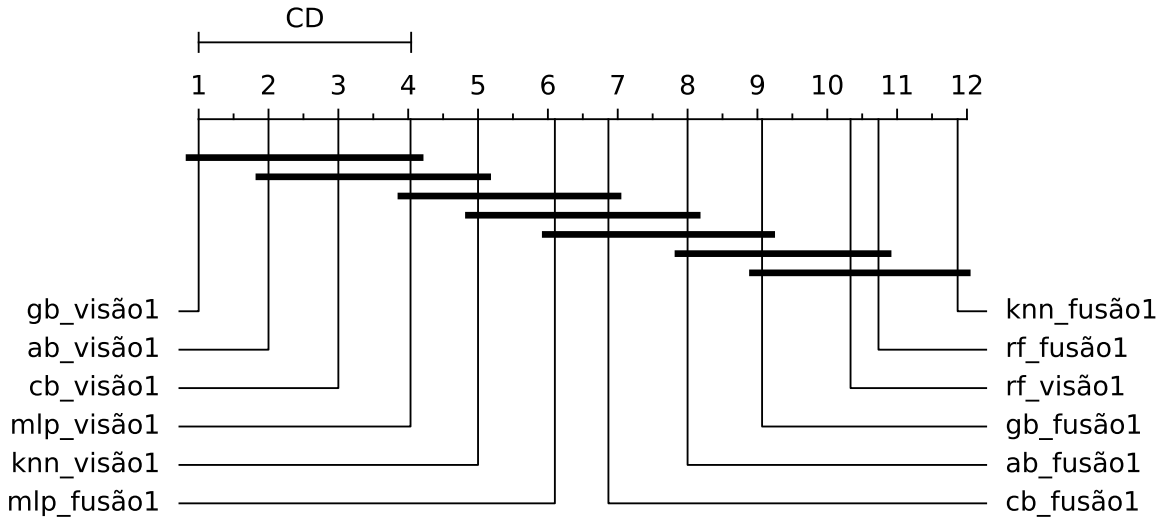
Fonte: Elaborada pela autora.

Analisando o diagrama das Figuras 5.3 e 5.4, concluímos que os algoritmos melhor ranqueados são os que utilizam a visão fundida pelo método de fusão precoce ($Fusão_2$). Podemos ver no diagrama que, para o conjunto de dados da Plos One, os algoritmos melhor ranqueados são AB, MLP e GB usando $Fusão_2$. Como podemos ver, esses algoritmos foram agrupados pelo teste *post hoc* Nemenyi, isso significa que não existe uma diferença estatisticamente significante entre eles. O mesmo ocorre para o conjunto de dados do ArXiv, porém agrupando os algoritmos GB, AB e o kNN usando $Fusão_2$. Acredita-se que o desempenho superior dos algoritmos usando $Fusão_2$ tenha sido ocasionado pelo fato do algoritmo ter atribuído uma maior importância para os atributos de $Visão_1$. Sendo assim, o desempenho de $Fusão_2$ seria mais semelhante ao desempenho de $Visão_1$, como de fato ocorreu. Vale ressaltar que, nessa análise, consideramos apenas o desempenho individual de cada classificador, não analisando outros fatores, como a consistência dos classificadores treinados com cada uma das representações.

Por fim, no terceiro experimento, comparamos os modelos que utilizam a representação de atributos com o modelo que utiliza a representação fundida com a nossa abordagem de fusão. Com base nos resultados apresentados nas Tabelas 5.5 e 5.6, verificamos que o melhor desempenho tanto em termos de ROUGE quanto em termos de BS foram dos algoritmos que utilizaram a visão de atributos ($Visão_1$). Contudo, a diferença entre a métrica BS em cada abordagem é pequena. A fim de verificarmos se a diferença entre a média dos algoritmos é estatisticamente significativa, realizamos, novamente, o

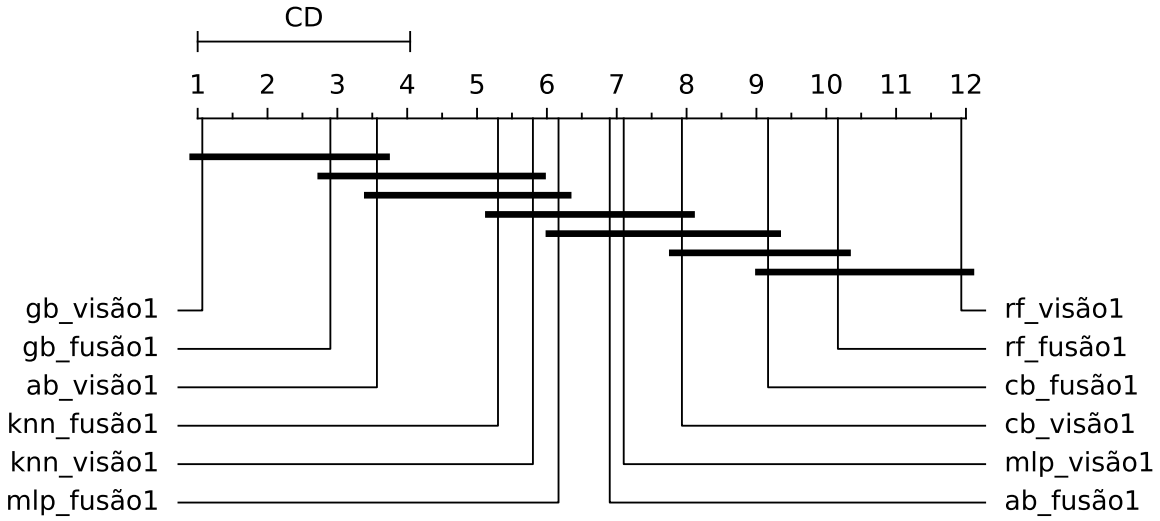
teste de post-hoc Nemenyi para inferir quais diferenças são significativas. As Figuras 5.5 e 5.6 apresentam o resultado obtido.

Figura 5.5: Diagrama de diferença crítica para o conjunto de dados da Plos One. Comparação entre a média dos algoritmos utilizando como entrada a visão de atributos e a média dos algoritmos utilizando a visão fundida ($Fusão_1$). Quanto menor a posição no rank melhor o desempenho do algoritmo. Há uma linha cruzando um ou mais algoritmos quando não há uma diferença estatisticamente significante entre eles.



Fonte: Elaborada pela autora.

Figura 5.6: Diagrama de diferença crítica para o conjunto de dados do ArXiv. Comparação entre a média dos algoritmos utilizando como entrada a visão de atributos e a média dos algoritmos utilizando a visão fundida ($Fusão_1$). Quanto menor a posição no rank melhor o desempenho do algoritmo. Há uma linha cruzando um ou mais algoritmos quando não há uma diferença estatisticamente significante entre eles.



Fonte: Elaborada pela autora.

Para o conjunto de dados da Plos One, podemos ver que os algoritmo melhor ranqueado foi o GB usando $Visão_1$. Entretanto, o teste de *post hoc* de Nemenyi agrupa os algoritmos GB, AB e CB usando $Visão_1$, isso significa que não há uma diferença estatisticamente significante entre eles. O algoritmo com pior desempenho nesse caso foi o kNN usando $Fusão_1$. Para o conjunto de dados do ArXiv, o algoritmo melhor ranqueado foi o GB com $Visão_1$. Vale notar que o segundo algoritmo melhor ranqueado é o GB com $Fusão_1$ e, de acordo com o teste, não há uma diferença estatisticamente significante entre GB com $Visão_1$ e GB com $Fusão_1$. Para esse conjunto de dados, o algoritmo pior ranqueado foi o RF com $Visão_1$.

5.3.1 *Baselines* e modelos do estado-da-arte

Nesta seção, comparamos o desempenho dos modelos obtidos com *baselines* e melhores modelos do estado-da-arte em sumarização. Criamos dois experimentos. O primeiro experimento visa gerar *baselines* de comparação. No segundo experimento, os modelos propostos por Beltagy et al. [2020], Zaheer et al. [2020], Zhang et al. [2020], Lewis et al. [2019] e Raffel et al. [2020] foram executados usando nosso conjunto de dados de teste.

Usamos três estratégias de sumarização extrativa como *baseline*. A primeira estratégia é chamada Max. Essa estratégia cria o resumo selecionando as três sentenças com a pontuação ROUGE-1 mais alta de cada seção. Este resumo é o nosso resumo padrão ouro. Então, supondo que nosso classificador acerte todas as classificações, esse seria o melhor resultado que podemos obter com o conjunto de rótulos criado. A segunda estratégia é Min, que cria o resumo com as sentenças com a menor pontuação R1 de cada seção. Este resumo é o pior que poderíamos obter usando nosso conjunto de rótulos. A terceira estratégia é a *First Three* (FT) que cria um resumo com as três primeiras sentenças de cada seção. Esta é o nosso *baseline vanilla*.

Para comparação com o estado-da-arte, selecionamos cinco modelos de sumarização, incluindo sumarização abstrativa e híbrida. Dois dos modelos comparados são modelos de sumarização abstrativos baseados em *transformers* para sequências longas, que são BigBird Pegasus e LED. Outros modelos abstrativos comparados foram os modelos Pegasus, T5 e BART. Entre os modelos baseados em *transformers* comparados, o Pegasus, o T5 e o BART possuem o menor limite para o comprimento da sequência de entrada, sendo este limite de 512 *tokens*. Na Tabela 5.9 e na Tabela 5.10, mostramos os resultados obtidos com os dois experimentos.

Tabela 5.9: Comparação do melhor modelo obtido usando a nossa abordagem, utilizando conjunto de dados da Plos One, com modelos do estado-da-arte e *baselines*.

Alg.	R1 (%)	R2 (%)	RL (%)	BS (%)
Max (Padrão Ouro)	55.65	27.44	38.16	87,55
Min	17.65	2.24	13.76	70,87
FT (Vanilla)	34.61	9.12	22.70	76,29
PEGASUS [Zhang et al., 2020]	28,01	7,62	20,77	75,80
LED [Beltagy et al., 2020]	28,72	7,20	26,90	76,65
BIGBIRD-Pegasus [Zaheer et al., 2020]	30,01	5,84	22,54	75,52
T5 [Raffel et al., 2020]	7,05	2,18	9,77	70,36
BART [Lewis et al., 2019]	6,94	2,15	9,68	69,86
Nossa abordagem	**46,42**	**16,45**	**28,62**	**80,49**

Fonte: Elaborada pela autora.

Tabela 5.10: Comparação do melhor modelo obtido usando a nossa abordagem, utilizando conjunto de dados do ArXiv, com modelos do estado-da-arte e *baselines*.

Alg.	R1 (%)	R2 (%)	RL (%)	BS (%)
Max (Padrão Ouro)	36.92	14.42	25.53	76,43
Min	22.57	4.58	16.62	60,69
FT (Vanilla)	31.39	8.89	20.91	76,08
PEGASUS [Zhang et al., 2020]	30,95	8,33	23	75,8
LED [Beltagy et al., 2020]	30,12	7,81	27,43	75,11
BIGBIRD-Pegasus[Zaheer et al., 2020]	38,01	**13,46**	**27,82**	**80,43**
T5 [Raffel et al., 2020]	8,54	2,33	11,17	70,44
BART [Lewis et al., 2019]	9,06	2,35	11,49	70,22
Nossa abordagem	**38,67**	12,42	25,08	77,76

Fonte: Elaborada pela autora.

Comparando os resultados do nosso melhor modelo com os resultados dos algoritmos do estado-da-arte, verificamos que, para o conjunto de dados da Plos One, o modelo com melhor desempenho, em termos de BS, foi o modelo que utiliza a nossa abordagem de treinamento multi-seção com a representação de atributos. Para o conjunto de dados do ArXiv, a nossa abordagem teve um desempenho inferior em termos de R2, RL e BS do modelo BIGBIRD-Pegasus, mas, superou esse mesmo modelo em termos de R1. Vale ainda ressaltar que os modelos de sumarização que possuem limitação de 512 *tokens* foram os que apresentaram o pior desempenho. Isso mostra a necessidade de explorar outras abordagens de sumarização que não possuem limitações do número de *tokens*. Diante disso, podemos concluir que a abordagem de sumarização aqui proposta pode obter resultados competitivos e até superiores a abordagens do estado-da-arte em sumarização de textos.

Capítulo 6

Conclusão e Trabalhos Futuros

Atualmente, a sumarização de textos longos tem sido cada vez mais explorada. No entanto, neste trabalho, mostramos que a literatura de sumarização de textos longos ainda possui algumas lacunas que precisam ser exploradas. Dentre essas lacunas está a limitação de *tokens* de entrada dos modelos de última geração. Para mitigar esse problema, propusemos e avaliamos uma abordagem ETS capaz de trabalhar com todo o texto. Nesta dissertação, exploramos diferentes estratégias de treinamento de algoritmos para sumarização extrativa de textos longos e avaliamos, também, o desempenho desses algoritmos com diferentes formas de representação das sentenças do texto. Dentre as estratégias de treinamento, exploramos o treinamento utilizando um único texto como entrada para os algoritmos e também segmentando os textos e submetendo para os algoritmos cada seção, separadamente. Como estratégia de representação das sentenças, utilizamos representações de visão única e representações de multi-visões. Com esses experimentos, chegamos às seguintes respostas para as perguntas de pesquisa colocadas na seção de introdução.

Qual a melhor estratégia de treinamento de modelos de sumarização extrativa de textos longos, seção única ou multi-seção? Nossos experimentos mostram que a abordagem de sumarização multi-seção apresenta, em todos os casos, resultados superiores a abordagem de seção única. Houve um ganho de desempenho de, aproximadamente, 14% de BS para a base da Plos One e um ganho de, aproximadamente, 5% de BS para a base do ArXiv. Isso mostra que a abordagem multi-seção é capaz de produzir resultados de melhor qualidade do que a abordagem de seção única quando se trata de sumarizar textos longos.

Qual a melhor forma de representar as sentenças de modo que permita identificar a sua importância para o texto? Nossos experimentos mostram que a melhor forma de representação das sentenças do texto, dentre as testadas, é a representação com matrizes de atributos. Isso nos permite concluir que a importância das sentenças do texto é melhor detectada usando atributos posicionais, medidas de centralidade e ranqueamento de outros algoritmos do que usando, por exemplo, *embeddings* contextuais extraídos de modelos *transformers* pré-treinados com grandes conjuntos de dados. Vale ainda ressaltar que a representação fundida teve, em geral, desempenho inferior a representação de atributos.

Acreditamos que isso tenha ocorrido pois a matriz de atributos e a matriz de *embeddings* não se complementam, tornando a fusão dessas visões de pior qualidade. Além disso, podemos concluir com os experimentos, que a visão de *embeddings* teve, em todos os casos, desempenho inferior à representação de atributos e à representação fundida. Isso nos mostra que a sumarização de textos não pode ser realizada apenas com os *embeddings*.

Além de analisar o percentual de ganho dos algoritmos com cada estratégia de treinamento e cada forma de representação, conduzimos testes estatísticos a fim de identificar se a diferença entre as médias dos algoritmos é estatisticamente significante. Para isso, inicialmente, usando o teste de Friedman e, posteriormente, o teste *post hoc* de Nemenyi. Os testes estatísticos foram conduzidos de modo a realizar três comparações. A primeira comparação tem como objetivo identificar qual abordagem de visão única apresenta o melhor desempenho. A segunda comparação tem como objetivo avaliar a diferença entre os métodos de fusão testados. Por fim, a terceira comparação foi conduzida a fim de comparar o método de fusão proposto com a melhor representação de visão única que é a representação de atributos. Para a primeira comparação os resultados do teste *post hoc* de Nemenyi permitiu que concluíssemos que a representação de atributos é melhor que a representação de *embeddings* na tarefa de identificar a importância das sentenças do texto. Para o segundo experimento nós concluímos que o método $Fusão_2$ apresentou melhores resultados que o método $Fusão_1$. Por fim, para o terceiro experimento, concluímos que a $Visão_1$ apresentou o melhor resultado para o conjunto de dados da Plos One e, para o conjunto de dados do ArXiv, não foi possível afirmar que há uma diferença estatisticamente significante entre GB com $Visão_1$ e GB com $Fusão_1$.

Para trabalhos futuros, pretendemos estudar e testar diferentes formas de combinação das sentenças na criação dos resumos. Atualmente, a estratégia utilizada é uma estratégia gulosa, que seleciona as sentenças melhor ranqueadas e as concatena gerando um resumo. Porém, não é claro para nós que essa é a melhor estratégia. Sendo assim, pretendemos estudar outras formas de combinar essas sentenças considerando a coesão, coerência e diversidade das sentenças do resumo. Além disso, pretendemos expandir o número de métricas utilizadas na avaliação de qualidade dos resumos. De acordo com Souza et al. [2021], as métricas ROUGE não são as melhores métricas para avaliação da qualidade de resumos. Tendo isso em vista, adicionamos, nesta dissertação, a métrica BERTScore. Contudo, nós acreditamos que há uma necessidade de estudarmos e avaliarmos o desempenho dos nossos resumos com outras métricas como *NeUral Based Interchangeability Assessor* (NUBIA) [Kane et al., 2020] e BLEURT [Sellam et al., 2020] a fim de obtermos diferentes avaliações de qualidade. Pretendemos, também, realizar uma análise da importância dos atributos para os algoritmos de classificação. Por fim, há a necessidade de explorar métricas de coesão e diversidade. Espera-se que, ao explorar a coesão dos textos gerados, teremos resumos de melhor qualidade sem sentenças truncadas. Além disso, espera-se que a exploração da diversidade das sentenças nos proporcione

resumos de melhor qualidade sem sentenças redundantes.

Referências

Nouf Ibrahim Altmami e Mohamed El Bachir Menai. Automatic summarization of scientific articles: A survey. *Journal of King Saud University-Computer and Information Sciences*, 2020.

Tadas Baltrušaitis, Chaitanya Ahuja, e Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

Iz Beltagy, Matthew E Peters, e Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jiaao Chen e Diyi Yang. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. *arXiv preprint arXiv:2010.01672*, 2020.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, e Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. *arXiv preprint arXiv:1804.05685*, 2018.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine learning research*, 7:1–30, 2006.

Yue Dong, Andrei Mircea Romascanu, e Jackie Chi Kit Cheung. Discourse-aware unsupervised summarization for long scientific documents. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1089–1102, 2021.

Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, e Hoda K Mohamed. Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, 165: 113679, 2021.

Günes Erkan e Dragomir R Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, 22:457–479, 2004.

Mohamed Abdel Fattah. A hybrid machine learning model for multi-document summarization. *Applied intelligence*, 40(4):592–600, 2014.

Alexios Gidiotis e Grigorios Tsoumakas. A divide-and-conquer approach to the summarization of long documents. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:3029–3040, 2020.

Luís Gonçalves e Renato Vimieiro. Approaching authorship attribution as a multi-view supervised learning task. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2021. doi: 10.1109/IJCNN52387.2021.9533360.

Hongyu Guo, Colin Cherry, e Jiang Su. End-to-end multi-view networks for text classification. *arXiv preprint arXiv:1704.05907*, 2017.

Mandy Guo, Joshua Ainslie, David Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, e Yinfei Yang. Longt5: Efficient text-to-text transformer for long sequences. *arXiv preprint arXiv:2112.07916*, 2021.

Hassan Kane, Muhammed Yusuf Kocyigit, Ali Abdalla, Pelkins Ajanoh, e Mohamed Coulibali. Nubia: Neural based interchangeability assessor for text generation. *arXiv preprint arXiv:2004.14667*, 2020.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, e Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Jinxing Li, Zhaoqun Li, Guangming Lu, Yong Xu, Bob Zhang, e David Zhang. Asymmetric gaussian process multi-view learning for visual classification. *Information Fusion*, 65:108–118, 2021. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2020.08.020. URL https://www.sciencedirect.com/science/article/pii/S156625352030350X.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics. Disponível em: ¡https://www.aclweb.org/anthology/W04-1013.pdf¿. Acesso em: 4 de abr. 2020.

Yang Liu e Mirella Lapata. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*, 2019.

Rada Mihalcea e Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004.

Derek Miller. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*, 2019.

Mudasir Mohd, Rafiya Jan, e Muzaffar Shah. Text document summarization using word embedding. *Expert Systems with Applications*, 143:112958, 2020. Disponível em: ¡https://www.sciencedirect.com/science/article/abs/pii/S0957417419306761¿. Acesso em: 4 de abr. 2020.

N Moratanch e S Chitrakala. A survey on extractive text summarization. In *2017 international conference on computer, communication and signal processing (ICCCSP)*, pages 1–6. IEEE, 2017.

N Nazari e MA Mahdavi. A survey on automatic text summarization. *Journal of AI and Data Mining*, 7(1):121–135, 2019.

Bo Pang, Erik Nijkamp, Wojciech Kryściński, Silvio Savarese, Yingbo Zhou, e Caiming Xiong. Long document summarization with top-down and bottom-up inference. *arXiv preprint arXiv:2203.07586*, 2022.

Jason Phang, Yao Zhao, e Peter J Liu. Investigating efficiently extending transformers for long input summarization. *arXiv preprint arXiv:2208.04347*, 2022.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

Lawrence H Reeve, Hyoil Han, e Ari D Brooks. The use of domain-specific concepts in biomedical text summarization. *Information Processing & Management*, 43(6):1765–1776, 2007.

Jesus M Sanchez-Gomez, Miguel A Vega-Rodríguez, e Carlos J Pérez. Extractive multi-document text summarization using a multi-objective artificial bee colony optimization approach. *Knowledge-Based Systems*, 159:1–8, 2018.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

Thibault Sellam, Dipanjan Das, e Ankur P Parikh. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*, 2020.

Alan Silva Romualdo, Livy Real, e Helena de Medeiros Caseli. Classificação multimodal para detecção de produtos proibidos em uma plataforma marketplace. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 111–120. SBC, 2021.

Shashi Pal Singh, Ajai Kumar, Abhilasha Mangal, e Shikha Singhal. Bilingual automatic text summarization using unsupervised deep learning. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 1195–1200. IEEE, 2016.

Cinthia Souza e Renato Vimieiro. A long texts summarization approach to scientific articles. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 182–189, Porto Alegre, RS, Brasil, 2021. SBC. doi: 10.5753/stil.2021.17797. URL https://sol.sbc.org.br/index.php/stil/article/view/17797.

Cinthia M Souza, Magali RG Meireles, e Paulo EM Almeida. A comparative study of abstractive and extractive summarization techniques to label subgroups on patent dataset. *Scientometrics*, 126(1):135–156, 2021.

Amy Trappey, Charles Trappey, e Chun-Yi Wu. Automatic patent document summarization for collaborative knowledge systems and services. *Journal of Systems Science and Systems Engineering*, 18(1):71–94, 2009. Disponível em: ¡https://ir.nctu.edu.tw/bitstream/11536/7574/1/000264362800005.pdf¿. Acesso em: 29 de mar. 2020.

Dingding Wang, Shenghuo Zhu, Tao Li, Yun Chi, e Yihong Gong. Integrating document clustering and multidocument summarization. *ACM Trans. Knowl. Discov. Data*, 5 (3), August 2011. ISSN 1556-4681. doi: 10.1145/1993077.1993078. Disponível em: ¡https://dl.acm.org/doi/abs/10.1145/1993077.1993078¿. Acesso em: 29 de mar. 2020.

Wen Xiao e Giuseppe Carenini. Extractive summarization of long documents by combining global and local context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 3011–3021. Association for Computational Linguistics, 2019.

Chang Xu, Dacheng Tao, e Chao Xu. A survey on multi-view learning. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 2531–2544, 2013.

Shansong Yang, Weiming Lu, Zhanjiang Zhang, Baogang Wei, e Wenjia An. Amplifying scientific paper's abstract by leveraging data-weighted reconstruction. *Information Processing & Management*, 52(4):698–719, 2016.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. Big bird: Transformers for longer sequences. In *NeurIPS*, 2020.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, e Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

Yong Zhang, Meng Joo Er, Rui Zhao, e Mahardhika Pratama. Multiview convolutional neural networks for multidocument extractive summarization. *IEEE transactions on cybernetics*, 47(10):3230–3242, 2016.

Jing Zhao, Xijiong Xie, Xin Xu, e Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2017.02.007. URL https://www.sciencedirect.com/science/article/pii/S1566253516302032.

Apêndice A

Lista de siglas

AP Aprendizado Profundo

ETS Extractive Text Summarization

ATS Abstractive Text Summarization

HTS Hybrid Text Summarization

ROUGE Recall-Oriented Understudy for Gisting Evaluation

R1 ROUGE-1

R2 ROUGE-2

RL ROUGE-L

LCS Longest Common Substring

BS BERTScore

LSA Latent semantic analysis

TF-ISF Term Frequency - Inverse Sentence Frequency

TF-IDF Term Frequency - Inverse Document Frequency

Leaky ReLu Leaky Rectified Linear Unit

kNN K Nearest Neighbor

RF Random Forest

AB Adaptive Boost

CB Cat Boost

GB Gradient Boost

MLP Multi Layer Perceptron

FT First Three

NUBIA NeUral Based Interchangeability Assessor

Apêndice B

Hiperparâmetros testados no *Randomized Search*

Tabela B.1: Hiperparâmetros testados no *Randomized Search*

Alg.	Parâmetros	Valores
kNN	n_neighbors	3, 5, 10
AB	n_estimators	10, 25, 50, 100, 200
RF, GB	n_estimators	10, 25, 50, 100, 200
	min_samples_leaf	1000, 2500, 4000, 5000
	min_samples_split	2000, 5000, 10000
	max_depth	2, 3, 5, 10
CB	iterations	10, 25, 50, 100
	learning_rate	0.01, 0,001
	depth	2, 3, 5, 10
	min_data_in_leaf	1000, 2500, 5000

Fonte: Elaborada pela autora.

Apêndice C

Hiperparâmetros selecionados

Tabela C.1: Hiperparâmetros selecionados usando *Randomized Search* para a base de dados da Plos One

Alg.	Parâmetros	Introdução	Materiais e Métodos	Resultados e Conclusão	Concatenação
kNN	*n_neighbors*	10	10	10	10
AB	*n_estimators*	200	200	200	200
RF	*n_estimators*	200	100	100	100
	min_samples_split	4000	4000	4000	4000
	min_samples_leaf	2000	2000	2000	2000
	max_depth	10	5	5	5
CB	*min_data_in_leaf*	2500	1000	2500	2500
	learning_rate	0,001	0,001	0,001	0,001
	iterations	25	10	100	10
	depth	10	2	5	2
GB	*n_estimators*	200	200	200	100
	min_samples_split	4000	4000	4000	4000
	min_samples_leaf	2000	2000	2000	2000
	max_depth	2	10	10	5

Fonte: Elaborada pela autora.

Tabela C.2: Hiperparâmetros selecionados usando *Randomized Search* para a base de dados do ArXiv

Alg.	Parâmetros	Introdução	Materiais e Métodos	Resultados e Conclusão	Concatenação
kNN	*n_neighbors*	10	10	10	10
AB	*n_estimators*	200	200	200	200
RF	*n_estimators*	200	100	100	100
	min_samples_split	4000	4000	4000	4000
	min_samples_leaf	2000	2000	2000	2000
	max_depth	10	5	5	5
CB	*min_data_in_leaf*	2500	2500	2500	2500
	learning_rate	0.001	0.001	0.001	0.001
	iterations	100	10	100	10
	depth	10	2	10	2
GB	*n_estimators*	200	200	200	100
	min_samples_split	4000	4000	4000	4000
	min_samples_leaf	2000	2000	2000	2000
	max_depth	2	5	10	5

Fonte: Elaborada pela autora.

Apêndice D

Exemplos de resumos de seção única e multi-seção

Neste apêndice são apresentados exemplos de resumos gerados com a abordagem de seção única e multi-seção com os conjuntos de dados da Plos One e do ArXiv. Os exemplos reportados neste apêndice foram obtidos usando o algoritmo que obteve o melhor resultado na comparação, que foi o GB.

Quadro D.4: Resumos obtidos utilizando a abordagem de seção única e multi-seção no conjunto de dados da Plos One.

Referência	the hematopoietic and neurologic expressed sequence 1 (hn1) gene encodes a highly conserved protein that is expressed in developing and regenerating tissues. in this study, hn1 expression was evaluated in human and murine malignant gliomas. hn1 mrna and protein were detected in the murine gl261 glioma cell line and in gl261 brain tumors in vivo. hn1 is also expressed in human u118mg and u87mg cell lines. evaluation of human brain tumors using an anti hn1 polyclonal antibody detected strong immunoreactivity in high grade (who iii and iv) malignant gliomas. the rate of gl261 cell proliferation in vitro was unaltered by hn1 depletion using an anti hn1 sirna. however, tumors established from hn1 depleted gl261 cells formed significantly smaller volumes than those established from control treated cells. these data suggest a role for hn1 in the biology of malignant brain tumors.'
Seção única	GL261 cells were seeded on a 12 well cell culture plate at a density of 300 000 cells per well . Two days later cells were rinsed with serum free RPMI and transduced with serum free media containing either the Hn1 siRNA AAV6 or the control H1 AAV6 . Twenty days after GL261 cell implantation the mice were given an injection of sodium pentobarbital 32 mg kg and euthanized by transcardial perfusion with 0 9 saline followed by buffered 4 paraformaldehyde PFA . This was followed by incubation with a biotinylated secondary antibody DakoCytomation Copenhagen Denmark . vitro cell proliferation assay . Two days later cells were transduced at an M O I of 5 000 with Hn1 siRNA AAV6 control H1 AAV6 or no virus . Two days after transduction triplicate wells of cells were collected by trypsinization and counted manually using a hemacytometer . tumor volume calculation . The volumes from all the sections pertaining to a single tumor were summed to obtain the final tumor volume .
Multi-seção	Identifying the roles of the various genes involved in the aberrant growth properties of tumors is a valuable goal toward uncovering mechanisms of tumorigenesis . The GL261 murine glioma model is appropriate for the study of treatments against glioma because it shares numerous characteristics with human gliomas . Moreover HN1 expression was detected in the human glioma cell lines U118MG and U87MG as well as high grade malignant human brain gliomas .GL261 cells obtained from the NCI Frederick MD were grown in RPMI Gibco BRL containing 10 FBS 1 penicillin streptomycin and 4 mM L glutamine . B16 F10 and HEK293 cells were grown in DMEM Gibco BRL containing 10 FBS 1 penicillin streptomycin and 1 sodium pyruvate . The sodium bicarbonate content in DMEM was 3 7 g L for HEK293 cells and 1 5 g L for B16 F10 cells .Hn1 protein in GL261 cells cultured in vitro was detected by Western blot analysis using a rabbit polyclonal anti murine Hn1 antibody . A human brain tumor tissue microarray containing multiple WHO tumor grade III and IV infiltrating gliomas was subjected to immunohistochemical analysis using the polyclonal anti Hn1 antibody . As such we explored the expression of Hn1 in a murine model of malignant glioma as well as in high grade human gliomas and determined that each of these brain tumors express the Hn1 protein .

Fonte: Elaborado pela autora.

Tabela D.1: Métricas para o exemplo com o conjunto de dados da Plos One

Abordagem	R1 (%)	R2 (%)	RL (%)	BS (%)
Seção única	25,76	2,65	19,98	76,09
Multi-seção	51,23	14,56	29,87	83,01

Fonte: Elaborada pela autora.

Quadro D.5: Resumos obtidos utilizando a abordagem de seção única e multi-seção no conjunto de dados da ArXiv.

Referência	experimental and theoretical studies are made of brownian particles trapped in a periodic potential , which is very slightly tilted due to gravity . in the presence of fluctuations , these will trigger a measurable average drift along the direction of the tilt . the magnitude of the drift varies with the ratio between the bias force and the trapping potential . this can be closely compared to a theoretical model system , based on a fokker - planck - equation formalism . we show that the level of control and measurement precision we have in our system , which is based on cold atoms trapped in a 3d dissipative optical lattice , makes the experimental setup suitable as a testbed for fundamental statistical physics . we simulate the system with a very simplified and general classical model , as well as with an elaborate semi - classical monte - carlo simulation . in both cases , we achieve good qualitative agreement with experimental data .
Seção única	closely related to this are systems where noise or fluctuations is the source for directed drift so called brownian motors see . a classical particle in the above predicament with vertical position coordinate xmath0 will follow the langevin equation xmath1 here xmath2 is the mass xmath3 is a uniform damping constant xmath4 is a uniform external force and xmath5 is a langevin stochastic force . we let the noise term and the friction scale linearly with the potential depths for simplicity and generalization. this leaves the potential depth xmath19 as the main variable parameter but it is only accessible through the laser irradiance which also modifies diffusion and friction . a very general problem in physics is that of a brownian particle moving in a periodic potential a seminal treatment using fokker planck formalism is given by risken . however these dissipative optical lattices put the atoms in a regime where fluctuations play a dominating role and where dissipation is also present . in addition the setup used here is very close to the double optical lattice arrangement that has been used to create a brownian motor . of particular interest is the tilted washboard potential where the brownian particle is also subjected to a constant force which can actually be used to model a wide variety of physical systems see . however these dissipative optical lattices put the atoms in a regime where fluctuations play a dominating role and where dissipation is also present .
Multi-seção	however in actual experiments with dissipative optical lattices this cooling mechanism may be relevant for the initial damping of the thermal energy and for the first phases of the route to equilibrium but the atoms will quickly loose enough thermal energy in order to be trapped in the potential wells of the lattice and at equilibrium they indeed typically get localized close to the bottom of the potentials c f . when it gets unlocked it will be again be exposed to laser cooling it will loose its kinetic energy and it will be trapped again in some bound state as this goes on there will be a gradual accumulation towards lower lying and more deeply trapped states from which the escape probability is low and eventually an equilibrium will be reached furthermore the deeper the potentials are the larger the portion of atoms that are trapped but even for very shallow potentials the majority of the atoms are trapped or correspondingly one atom spends most of its time being trapped interrupted by short periods of inter well flight where it can travel over several wells in the current work . we make the working hypothesis that when an atom is trapped locked state its motion is undamped if and when it becomes untrapped running state an effective friction xmath13 with lc standing for laser cooling turns on which we assume can be reasonably well approximated by the spatial average used in i e an untrapped atom is subjected to dissipation of its momentum as in the traditional picture of laser cooling .for the vast majority of experiments done with dissipative optical lattice the holding time in the optical lattice has been rather short . the potential depth is varied by adjusting the irradiances and the detunings of the optical lattice laser beams from the time of flight data. the solid line is from a semi classical monte carlo simulation while the dashed line is from a simplified classical simulation in fig .to conclude we have made a quantitative study of how random isotropic fluctuations together with a very small bias force gives rise to an average drift . when the fraction between bias force and trapping potential xmath20 is varied the magnitude of this drift can change . we qualitatively reproduce our data with a simplified classical simulation as well as with a careful semi classical monte carlo simulation of the laser cooling setup .

Fonte: Elaborado pela autora.

Tabela D.2: Métricas para o exemplo usando o conjunto de dados do ArXiv

Abordagem	R1 (%)	R2 (%)	RL (%)	BS (%)
Seção única	29,83	7,86	21,15	77,14
Multi-seção	35,43	9,08	23,97	77,72

Fonte: Elaborada pela autora.

Para ambos os conjuntos de dados, podemos verificar que a abordagem multi-seção produz resumos de melhor qualidade que a abordagem de seção única. Podemos ver, a partir dos exemplos, que os resumos de seção única capturam menos informações que os resumos de múltiplas seções, embora o número de sentenças nesses resumos seja o mesmo. Quando analisamos os resumos criados usando múltiplas seções, verificamos que a qualidade do resumo melhora. Em geral, os resumos de múltiplas seções cobrem melhor o conteúdo do texto e, por isso, proporcionam uma visão geral do texto de modo mais abrangente. Esse comportamento pode ser verificado, também, a partir das métricas obtidas com cada resumo. Em todos os casos os resumos multi-seção apresentaram desempenho superior aos de seção única. Mostrando assim que o resumo multi-seção é melhor em capturar as informações do resumo referência.

Apêndice E

Exemplos de resumos de visão única e multi-visão

Neste apêndice são apresentados exemplos de resumos gerados usando a abordagem de visão única e a abordagem multi-visão associado ao algoritmo que obteve o melhor desempenho, que foi o GB. Para cada exemplo comparamos os resultados da abordagem de visão única que usa a representação de atributos com a abordagem multi-visão que usa $Fusão_1$.

Quadro E.6: Resumo gerado com o algoritmo GB usando a representação de atributos e a representação gerada com $Fusão_1$ com o conjunto de dados da Plos One.

Referência	purpose studies herein explore paclitaxel enhancement of the therapeutic efficacy of particle targeted radiation therapy. experimental design athymic mice bearing 3 d i.p. ls 174t xenografts were treated with 300 or 600 g of paclitaxel at 24 hr prior to, concurrently or 24 hr post 213bi or 212pb trastuzumab. results paclitaxel (300 or 600 g) followed 24 h later with 213bi trastuzumab (500 ci) provided no therapeutic enhancement. paclitaxel (300 g) administered concurrently with 213bi trastuzumab or 213bi huigg resulted in median survivals (ms) of 93 d and 37 d, respectively; no difference was observed with 600 g of paclitaxel. mice receiving just 213bi trastuzumab, 213bi huigg, or left untreated had a ms of 31, 21 and 15 d, respectively; 23 d for just either paclitaxel dose alone. paclitaxel (300 or 600 g, respectively) given 24 hr post 213bi trastuzumab increased ms to 100 and 135 d, respectively. the greatest improvement in ms (198 d) was obtained with two weekly doses of paclitaxel (600 g) followed by 213bi trastuzumab. studies were also conducted investigating paclitaxel administered 24 hr pre , concurrently, or 24 hr post 212pb trastuzumab (10 ci). the 300 g dose of paclitaxel 24 hr pre rit failed to provide benefit while 600 g extended the ms from 44 d to 171 d. conclusions these results suggest that regimens combining chemotherapeutics and high let radioimmunotherapy may have tremendous potential in the management and treatment of cancer patients. dose dependency and administration order appear to be critical factors requiring careful investigation. statement of clinical relevance these investigations reported herein demonstrate the potential of combining chemotherapeutics with high let radioimmunotherapy for the management and treatment of cancer patients who present with disseminated peritoneal disease at the time of their diagnosis. these studies are a natural progression to prior studies that established the efficacy of paclitaxel administered in conjunction with radiation radioimmunotherapy for the treatment of ovarian patients. chemotherapy in conjunction with particle radioimmunotherapy using the appropriate targeting vehicle would be an effective adjuvant therapy following procedures such as cytoreductive surgery or peritioneal external beam radiation therapy. the intent of developing a treatment regimen utilizing targeted radiation is to expand the repertoire to patient populations. such a strategy would be potentially beneficial for not only those with pancreatic or ovarian cancer but also those with cancers of the colon, stomach and small intestine, that result in peritoneal carcinomatosis as well as those with peritoneal mesothelioma. in the general, the results obtained define a potentiating interaction between paclitaxel and the high let radiation labeled trastuzumab. the studies also illustrate the necessity of establishing the optimal administration sequence of the treatment components, and that dose dependency and administration order are critical factors that require careful investigation.
Visão única	learly development of new tactics for the treatment and management of patients with pancreatic or ovarian cancer remains a high priority 1 . Paclitaxel has activity against ovarian and pancreatic cancer and is a recognized radiosensitizer which has been reported to sensitize ovarian and pancreatic cancer cells tumors to the cytotoxic effects of radiation 23 25 . The studies described herein are an evaluation of the ability of paclitaxel to potentiate the therapeutic efficacy of HER2 targeting emitting high LET 213Bi and 212Pb labeled trastuzumab in a multimodality regimen for the management of disseminated intraperitoneal diseaseAll therapy studies were conducted using the LS 174T a human colon carcinoma cell line SKOV 3 a human ovarian carcinoma cell line that expresses 1106 HER2 molecules per cell was used for in vitro analysis 32 . Elution of the 213Bi and radiolabeling of trastuzumab CHX A was performed as previously described 10 . Additional groups of mice included those injected with 1 2 or 3 doses of paclitaxel at weekly intervals with either 213Bi trastuzumab or 213Bi HuIgG or no radiation treatment .The corresponding groups of mice treated with 213Bi HuIgG and paclitaxel had a median survival of only 36 38 43 and 63 d for zero one two or three doses of paclitaxel . Since therapeutic efficacy was observed with paclitaxel administered concurrently with the 213Bi trastuzumab the experiment combining paclitaxel and 212Pb trastuzumab evaluated the effects of paclitaxel given prior to concomitantly and post 212Pb trastuzumab . Therapeutic Index treatment group median survival divided by the untreated group median survival
Multi-visão	The outlook for patients with ovarian cancer is not appreciably better . Over the years there has been a trend towards improvement in this situation by a few percentage points though that seems imperceptible in the 5 yr RSRs for both diseases . Paclitaxel has activity against ovarian and pancreatic cancer and is a recognized radiosensitizer which has been reported to sensitize ovarian and pancreatic cancer cells tumors to the cytotoxic effects of radiation 23 25 .The number of CHX A DTPA or TCMC molecules linked to the monoclonal antibody mAb was determined using spectrophotometric assays based on the titration of either ytrrium or lead Arsenazo III complex respectively 37 38 . Additional groups of mice included those injected with 1 2 or 3 doses of paclitaxel at weekly intervals with either 213Bi trastuzumab or 213Bi HuIgG or no radiation treatment . These treatment groups were compared to sets of mice that received only paclitaxel 212Pb trastuzumab 212Pb HuIgG or to those without any treatment .As shown in Table 2 the median survival of the untreated group was 25. d There was no increase P 0 82 in median survival for those mice receiving one two or three doses of paclitaxel only median survival was 45 63 and 46 d in these three groups respectively . Since therapeutic efficacy was observed with paclitaxel administered concurrently with the 213Bi trastuzumab the experiment combining paclitaxel and 212Pb trastuzumab evaluated the effects of paclitaxel given prior to concomitantly and post 212Pb trastuzumab . Therapeutic Index treatment group median survival divided by the untreated group median survival

Fonte: Elaborado pela autora.

Tabela E.1: Métricas para o exemplo com o conjunto de dados do Plos One

Abordagem	R1 (%)	R2 (%)	RL (%)	BS (%)
Visão única	46,11	13,90	20,36	82,16
Multi-visão	41,07	8,74	18,95	80,23

Fonte: Elaborada pela autora.

Quadro E.7: Resumo gerado com o algoritmo GB usando a representação de atributos e a representação fundida com $Fusão_1$ com o conjunto de dados do ArXiv.

Referência	we studied the formation process of star clusters using high resolution @xmath0 body / smoothed particle hydrodynamcs simulations of colliding galaxies . the total number of particles is @xmath1 for our high resolution run . the gravitational softening is @xmath2 and we allow gas to cool down to @xmath3 . during the first encounter of the collision , a giant filament consists of cold and dense gas found between the progenitors by shock compression . a vigorous starburst took place in the filament , resulting in the formation of star clusters . the mass of these star clusters ranges from @xmath4 . these star clusters formed hierarchically : at first small star clusters formed , and then they merged via gravity , resulting in larger star clusters .
Seção única	there have been however only a few numerical studies of star cluster formation in merging galaxies even though it has been shown that resolving a cloudy multiphase interstellar medium ism and or clustered star formation can have important consequences for the formation history of early type galaxies . some of the existing studies adopted sub grid models of star cluster formation . we report the result of merger simulations that capture the multiphase nature of the ism and include realistic models of star formation and feedbackwe prepared two identical progenitor galaxies and then let them merge from a parabolic and coplanar configuration . the mass of components in one progenitor galaxy is xmath5 for the dark matter halo xmath6 for the stellar disk and xmath7 for the gas disk . the galaxies are modeled using both xmath0 body and smoothed particle hydrodynamics sph particles .at the first encounter after xmath16 from the beginning of the simulations strong shocks took place . width 480 figure shows the snapshots of the early phase of the star cluster formation where we can see that a number of small star clusters formed along the gas filament . the late phase of the star cluster formation xmath22 was mainly driven by mergers of star clusters without gas since gas in the filament was blown out by sne .
Multi-seção	these star clusters could potentially evolve into the present day metal rich globular clusters and so they are widely accepted to be a good candidate for globular cluster progenitors . there have been however only a few numerical studies of star cluster formation in merging galaxies even though it has been shown that resolving a cloudy multiphase interstellar medium ism and or clustered star formation can have important consequences for the formation history of early type galaxies . some of the existing studies adopted sub grid models of star cluster formation .we prepared two identical progenitor galaxies and then let them merge from a parabolic and coplanar configuration . the galaxies are modeled using both xmath0 body and smoothed particle hydrodynamics sph particles . we employed xmath8 particles for the two progenitor galaxies for the finest runs where the corresponding mass of each particle is xmath9 for both xmath0 body and sph particles .width 480 figure shows the snapshots of the early phase of the star cluster formation where we can see that a number of small star clusters formed along the gas filament . the late phase of the star cluster formation xmath22 was mainly driven by mergers of star clusters without gas since gas in the filament was blown out by sne . since the formation of star clusters is driven by mergers the shape of the mass function becomes a power law reflecting the scale free nature of gravity .

Fonte: Elaborado pela autora.

Tabela E.2: Métricas para o exemplo com o conjunto de dados do ArXiv

Abordagem	R1 (%)	R2 (%)	RL (%)	BS (%)
Visão única	41,54	13,73	28,43	82,44
Multi-visão	37,43	13,48	27,08	81,52

Fonte: Elaborada pela autora.

Ao analisar os resultados apresentados neste apêndice verificamos que, para ambos os exemplos apresentados, a abordagem de visão única apresentou um desempenho

superior ao da abordagem de multi-visão tanto em termos de métricas ROUGE quanto de BS. Ao analisar o conteúdo dos resumos podemos constatar que o resumo gerado pela abordagem de visão única é melhor em capturar as informações do resumo referência que a abordagem de multi-visão. No entanto, podemos verificar que, em geral, os resumos gerados não possuem uma coesão entre as sentenças, mostrando assim a necessidade de estudar outras estratégias para combinação dos resumos.