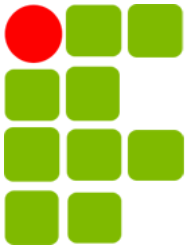


INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

CAMPUS CAMPINA GRANDE

COORDENAÇÃO DA ÁREA DE INFORMÁTICA

CURSO DE GRADUAÇÃO EM ENGENHARIA DE COMPUTAÇÃO

MICAEL MARQUES RODRIGUES SILVA

SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS: DESAFIOS E AVANÇOS EM TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

CAMPINA GRANDE

2023

MICAEL MARQUES RODRIGUES SILVA

SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS: DESAFIOS E AVANÇOS EM TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Monografia apresentada ao Instituto Federal de Educação, Ciência e Tecnologia da Paraíba como requisito para obtenção do título de Bacharel em Engenharia de Computação.

Orientador: Prof. Dr. Alysson Filgueira Milanez- Universidade Federal Rural do Semi-Árido (UFERSA)

CAMPINA GRANDE

2023

S586s Silva, Micael Marques Rodrigues.
Sumarização automática de textos: desafios e avanços em técnicas de processamento de linguagem natural / Micael Marques Rodrigues Silva. - Campina Grande, 2023.
67 f. : il.

Trabalho de Conclusão de Curso (Graduação em Engenharia de Computação) - Instituto Federal da Paraíba, 2023.
Orientador: Prof. Dr. Alysson Filgueira Milanez.

1. Engenharia da Computação 2. Sumarização automática de texto 3. Algoritmo de Marques I. Milanez, Alysson Filgueira II. Título.

CDU 004

MICAEL MARQUES RODRIGUES SILVA

SUMARIZAÇÃO AUTOMÁTICA DE TEXTOS: DESAFIOS E AVANÇOS EM TÉCNICAS DE PROCESSAMENTO DE LINGUAGEM NATURAL

Monografia apresentada ao Instituto Federal de Educação, Ciência e Tecnologia da Paraíba como requisito para obtenção do título de Bacharel em Engenharia de Computação.

Aprovada em: 26 / 05 / 2023

BANCA EXAMINADORA

Prof. Dr. Alysson Filgueira Milanez (Orientador)
Universidade Federal Rural do Semi-Árido
(UFERSA)

Profa. Dra. Ianna Maria Sodré Ferreira De Sousa
Instituto Federal da Paraíba (IFPB)

Prof. Dr. Katyusco de Farias Santos
Instituto Federal da Paraíba (IFPB)

AGRADECIMENTOS

Em primeiro lugar, quero expressar minha profunda gratidão aos meus pais, José Expedito da Silva e Isabel Ana Rodrigues da Silva, por todo o amor, apoio e encorajamento que me proporcionaram ao longo de toda a minha vida e, especialmente, durante o desenvolvimento desta monografia.

Agradeço ao meu irmão, Misael Rodrigues Silva, pelo companheirismo e inspiração, que me motivaram a continuar meus estudos e a buscar meus objetivos.

Gostaria de estender meus sinceros agradecimentos ao meu primo Rubens de Oliveira Rodrigues e sua família, que generosamente me ofereceram sua casa, apoio e conforto, permitindo-me focar nos meus estudos enquanto estava longe da minha família. Serei eternamente grato pela sua hospitalidade e carinho.

Também quero homenagear meus avós, que já se foram, mas que sempre foram uma fonte inesgotável de incentivo e inspiração para que eu perseguisse meus sonhos e me tornasse a pessoa que sou hoje. Levo comigo os valores e a sabedoria que eles compartilharam comigo durante sua vida.

Além disso, sou extremamente grato à Igreja Batista Regular do Catolé por me acolher e cuidar de mim durante todo esse tempo. A comunidade e o apoio espiritual que encontrei nesta igreja foram fundamentais para o meu desenvolvimento pessoal e acadêmico.

Por fim, gostaria de agradecer a todos os amigos, colegas e professores que, de alguma forma, contribuíram para a realização deste trabalho e estiveram ao meu lado durante essa jornada.

A todos vocês, meu mais sincero agradecimento.

EPÍGRAFE

"Alguns homens vêem as coisas como são, e dizem 'Por quê?' Eu sonho com as coisas que nunca foram e digo 'Por que não?"

(Geroge Bernard Shaw)

RESUMO

Neste trabalho, explorou-se a área de sumarização automática de texto, um campo em constante evolução com desafios significativos a superar. O principal foco do estudo foi a comparação do desempenho de seis algoritmos de sumarização automática de textos, a fim de identificar o mais eficaz. Dentre as principais contribuições deste trabalho, destacam-se: a análise comparativa de seis algoritmos de sumarização automática – Algoritmo de Luhn, GistSumm, ChatGPT, Algoritmo de Programação Linear Inteira, Algoritmo de Regressão Bayesiana e Algoritmo de Marques – em um texto sobre a COVID-19 e seus impactos; a implementação de uma metodologia de avaliação pautada em métricas de qualidade da sumarização; e a realização de um estudo com usuários para avaliar a utilidade e relevância dos resumos gerados. Os resultados obtidos indicam que o Algoritmo de Marques superou os demais algoritmos em até 20% em métricas como precisão, coerência, coesão e tempo de processamento. Isso sinaliza que a aplicação do Algoritmo de Marques na sumarização de textos pode ser promissora. Ainda assim, enfatiza-se que a área de sumarização automática de texto enfrenta desafios significativos, tais como a adaptação a diferentes tipos de textos e domínios de conhecimento, a avaliação rigorosa e consistente dos resumos gerados e a geração de resumos personalizados para atender às necessidades específicas dos usuários.

Keywords: Sumarização Automática de Texto. Algoritmo de Marques. Análise Comparativa de Algoritmos. Aprendizado de Máquina.

ABSTRACT

In this work, we explore the field of automatic text summarization, an evolving field with significant challenges to overcome. The main focus of the study was to compare the performance of six summarization algorithms in order to identify the most effective one for automatic summary generation. The main contributions of this work include: the comparative analysis of six automatic summarization algorithms - Luhn Algorithm, GistSumm, ChatGPT, Integer Linear Programming Algorithm, Bayesian Regression Algorithm, and Marques Algorithm - on a text about COVID-19 and its impacts; the implementation of an evaluation methodology based on summarization quality metrics; and a user study to evaluate the usefulness and relevance of the generated summaries. The results indicate that the Marques Algorithm outperformed the other algorithms by up to 20% in metrics such as precision, coherence, cohesion, and processing time. This signals that the application of the Marques Algorithm in automatic document summarization is promissing. However, we emphasize that the field of automatic text summarization still faces significant challenges, such as adapting to different types of texts and domains of knowledge, rigorously and consistently evaluating the generated summaries, and generating personalized summaries to meet specific user needs.

Keywords: Automatic Text Summarization. Marques Algorithm. Summarization Quality Metrics. Machine Learning.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Este capítulo, apresenta a introdução da presente pesquisa. Na Seção 1.1 são demonstradas as justificativas do trabalho, a importância para a área e o tema da pesquisa escolhido. Na Seção 1.2, são descritos os objetivos gerais e específicos. Por sua vez, a Seção 1.3 discute a relevância da pesquisa. A Seção 1.4 lista as principais contribuições que se deseja alcançar com o trabalho. Por fim, a Seção 1.5, sumariza as principais descobertas realizadas no presente trabalho.

Diante da necessidade de usar o computador para ler e editar textos, foi preciso "ensinar" ao computador a identificar textos em uma linguagem natural, e codificá-los para uma linguagem computacional. De acordo com Chowdhary (2020), ocorreram vários avanços por meio de pesquisas na área de Inteligência Artificial (IA), o que possibilitou, a partir de uma de suas subáreas, uma forma para ensinar o computador a identificar palavras, chamada de Processamento de Linguagem Natural (PLN) (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

O PLN (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011) busca soluções para questões computacionais, por meio de uma aprendizagem automática no processamento de linguagem e se dedica a propor e desenvolver modelos computacionais para a realização de tarefas que dependem da língua humana, escrita, como objeto primário. Para isso, linguistas e cientistas da computação, buscam fundamentos em várias disciplinas: Filosofia da Linguagem, Psicologia, Lógica, Inteligência Artificial, Matemática, Ciência da Computação, Linguística Computacional e Linguística (NADKARNI; OHNO-MACHADO; CHAPMAN, 2011).

Segundo Gariba *et al.* (2005), o PLN busca facilitar a interação do *software* com o usuário, o que possibilita além do melhor entendimento, uma forma alternativa de alcançar o que se está procurando.

1.1 Justificativa

A sumarização automática de texto é um processo que envolve a compressão de um texto longo em um resumo mais curto, preservando as informações essenciais e o sentido geral do texto original (GONZALEZ, 2003). Esta técnica permite um acesso mais fácil à informação ao leitor, sendo, portanto, de fundamental importância (LEITE, 2010). Para tal, este trabalho de conclusão de curso se propõe a analisar técnicas de Processamento de Linguagem Natural (PLN) e o uso de dados estatísticos específicos, como frequência de palavras-chave e número

de ocorrências de entidades nomeadas, por exemplo. O desafio consiste em uma redução da dimensionalidade e condensação semântica sem gerar prejuízo ao entendimento (SOUZA *et al.*, 2017).

De acordo com Filho *et al.* (2006), Carlson, Marcu e Okurowski (2003), Lin e Hovy (2003), o processo de sumarização automática de textos adiantaria os estudos e trabalhos de diversos estudantes e pesquisadores. Uns para seus estudos, e outros para elaboração de artigos, dissertações, teses e pesquisas. Esse processo será explicado no decorrer desta pesquisa.

De acordo com Ribaldo, Pardo e Rino (2011), identificar segmentos relevantes em um texto e compô-los para gerar sumários representa um dos principais desafios da sumarização automática. Os sumários, nesse contexto, são os textos resumidos. A relevância desta tarefa está na possibilidade de fornecer, de maneira automatizada, um resumo conciso e informativo de um texto, preservando todas as informações essenciais do texto original (BLACK; JOHNSON, 1988; TORRES-MORENO, 2014; BARZILAY, 1997). Entretanto, a avaliação desses sumários exige um esforço significativo, já que depende fortemente do julgamento humano. Os métodos automáticos de avaliação existentes, embora úteis, não conseguem capturar totalmente as nuances e a subjetividade inerentes à avaliação humana, tornando-os menos precisos e completos em comparação. Além disso, o próprio julgamento humano, embora mais preciso, apresenta desafios devido à subjetividade e interpretação individual que cada tarefa de avaliação pode exigir (KAHNEMAN; SIBONY; SUNSTEIN, 2021).

1.2 Objetivos

Nesta seção, serão explicados os objetivos geral e específicos que esta pesquisa se propõe a alcançar.

1.2.1 Objetivo Geral

O objetivo geral deste trabalho é desenvolver e avaliar um novo algoritmo de sumarização automática destinado especificamente para a produção de resumos de textos acadêmicos, como monografias, artigos e pesquisas científicas.

1.2.2 Objetivos Específicos

Para alcançar o objetivo geral, os seguintes objetivos específicos foram estabelecidos:

- Discutir e identificar os principais desafios enfrentados na área de sumarização automática de textos acadêmicos;
- Revisar e analisar criticamente os algoritmos de sumarização automática existentes na literatura;
- Desenvolver um novo algoritmo de sumarização automática, especificamente voltado para a produção de resumos de textos acadêmicos;
- Avaliar o desempenho e a qualidade dos resumos gerados pelo novo algoritmo em comparação com os algoritmos existentes;
- Discutir os resultados obtidos e como eles contribuem para o avanço da área de sumarização automática de texto.

1.3 Relevância

Ao atingir os objetivos citados na Seção 1.2, esta pesquisa pretende ajudar pesquisadores, acadêmicos, alunos e professores que precisam ler diversos textos em pouco tempo, e com a ajuda dessa pesquisa, conseguirão ter acesso a um resumo inteligente (SOUZA, 2004), facilitando assim o processo de leitura e escrita.

A presente pesquisa também ajudará estudantes da área de ciência da computação, pois pode ser um ponto de partida para aprender, na prática, como a Tecnologia da Informação (TI) está presente no dia a dia da população (MARGARIDO; PARDO; ALUíSIO, 2008).

Além disso, a pesquisa ajudará estudantes de outros cursos superiores que precisam escrever textos, artigos, pesquisas e monografias, e, ao se depararem com o processo de criação do resumo, não precisarão mais criar seu resumo tendo que ler tudo que escreveram, pois os algoritmos citados nessa monografia poderão ser usados para realizar essa tarefa.

1.4 Contribuições

Uma vez que não raros são os casos que se faz necessária a obtenção de resumos de determinados textos e o quanto pode ser um processo trabalhoso a realização dos mesmos, principalmente dependendo do tamanho, torna-se relevante um método que possibilite a sumarização automática (BARBIERI, 2021).

Para isso, existem alguns algoritmos de mineração de textos que proporcionam resumos automatizados. Destaca-se o clássico algoritmo de Luhn (LUHN, 1957), além dos algoritmos de

GistSumm (MULLER; GRANATYR; LESSING, 2015b), o algoritmo de Programação Linear Inteira (PLI) (OLIVEIRA, 2018), o algoritmo de regressão Bayesiana (SODRé; OLIVEIRA, 2019) e o *ChatGPT* (RUDOLPH; TAN; TAN, 2023).

Neste trabalho, serão avaliados seis algoritmos de sumarização automática em um texto sobre a COVID-19 e seus impactos na pandemia, utilizando métricas de qualidade de sumarização. Além disso, serão discutidos os principais desafios enfrentados na área de sumarização automática de texto, contribuindo assim para o avanço do conhecimento e a superação de obstáculos nesse campo de pesquisa.

1.5 Descobertas Relevantes do Trabalho

Neste trabalho de conclusão de curso, aborda-se o desafio da sumarização automática de texto, comparando o desempenho de seis algoritmos de sumarização. Os principais achados obtidos durante a execução deste trabalho são apresentados a seguir:

- A aplicação e análise dos seis algoritmos de sumarização automática (Algoritmo de Luhn, *GistSumm*, *ChatGPT*, Algoritmo de Programação Linear Inteira, Algoritmo de Regressão Bayesiana e Algoritmo de Marques) em um texto sobre a COVID-19 e seus impactos na pandemia, revelou diferenças significativas no desempenho desses algoritmos em termos de precisão, coerência, coesão e tempo de processamento.
- Com base em métricas de qualidade de sumarização, o Algoritmo de Marques demonstrou desempenho levemente superior em relação aos outros algoritmos, sendo identificado como uma escolha com potencial para a geração de resumos automáticos de documentos. Este algoritmo apresentou resumos mais relevantes, coesos e coerentes, com menor tempo de processamento.
- Um estudo com usuários foi conduzido para verificar a utilidade e relevância dos resumos gerados pelos algoritmos de sumarização. Os resultados indicaram uma preferência geral pelo Algoritmo de Marques, confirmando os achados obtidos por meio das métricas de qualidade de sumarização.
- A pesquisa também destacou a necessidade de aprimorar os algoritmos de sumarização para lidar com diferentes tipos de textos e domínios de conhecimento, além de buscar a geração de resumos personalizados de acordo com as necessidades e expectativas dos usuários.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, apresenta-se na Seção 2.1, o significado do termo Processamento de Linguagem Natural (PLN), que é a base para a criação dos algoritmos que se dispõem a fazer a análise e sumarização do texto. Na Seção 2.2, são explicitadas as terminologias utilizadas ao longo dessa pesquisa. Por fim, na Seção 2.3, são discutidos os trabalhos relacionados.

2.1 Processamento de Linguagem Natural (PLN)

O Processamento de Linguagem Natural consiste no desenvolvimento de modelos computacionais para a realização de tarefas que dependem de informações expressas em alguma língua natural (e.g. tradução e interpretação de textos, busca de informações em documentos e interface homem-máquina) (MORO *et al.*, 2018; PEREIRA, 2019).

De acordo com Covington, Nute e Vellino (1997), a pesquisa em PLN está voltada, essencialmente, a três aspectos da comunicação em língua natural:

- som: fonologia;
- estrutura: morfologia e sintaxe;
- significado: semântica e pragmática.

A fonologia está relacionada ao conhecimento dos sons que compõem as palavras de uma língua (JUNIOR *et al.*, 2022). A morfologia reconhece as palavras em termos das unidades produtivas que a compõem (e.g. caçou $\rightarrow$ caç + ou) (LACOTIZ, 2020). A sintaxe define a estrutura de uma frase, com base na forma como as palavras se relacionam nessa frase (Figura 1).

Figura 1 – Árvore Sintática

Fonte: Extraído de (PEREIRA, 2019).

A semântica associa significado a uma estrutura sintática, em termos dos significados das palavras que a compõem (e.g. à estrutura da Figura 1, podemos associar o significado "um animal perseguiu/capturou outro animal") (SAMPAIO; RIBEIRO, 2019). Finalmente, a pragmática verifica se o significado associado à uma estrutura sintática é realmente o significado mais apropriado no contexto considerado (e.g. no contexto predador-presa, "perseguiu/capturou" $\rightarrow$ "comeu") (MORO *et al.*, 2018).

Mesmo com o avanço no relacionamento homem-máquina, a comunicação via linguagem natural continua sendo um desafio: como criar programas capazes de interpretar mensagens codificadas em linguagem natural e decifrá- las para a linguagem de máquina? (SILVA, 2021) Ao longo dos anos, muitas pesquisas e desenvolvimentos ocorreram nos mais diversos ramos do processamento de linguagem natural, com ênfase na sumarização automática, que a maioria considera ser o ponto de partida para o estudo da linguagem natural por meio de computadores (RODRIGUEZ; BEZERRA, 2020).

Para modelar uma linguagem e permitir que as máquinas a compreendam, é necessário um processo de pré-processamento abstrato e estruturado para deixar apenas informações relevantes. Esse pré-processamento reduz o tamanho do vocabulário e torna os dados menos esparsos, um recurso conveniente do processamento computacional (SANTOS; CLEMENTINO; PUGLIESI, 2020).

A seguir serão descritas as técnicas de processamento da linguagem natural, que são a Normalização (Seção 2.1.1), Remoção de *Stopwords* (Seção 2.1.2), Remoção de Numerais (Seção 2.1.3), *Stemização* e Lematização (Seção 2.1.4), Compreensão da Linguagem Natural (seção 2.1.5), e na Seção 2.2 serão descritas as terminologias utilizadas, que são *Tokenização* (Seção 2.2.1), Mineração de Textos (Seção 2.2.2) e a Sumarização Automática (Seção 2.2.3).

2.1.1 Normalização

A normalização abrange tratativas como a *tokenização* (AVILA *et al.*, 2022), transformação de letras maiúsculas para minúsculas, remoção de caracteres especiais, remoção de *tags* HTML/Javascript/CSS, dentre outras (MOTTA, 2018). O processo de *tokenização* visa dividir palavras ou frases em unidades. A *tokenização* lexical *tokeniza* cada palavra como um *token* no texto, reconhecendo-a mesmo que seja tocada por sinais de pontuação (CIDRIM; LOPES; MADEIRO, 2019). Um exemplo de texto *tokenizado* lexicalmente seria:

- Esta é uma sentença.

['Esta', 'é', 'uma', 'sentença', '.']

A *tokenização* sentencial identifica e marca sentenças. Um exemplo seria:

- Esta é a primeira sentença. Esta é a segunda. Esta é a terceira!

['Esta é a primeira sentença.', 'Esta é a segunda.', 'Esta é a terceira!']

A normalização é importante por começar a estruturar o texto, já que os processamentos seguintes atuam a partir de unidades sentenciais e lexicais (PINHO *et al.*, 2021).

2.1.2 Remoção de Stopwords

Uma das tarefas muito utilizadas no pré-processamento de textos é a remoção de *stopwords*. Esse método consiste em remover palavras muito frequentes, tais como "a", "de", "o", "da", "que", "e", "do" entre outras, pois na maioria das vezes não são informações relevantes para a construção do modelo (CARDOZO; FREITAS, 2021). Esse processo, só pode ser aplicado quando realmente as palavras não forem importantes para a compreensão do sentido do texto.

2.1.3 Remoção de Numerais

Outra remoção necessária é a dos numerais presentes no texto. Os numerais não agregam informação relevante por não trazerem carga semântica (JURAFSKY; MARTIN, 2020). O PLN remove também os símbolos que os acompanham, como "R$", "$", "US$", "km", "milhões", "bilhões", dentre outros.

2.1.4 Stemização e Lematização

O processo de *stemização* (do inglês, *stemming*) consiste em reduzir uma palavra ao seu radical (GOBBO, 2019). A palavra "meninas" se reduziria a "menin", assim como "meninos" e "menininhos". As palavras "gato", "gata", "gatos" e "gatas" reduzem-se para "gat". A lematização reduz a palavra ao seu lema, que é a forma no masculino e singular. No caso de verbos, o lema é o infinitivo (CAMPOS; FIGUEIREDO, 2021).

Por exemplo, as palavras "gato", "gata", "gatos" e "gatas" são todas formas do mesmo lema: "gato". Igualmente, as palavras "tiver", "tenho", "tinha", "tem" são formas do mesmo lema "ter". A vantagem de aplicar a *stemização* ou lematização é clara: redução de vocabulário e abstração de significado (SANTOS *et al.*, 2022).

2.1.5 Compreensão da Linguagem Natural

Essa parte do processamento de linguagem natural é responsável por transformar sentenças de um texto em estruturas lógicas, ou seja, é compreender uma frase que carrega um valor que pode ser verdadeiro ou falso (FRUTUOSO; BEDREGAL, 2018).

A compreensão da linguagem natural tem como objetivo facilitar a manipulação de texto por computadores, além de identificar instruções recebidas por humanos e até por outras máquinas. É o processo de construção de uma base semântica formal da linguagem. O significado atribuído a sentenças pode ser interpretado pelo computador, da mesma forma que os humanos o fazem (D'ADDARIO, 2022).

2.2 Terminologia Utilizada

Esta seção, apresenta as principais terminologias utilizadas no decorrer dessa pesquisa.

2.2.1 Tokenização

Uma das principais etapas da operação de normalização é a *tokenização*, que é realizada com o objetivo de quebrar um documento de texto nas menores unidades, mas que representem a mesma semântica original do texto. É usado durante o processamento de linguagem natural para segmentação de palavras, localizando caracteres para quebrar sequências de caracteres no texto Limites por palavra, ou seja, palavras separadas ou frases em unidades (RODRIGUEZ; BEZERRA, 2020).

A *tokenização* também foi exemplificada na Seção 2.1.1.

2.2.2 Mineração de Textos

A mineração de textos é um conjunto de métodos utilizado para navegar, organizar, encontrar e descobrir informações em bases textuais (RODRIGUES; SILVA; GAVA, 2018).

A tecnologia de mineração de textos deriva das técnicas de recuperação de informações e da descoberta de informações estruturadas, por meio do uso de bancos de dados e de procedimentos estatísticos (LIMA, 2022). É uma subárea da extração de informações, porém é utilizada somente para análise em textos.

Por mais que possa parecer similar, a mineração de textos é diferente de mecanismos de

busca, uma vez que na busca o usuário já sabe o que quer encontrar; enquanto na mineração de textos, o usuário descobrirá conhecimento e padrões até então, por ele desconhecidos (SOUZA *et al.*, 2021). Em suma, na busca o usuário pesquisa determinada informação e na mineração é realizada a coleta de novos conhecimentos "escondidos" nos textos.

Mineração de textos (FERREIRA; CORREA, 2021) é o processo de extrair conhecimento não conhecido previamente a partir de fontes textuais, tais como correio, imprensa, transações, *websites*, *newsgroups*, fóruns, listas de correspondência, redes sociais, dentre outros.

2.2.3 Sumarização Automática

A sumarização automática é o processo de selecionar as informações mais importantes de um conjunto de fontes, seja um único texto ou um corpus, para produzir uma versão resumida (MARTINS *et al.*, 2001).

Os textos podem ser de qualquer tipo: notícias, artigos científicos, postagens em blogs, resenhas de filmes, atas de reuniões e muito mais. Eles também podem ser escritos em qualquer idioma sem a necessidade de uma escrita formal ou casta. Obviamente, os métodos de resumo, incluindo algum pré-processamento, podem variar de idioma para idioma, com a grande maioria existente para o inglês (CABRAL *et al.*, 2014).

Computacionalmente explicando, existem duas formas de se abordar o problema da sumarização, a superficial e a profunda (SALVINO *et al.*, 2019). Neste trabalho será abordada a primeira, que utiliza métodos estatísticos e/ou empíricos para obter o sumário. Essa técnica é a mais simples de ser implementada e é utilizada por grande parte dos pesquisadores, porém, pode produzir sumários com problemas de coesão e principalmente de coerência, o que pode deixar o resumo sem um sentido lógico da ordem das frases, apresentando deficiências no sentido das frases (ANTUNES, 2018). Por outro lado, a sumarização profunda realiza uma análise semântica frase a frase no texto, analisando a forma que as frases são construídas e o relacionamento de uma frase a outra (PINHO *et al.*, 2021).

2.3 Trabalhos Relacionados

Esta seção, visa apresentar alguns trabalhos relacionados ao Processamento de Linguagem Natural. Em Singh (2018), são apresentadas as sub-tarefas da tecnologia de extração de informações, além de destacar pesquisas de ponta em várias tarefas onde a extração de informa-

ção é utilizada. Além disso, o artigo destaca os desafios atuais de lidar com o Processamento de Linguagem Natural, em razão da explosão de informações na forma de notícias, artigos, redes sociais, mídia, de uma forma que todo o texto interpretado pela máquina consiga ser identificado, extraído e repassado para o usuário de uma forma entendível.

Rodriguez e Bezerra (2020) apresentam uma forma alternativa de integração do meio jurídico com a tecnologia, expondo meios de como o Processamento de Linguagem Natural poderia ajudar a automatizar o reconhecimento de Entidades Nomeadas (Agentes Públicos) em uma base de portaria. Eles utilizam o NLTK (*Natural Language Toolkit*), que junto da linguagem de programação *Python*, trabalham com dados de linguagem humana para aplicação do PNL. O NLTK é útil para separar as sentenças em um paragrafo, separar as palavras dentro de cada sentença, reconhecer padrões no texto e criar modelos de classificação que permitam identificar nomes próprios dentro de um conjunto de dados.

Vieira, Silva e Cordeiro (2019) apresentam uma análise descritiva acerca do conteúdo das notícias publicadas no Portal da Saúde, através da utilização da técnica de mineração de textos, onde tentam gerar insumos e informações para discussões sobre o impacto da *fake news* na área da saúde. A ideia vem da análise de uma iniciativa lançada pelo Ministério da Saúde sobre o enfrentamento de notícias falsas que tem se espalhado nas mais diversas áreas, como política, finanças e a própria saúde, que se chama "Saúde sem *fake news*".

Tabosa *et al.* (2020) apresenta uma pesquisa iniciada no ano de 2014 sobre o desenvolvimento de um *software* que fosse capaz de criar resumos automáticos de textos baseados em técnicas de PLN, baseando-se na frequência de palavras. Os primeiros testes dessa ferramenta geraram resultados que indicavam uma significativa redução da dimensionalidade dos textos, preservando seu valor semântico. O artigo apresenta os resultados dessa pesquisa mostrando que existiu uma equivalência qualitativa entre os resumos produzidos pela ferramenta e por humanos, mas que deixa a desejar quando se trata do tamanho do resumo gerado, visto que o resultado da sumarização feita pela ferramenta denota textos muito longos.

3 MATERIAIS E MÉTODOS

Neste capítulo, são apresentadas as principais tecnologias e ferramentas empregadas ao longo deste trabalho, bem como uma descrição detalhada dos algoritmos estudados. A Seção 3.1 aborda a linguagem de programação empregada no desenvolvimento do algoritmo proposto por Marques. A Seção 3.2 discute os algoritmos que serão comparados ao método desenvolvido pelo autor. Por fim, a Seção 3.3 descreve a metodologia experimental adotada e explica como o experimento foi conduzido.

3.1 Python

Python é uma linguagem de programação interpretada de alto nível e que suporta múltiplos paradigmas de programação: imperativo, orientado a objetos e funcional. É uma linguagem com tipagem dinâmica e gerenciamento automático de memória (PEREIRA; MOREIRA, 2020).

A linguagem *Python* foi escolhida pela facilidade de criar e manusear uma Inteligência Artificial (IA), que deixa o desenvolvimento mais fluído e de forma mais orgânica (BRAGA; GATTI, 2018).

3.2 Algoritmos Trabalhados

Neste trabalho, foram analisados e avaliados os algoritmos de Luhn (LUHN, 1957), *GistSumm* (MULLER; GRANATYR; LESSING, 2015a), o algoritmo de Programação Linear Inteira (OLIVEIRA, 2018), um algoritmo de regressão Bayesiana (SODRé; OLIVEIRA, 2019), o *ChatGPT* (RUDOLPH; TAN; TAN, 2023) e um algoritmo criado pelo pesquisador dessa monografia, todos sumarizadores baseados na extração de palavras-chave do texto.

O algoritmo de Luhn é um dos trabalhos mais importantes na área de PLN, em Sumarização de Documentos (LUHN, 1957), é um algoritmo clássico que serviu como base para muitos algoritmos, seu método de extração é baseado na extração de palavras-chave.

Os algoritmos citados anteriormente são embasados no algoritmo de Luhn, mas desenvolvidos a partir de abordagens distintas, como técnicas de regressão, inferências Bayesianas, além de mostrarem que utilizando a sentença principal do texto é mais eficiente quando se trata de gerar resumos com as ideias principais de um texto (SALVINO *et al.*, 2019).

Os algoritmos analisados neste estudo são geralmente aplicados na sumarização de textos que não se enquadram na categoria de monografias, artigos, ou pesquisas científicas. Esta escolha

se deve, em grande parte, ao fato de que textos científicos possuem uma estrutura e linguagem complexa, com informações densas e especificidades que podem representar desafios adicionais para os algoritmos de sumarização. Dessa forma, a aplicação desses algoritmos a textos não científicos tem sido a preferência na área de sumarização automática.

Além de analisar os principais desafios da área de sumarização automática de textos, o objetivo deste trabalho é a criação de um algoritmo que se destine especificamente à produção de resumos de textos acadêmicos, como monografias, artigos e pesquisas. Dessa forma, busca-se contribuir com a comunidade científica, fornecendo uma ferramenta que facilite a criação literária e pesquisa, ao mesmo tempo em que se avança no conhecimento sobre a área de sumarização automática de textos.

3.3 Projeto do Experimento

Esta seção, descreve o planejamento e execução do experimento de comparação dos algoritmos, com o objetivo de determinar qual deles melhor atende às expectativas do usuário.

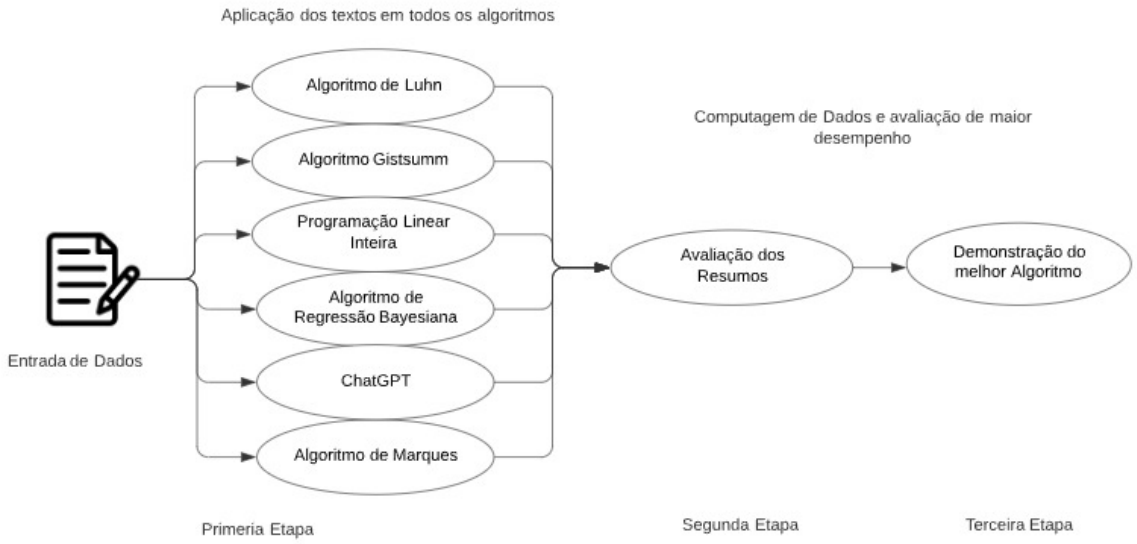
3.3.1 Instrumentação

A instrumentação dos experimentos é dividida em *hardware* e *software*. Será utilizado um computador equipado com um processador Intel Core i5 10500H (10ª geração) *Quad Core* – 12 MB de cache, 6 núcleos e 12 *threads* – com frequência de 2.50 GHz até 4.50 GHz e 16 GB de memória RAM para executar um conjunto de códigos-fonte em *Python* 3.10, que implementam os algoritmos a serem avaliados.

3.3.2 Etapas do Experimento

O experimento é composto por três etapas, conforme ilustrado na Figura 2:

Figura 2 – Etapas do Experimento



Fonte: Autoria própria.

A avaliação de desempenho dos algoritmos foi realizada com base em métricas como taxa de coerência, taxa de coesão, precisão e tempo de processamento. Essas métricas são amplamente utilizadas na literatura para avaliar a qualidade dos resumos gerados por algoritmos de sumarização automática de texto (PANDIAN, 2021). A seguir, serão apresentadas as definições e explicações detalhadas de cada métrica utilizada.

- Coerência: A taxa de coerência avalia a lógica e a clareza da estrutura do resumo gerado. Um resumo coerente deve apresentar as informações de maneira lógica, com uma sequência que faça sentido e seja fácil de seguir pelo leitor. A coerência pode variar de 0 a 1, sendo que valores mais próximos de 1 indicam maior coerência no resumo gerado.
- Coesão: A taxa de coesão mede o grau de conexão entre as ideias presentes no resumo gerado. Um resumo coeso deve apresentar ideias relacionadas de maneira integrada e harmoniosa, sem lacunas ou informações desconexas. A coesão também pode variar de 0 a 1, sendo que valores mais próximos de 1 indicam maior coesão no resumo gerado.
- Precisão: A precisão avalia a capacidade do algoritmo em selecionar as informações mais relevantes do texto original para compor o resumo. Um resumo preciso deve conter as informações essenciais e mais importantes do texto original, sem incluir informações desnecessárias ou irrelevantes. A precisão pode variar de 0 a 1, sendo que valores mais próximos de 1 indicam maior precisão na seleção de informações relevantes para o resumo.
- Tempo de processamento: O tempo de processamento é uma medida do tempo necessário

para que o algoritmo processe o texto original e gere o resumo correspondente. Essa métrica é importante para avaliar a eficiência dos algoritmos em termos de rapidez e capacidade de processar grandes volumes de dados. O tempo de processamento é geralmente expresso em segundos, sendo que menores valores indicam um algoritmo mais rápido e eficiente.

A análise dessas métricas, fundamentada em estudos prévios e na literatura científica sobre sumarização automática, possibilita uma compreensão aprofundada do desempenho dos algoritmos de sumarização automática. Essa análise fornece informações valiosas para a seleção do algoritmo mais adequado, levando em consideração as necessidades específicas dos usuários.

A Etapa 1 envolve a criação do algoritmo de Marques. Na Etapa 2, o objetivo é realizar a sumarização automática de um artigo sobre a pandemia da COVID-19 e as mudanças no estilo de vida dos brasileiros adultos (MALTA *et al.*, 2020). O texto que servirá de base para a geração dos resumos está disponível no Anexo A.1 e será processado através de seis algoritmos distintos. Por fim, na Etapa 3, os resumos serão avaliados por um grupo de avaliadores, que julgarão a qualidade dos resumos gerados por cada algoritmo, a fim de determinar o desempenho de cada um deles.

O cenário considerado para os experimentos envolve a simulação de resumos em uma máquina virtual com capacidade de processamento de um Intel Core i5 10500H (10ª geração) *Quad Core* – 12 MB de cache, 6 núcleos e 12 *threads* – com frequência de 2.50 GHz até 4.50 GHz e 16 GB de memória RAM. A análise das sumarizações obtidas será realizada dentro de um contexto educacional.

3.3.2.1 Etapa 1: Sumarização Automática através de Algoritmos

A primeira etapa do experimento, consiste em realizar a sumarização automática usando os algoritmos de Luhn, *Gistsumm*, PLI (OLIVEIRA, 2018), regressão Bayesiana, *ChatGPT* e o Algoritmo de Marques.

Foram realizados testes nas configurações dos parâmetros comuns modificáveis nos algoritmos, isto é, na quantidade de sentenças importantes, a fim de equipará-los para a etapa seguinte. Dessa forma, neste experimento foi utilizado o parâmetro de **cinco (5) sentenças importantes**.

Segundo Rino e Pardo (2003), para gerar um resumo eficiente, deve-se extrair entre 20 a 50% do texto e atingir uma taxa de compressão de 80%. Neste trabalho, seguimos essa orientação ao avaliar os algoritmos de sumarização automática.

Considerando os parâmetros mencionados, o texto utilizado no resumo seguiu uma taxa de extração de aproximadamente 40% para cada algoritmo, com aproximadamente 80% de compreensão. O algoritmo de Marques, por exemplo, selecionou as cinco sentenças mais importantes do texto original, que representam uma porcentagem significativa das informações relevantes. Essas cinco sentenças foram determinadas com base na frequência das palavras e na relevância das sentenças no contexto do texto.

As cinco sentenças importantes servem como um critério para avaliar a qualidade dos resumos gerados pelos algoritmos de sumarização automática. A escolha dessas sentenças permite obter um resumo conciso e informativo, mantendo a essência do texto original e assegurando que o leitor compreenda os principais pontos abordados. Esse critério contribui para a eficiência dos resumos gerados e auxilia na comparação dos algoritmos em relação à sua capacidade de extrair informações relevantes e concisas do texto original (OLIVEIRA *et al.*, 2022).

3.3.2.2 Etapa 2: Avaliação dos Resumos Gerados

Para avaliar a qualidade e coesão dos resumos gerados na segunda etapa, foram coletadas 49 respostas de indivíduos com diferentes graus de instrução acadêmica que responderam a questionários aleatórios. Essas pessoas avaliaram os resumos resultantes da inserção do artigo de Malta *et al.* (2020) distribuídos entre 6 resumos e 5 questionários, comparando cada algoritmo mencionado na segunda etapa com o algoritmo de Marques.

Os avaliadores julgaram qual texto apresentava maior coesão e coerência, sem saber quais algoritmos foram utilizados. Em todos os casos, o texto 1 representa o resumo proveniente do algoritmo de Marques, enquanto o texto 2 em cada questionário representa os algoritmos comparados: Luhn, *Gistsumm*, PLI (OLIVEIRA, 2018), regressão Bayesiana e *ChatGPT*.

Os questionários foram enviados por meio de formulários no *Google Docs*, conforme modelo apresentado no Apêndice A.

3.3.2.3 Etapa 3: Análise Qualitativa para Identificação do Melhor Algoritmo

A terceira etapa se concentra na análise e comparação dos resultados dos resumos produzidos pelos diferentes algoritmos de sumarização. Esta análise qualitativa tem o propósito de identificar o algoritmo mais eficaz na produção de resumos de alta qualidade.

Esta etapa envolve a aplicação de métricas para avaliar a qualidade, a relevância das informações retidas, a fluidez e coesão dos textos resumidos gerados na Etapa 1. Os resultados

são comparados e o algoritmo que produz o resumo com a melhor representação do texto original é considerado o mais bem-sucedido.

A avaliação se baseia tanto na análise objetiva dos resumos quanto nas respostas dos avaliadores obtidas na Etapa 2. Esta etapa resulta na seleção do melhor algoritmo de sumarização, com base na qualidade do resumo e na quantidade de informações essenciais retidas do texto original. A seleção do algoritmo mais eficiente contribui significativamente para os avanços na área de Processamento de Linguagem Natural.

3.3.3 Amostragem e Seleção dos Participantes

A escolha dos participantes para a avaliação dos algoritmos de sumarização automática neste trabalho foi feita através de uma amostragem por conveniência. A amostragem por conveniência é uma técnica não probabilística que seleciona os participantes com base na disponibilidade e na facilidade de acesso (NEUMAN, 2006).

Os questionários foram preenchidos pelos participantes após a divulgação dos formulários pelas redes sociais do autor e do orientador. Esse método de seleção permitiu alcançar um grupo diversificado de participantes, embora não seja uma amostra totalmente representativa da população em geral.

É importante mencionar que a amostragem por conveniência pode ter sido influenciada pelo efeito *snowball* (GOODMAN, 1961). O efeito *snowball* ocorre quando os participantes iniciais compartilham o questionário com seus contatos, que por sua vez compartilham com seus próprios contatos, aumentando o alcance da pesquisa. Esse efeito pode introduzir viés na amostra, uma vez que as pessoas que participam da pesquisa tendem a ter características semelhantes ou estarem relacionadas de alguma forma.

Apesar dessas limitações, a amostragem por conveniência e o possível efeito *snowball* permitiram obter um número suficiente de respostas para a avaliação dos algoritmos de sumarização automática, fornecendo informações valiosas sobre o desempenho dos algoritmos em relação às métricas estabelecidas.

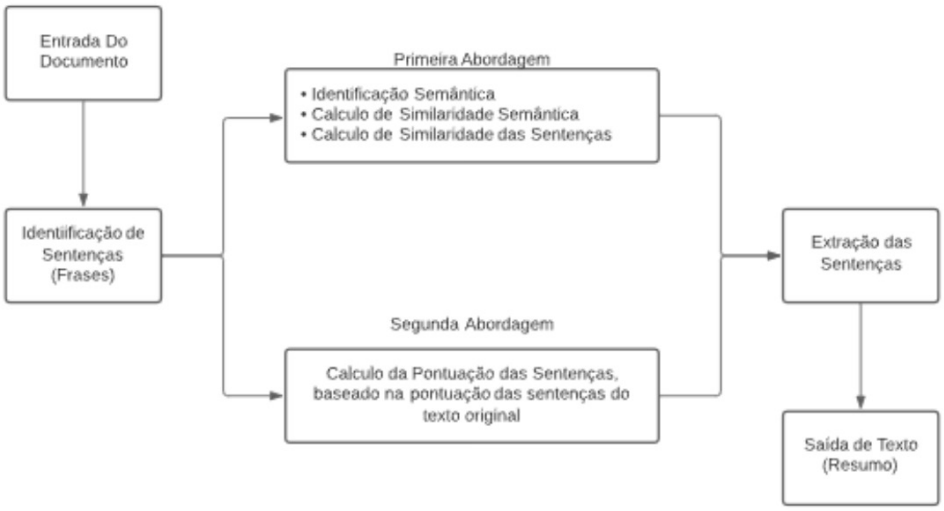
Após a coleta das respostas, os resultados foram processados e analisados com base em métricas específicas de qualidade de sumarização, como precisão, coerência e coesão, além do tempo de processamento. Essa análise permitiu determinar qual algoritmo apresentou melhor desempenho na geração de resumos coerentes e coesos. Os algoritmos utilizados na comparação e seus respectivos métodos de sumarização serão detalhados no capítulo 4.

4 ALGORITMOS DE SUMARIZAÇÃO DE TEXTO

Este capítulo, explica os algoritmos utilizados nessa monografia que, por meio de uma pesquisa, comparará cada um destes algoritmos consolidados com o proposto pelo autor do presente trabalho.

4.1 Algoritmo de Luhn

Figura 3 – Modelo de Funcionamento do Algoritmo de Luhn



Fonte: Adaptado de (RUSSELL, 2011).

O algoritmo de Luhn analisa as frases mais importantes de um documento, que são aquelas que mais aparecem no texto. Neste contexto, não são consideradas as *stopwords*, que são palavras como artigos, preposições, conjunções, entre outras que aparecem com frequência em um texto, porém são insignificantes em relação ao significado semântico do documento (ROCHA, 2022). Em suma, as *stopwords* são utilizadas apenas para dar um sentido gramatical correto na formação das frases. Como o algoritmo faz a sumarização baseada na frequência que as palavras ocorrem, elas são desconsideradas para não confundir o sumarizador.

O algoritmo não procura compreender os dados em um nível semântico, e simplesmente computa resumos com agrupamento de palavras que ocorrem com frequência no texto. A Figura 3 apresenta os passos desde quando o algoritmo de Luhn recebe um texto como parâmetro até a geração final do resumo. A primeira tarefa é identificar as frases, calculando a similaridade entre todas as frases do texto. Na segunda abordagem, outra tarefa é fazer o cálculo da pontuação, levando em conta que o resumo nunca pode ter mais pontos e vírgulas que o texto original.

Terminadas a primeira e segunda abordagens mostradas na Figura 3, o algoritmo finalmente extrai as sentenças escolhidas, as agrupa e mostra como resultado o resumo. O resumo gerado pelo algoritmo de Luhn está no anexo A.2

O algoritmo de Regressão Bayesiana é baseado no Algoritmo de Luhn, compartilhando muitas das suas características principais, como a identificação de sentenças importantes com base na frequência de palavras. No entanto, existem diferenças significativas na implementação e abordagem de cada um.

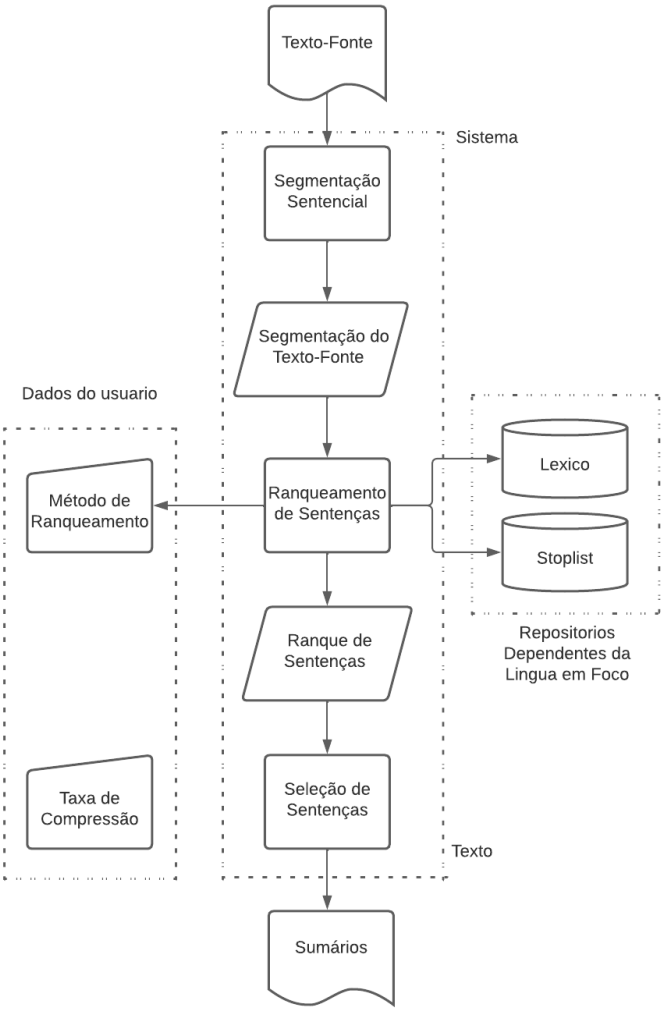
A Regressão Bayesiana, como sugere o nome, utiliza o conceito de inferência Bayesiana para estimar a relevância de cada sentença. Isso significa que, em vez de apenas contar a frequência das palavras, como no Algoritmo de Luhn, a Regressão Bayesiana também leva em conta a probabilidade de uma sentença ser importante, dado o restante do texto. Isso permite que o algoritmo identifique melhor as nuances e o contexto do texto.

Por outro lado, o Algoritmo de Luhn se baseia estritamente na frequência de palavras, sem considerar o contexto mais amplo. Isso torna o Algoritmo de Luhn mais simples, mas potencialmente menos preciso em alguns casos.

Portanto, embora ambos os algoritmos compartilhem uma abordagem comum à sumarização de texto, eles possuem diferenças fundamentais que podem afetar a qualidade e a precisão dos resumos gerados. O resumo gerado pelo algoritmo de Regressão Bayesiana está no anexo A.5.

4.2 Algoritmo Gistsumm

Figura 4 – Diagrama de Blocos Gistsumm



Fonte: Adaptado de (MILLER, 2015).

O *GistSumm* (*GistSumarizer*) é um sumarizador extrativo que usa técnicas estatísticas para determinar a ideia central dos textos por ele sumarizados. Baseia-se na simulação da sumarização humana, primeiro identificando a ideia principal do texto e, então, acrescenta informações adicionais ou complementares (BREWKA, 1996). Essas informações adicionais podem ser a segunda ou terceira frase mais importante do texto, seguindo em ordem crescente de acordo com a quantidade de frases que se deseja extrair do texto.

Dessa forma, o sumarizador primeiro procura a sentença que melhor expressa a ideia principal do texto e baseado nela são escolhidas as demais sentenças, que vão compor o extrato textual. Mesmo quando a sentença escolhida não for a sentença principal e há uma aproximação

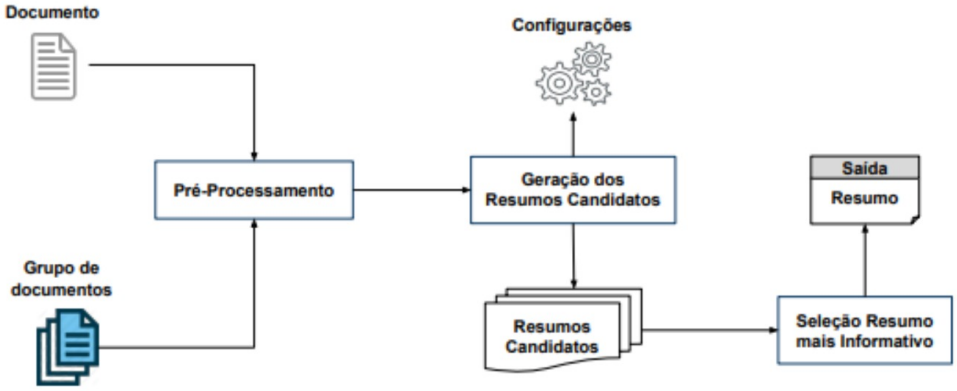
significativa da mesma, o extrato já pode ser gerado (TORRES-MORENO, 2014).

O *GistSumm* compreende três processos principais, e mais alguns secundários, os quais são descritos a seguir: segmentação textual, ranqueamento de sentenças e seleção de sentenças.

A segmentação textual delimita as sentenças do texto-fonte e procura pelos sinais de pontuação. O ranqueamento é uma ordenação a partir de pesos obtidos na aplicação de métodos estatísticos, sendo feita a análise léxica, extração das *stopwords* e aplicação do método de ranqueamento. Por fim, a seleção de sentença escolhe as sentenças mais relevantes, por meio de seus métodos extrativos, para, deste modo, gerar o sumário do documento analisado (MULLER; GRANATYR; LESSING, 2015b). Neste momento, o texto é transformado de acordo com a taxa de compressão definida. A taxa de compressão é a porcentagem do texto original que o algoritmo pode extrair para o resumo. O resumo gerado pelo algoritmo *Gistsumm* está no anexo A.3.

4.3 Algoritmo de Programação Linear Inteira (PLI)

Figura 5 – Diagrama de Blocos Programação Linear Inteira



Fonte: Extraido de (OLIVEIRA, 2018).

O Algoritmo de programação Linear Inteira tem como objetivo a sumarização automática de textos baseada em conceitos, utilizando programação linear inteira e regressão. Essa abordagem busca extrair as informações mais importantes de um texto e gerar um resumo coerente e relevante. O PLI segue as etapas: pré-processamento, geração dos resumos candidatos, seleção do resumo.

Na etapa de pré-processamento, o texto é pré-processado para eliminar ruídos e facilitar a extração de informações. O pré-processamento inclui a remoção de pontuação, números, caracteres especiais e *stopwords*, além da normalização do texto, como a conversão de todas as letras para minúsculas e a redução das palavras ao seu radical (*stemming*).

Com base no texto pré-processado, os conceitos são extraídos. Um conceito é uma palavra ou expressão que representa uma ideia importante no texto. A frequência de cada conceito no texto é calculada, e os conceitos são ordenados de acordo com sua importância, considerando suas frequências e pesos semânticos.

Na geração do resumos, as sentenças que contêm os conceitos mais importantes são selecionadas para compor o resumo. O algoritmo utiliza programação linear inteira para determinar a melhor combinação de sentenças que maximiza a cobertura dos conceitos importantes, mantendo a coerência e a concisão do resumo. A programação linear inteira é uma técnica matemática que permite a resolução de problemas de otimização com variáveis inteiras.

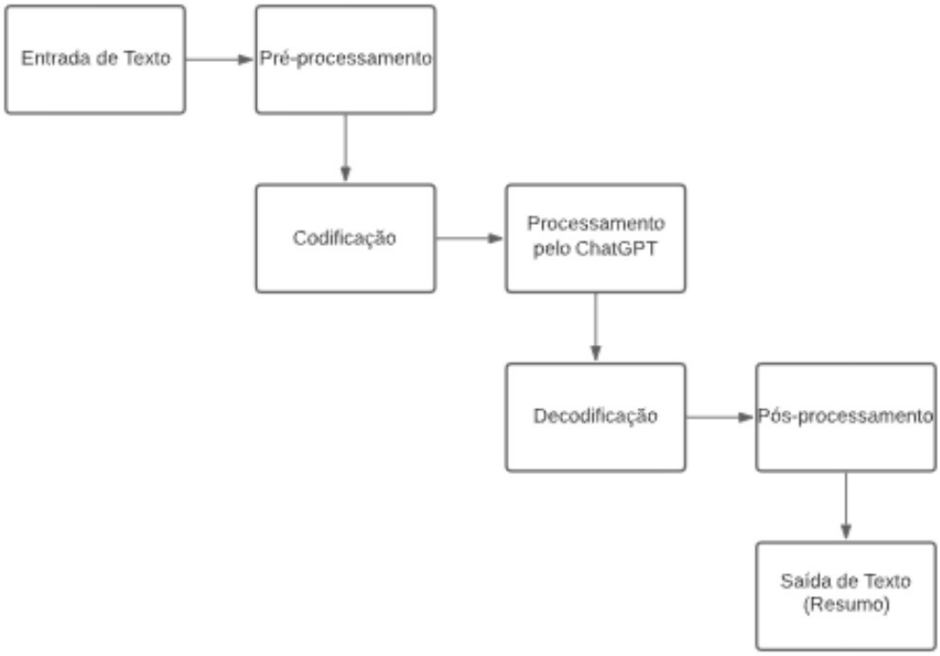
A regressão é aplicada para ajustar os pesos dos conceitos e das sentenças, de forma a melhorar a qualidade do resumo gerado. O algoritmo utiliza um conjunto de treinamento com textos e resumos previamente avaliados por humanos para aprender a importância relativa dos conceitos e sentenças no processo de sumarização. A regressão permite que o algoritmo ajuste seus parâmetros de acordo com os padrões observados nos dados de treinamento, melhorando a precisão e a relevância dos resumos gerados.

Com base na seleção de sentenças e nos pesos ajustados dos conceitos, o resumo final é gerado. As sentenças selecionadas são reorganizadas de acordo com a ordem em que aparecem no texto original, garantindo a coerência e a fluidez do resumo.

O PLI apresenta uma abordagem interessante para a sumarização automática de textos, combinando técnicas de programação linear inteira e regressão para extrair e ponderar conceitos importantes e selecionar as sentenças mais relevantes para compor o resumo. Essa abordagem busca gerar resumos mais precisos, coerentes e informativos, levando em consideração a importância dos conceitos e a estrutura do texto original. O resumo gerado está no anexo A.4.

4.4 *ChatGPT*

Figura 6 – Diagrama de Blocos ChatGPT



Fonte: Adaptado do ChatGPT.

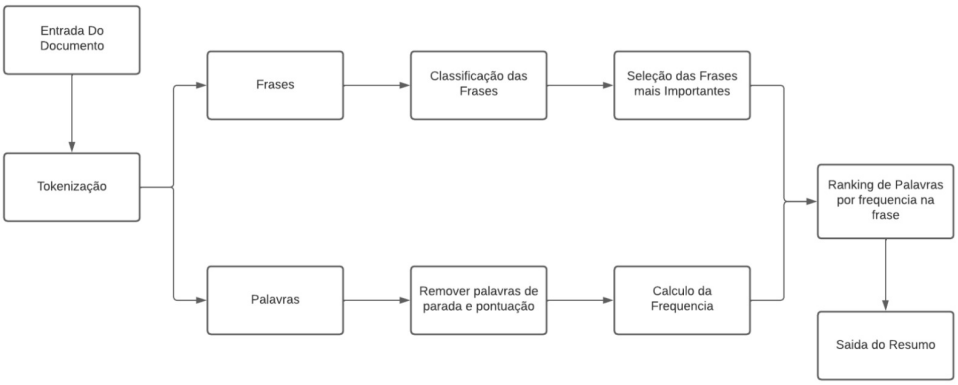
O *ChatGPT* é um algoritmo baseado em inteligência artificial, criado pela *OpenAI* (LUND; WANG, 2023). Seu nome vem é uma sigla de *Generative Pre-Trained Transformer*, que significa, em tradução livre, Transformador pré-treinado generativo (TRANSFORMER; ZHAVORONKOV, 2022). Esse algoritmo tem seu desenvolvimento traçado a partir de redes neurais e *Machine Learning*, e tem foco semelhante a *chatbots*, com conversação online. A ideia a partir do que ele foi criado, é para aprimorar a experiência e recursos oferecidos por alguns assistentes virtuais, como Alexa, Google Assistente, dentre outros. Grande parte de ter se tornado famoso e ter tanto sucesso se dá pela forma simples de conversação e obtenção de respostas (AYDIN; KARAARSLAN, 2023).

A sua arquitetura se baseia em uma rede neural chamada *Transformer*, que é projetada especialmente para lidar com textos. Seu modelo de várias camadas permite que a plataforma consiga identificar palavras-chave, nos mais diferentes contextos inseridos (NATARAJAN *et al.*, 2020). Esse algoritmo se alimenta de informações que colhe da Internet. Nesse caso, tudo que existe hoje disponível na rede pode ser usado como base para essa ferramenta (RAO *et al.*, 2023). Seguindo alguns padrões e no cruzamento de informações, o *ChatGPT* transforma as *querys*, perguntas feitas pelo usuário, em respostas. Seu diferencial se dá na criatividade que

está presente nessas respostas, já que, diferente dos métodos tradicionais de busca, ele traz a resposta contextualizada e em um texto elaborado que busca o maior entendimento do leitor, além da possibilidade de elaborar musicas, poesias, códigos de programação, receitas, dentre outras (RUDOLPH; TAN; TAN, 2023). O resumo gerado pelo *ChatGPT* está no anexo A.6.

4.5 Algoritmo de Marques

Figura 7 – Diagrama de Blocos Algoritmo de Marques



Fonte: Autoria própria.

O algoritmo proposto pelo autor tem embasamento no algoritmo *GistSumm* (*GistSummarizer*), que é embasado no algoritmo de Luhn, e seu método de extração de palavras-chave baseado na sentença principal do texto é muito eficiente quando trata-se de gerar resumos com as ideias principais de um texto (NETO, 2022). De acordo com Rino e Pardo (2003), o *GistSumm* atualmente encontra-se como o estado da arte de sumarização automática de documentos, com uma função para sumarização multi-documentos.

Segundo o trabalho de Oliveira e Guelpeli (2011), a eficiência de um algoritmo de sumarização está ligada ao desempenho de seu método de extração de palavras-chave. Para verificar essa hipótese, serão utilizadas tabelas com os resultados, que mostram todos os dados tanto em forma numérica e percentual, onde pode-se perceber a eficiência e demais características dos algoritmos, tais como taxa de erros e acertos. Avaliações de forma semelhante a essas foram feitas nos trabalhos de Luhn e Rino e Pardo (2003), sendo que somente o trabalho de Rino e Pardo (2003) aplica-se ao português do Brasil.

O algoritmo de Marques é um sumarizador extrativo que usa técnicas estatísticas para determinar a ideia central dos textos por ele sumarizados, que implementa bibliotecas da linguagem *Python* 3.10. Baseia-se na simulação da sumarização humana, primeiro identificando a

ideia principal do texto e, então, acrescenta informações adicionais ou complementares. Essas informações adicionais podem ser a segunda ou terceira frase mais importante do texto, seguindo em ordem crescente de acordo com a quantidade de frases que se deseja extrair do texto.

Dessa forma, o sumarizador primeiro procura a sentença que melhor expressa a ideia principal do texto, e baseado nela são escolhidas as demais sentenças que vão compor o extrato textual. O *GistSumm* trabalha da mesma forma: primeiro o GistSumm realiza a identificação da sentença principal com o uso de métodos estatísticos simples, e, por segundo, conhecendo-se as sentenças principais é possível produzir extratos coerentes. Mesmo quando a sentença escolhida não for a sentença principal e há uma aproximação significativa da mesma, o extrato já pode ser gerado. O código do algoritmo de Marques está no apêndice B.1. O resumo gerado pelo algoritmo de Marques está no apêndice B.2.

O algoritmo de Marques apresentado neste trabalho foi comparado com outros algoritmos de sumarização automática de texto presentes na literatura, citados acima. A Tabela 1 apresenta as principais diferenças entre o algoritmo de Marques e os algoritmos relacionados.

O algoritmo de Marques se destaca em relação aos algoritmos relacionados por utilizar um modelo de redes neurais treinado especificamente para a tarefa de sumarização automática de textos.

Tabela 1 – Comparação entre os algoritmos

Algoritmo	**Características**
GistSumm	Baseado em técnicas de extração
Luhn	Utiliza frequência de palavras e posição no texto
PLI	Baseado em programação linear inteira
Regressão Bayesiana	Abordagem baseada em aprendizado de máquina
ChatGPT	Geração de resumos por modelo pré-treinado de linguagem
Marques	Redes neurais treinadas especificamente para sumarização automática de textos

Fonte: Autoria própria.

5 RESULTADOS E DISCUSSÕES

Este capítulo, apresenta os resultados obtidos na análise comparativa dos seis algoritmos de sumarização automática de texto descritos no Capítulo 4. A metodologia utilizada consistiu na aplicação dos algoritmos em um texto sobre a COVID-19 e seus impactos na pandemia (MALTA *et al.*, 2020), e na avaliação dos resultados com base em métricas de qualidade de sumarização.

5.1 Comparativo dos Algoritmos

Esta seção, apresenta os resultados obtidos para cada um dos algoritmos, sendo comparados com o algoritmo de Marques. Foram obtidas 49 respostas ao total. É importante enfatizar que cada um dos respondentes é único e que os 49 participantes não responderam a todos os formulários. A distribuição das respostas para cada formulário é detalhada em sua respectiva seção.

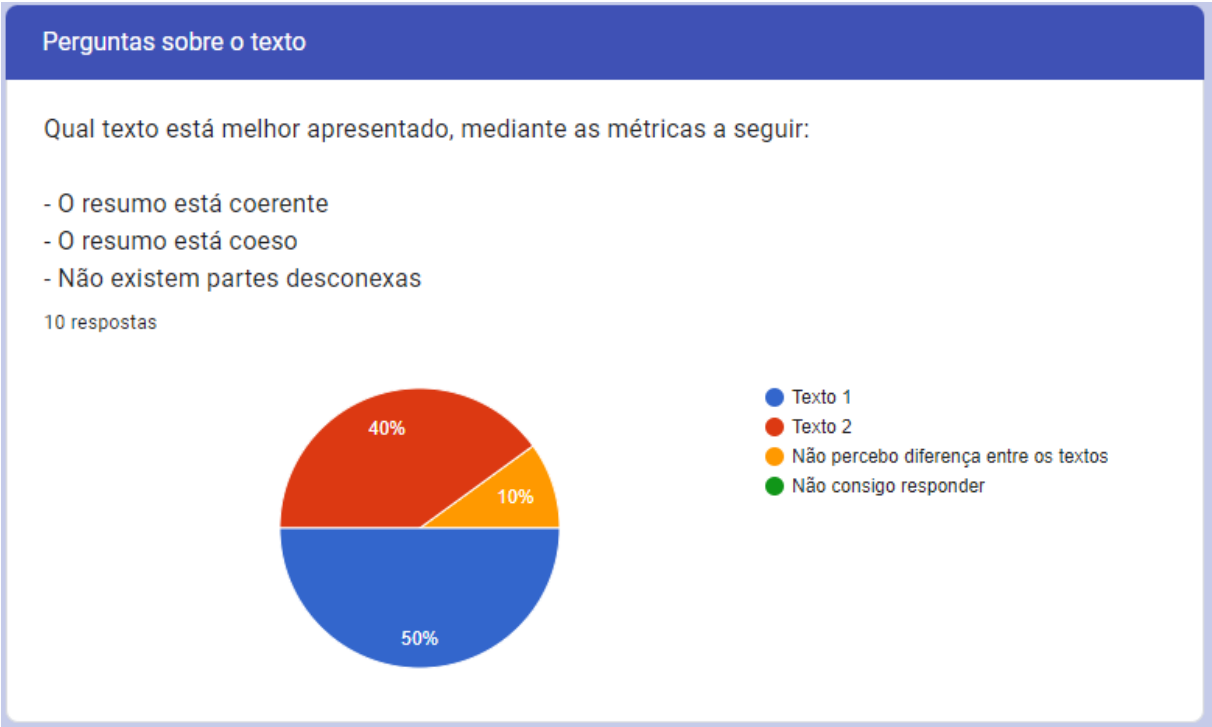
A fim de organização, nos gráficos, o texto **um** sempre será o algoritmo de Marques, e o texto **dois** o algoritmo daquela seção.

5.1.1 Algoritmo de Luhn

O formulário, que requeria que os participantes avaliassem os resumos gerados pelos algoritmos de Luhn e Marques, recebeu **10** respostas. Embora o algoritmo de Luhn tenha identificado algumas das principais ideias dos textos analisados e recebido quatro votos, seu desempenho foi considerado razoável em comparação com o algoritmo de Marques, que obteve cinco votos, conforme visto na Figura 8. O resultado era esperado, já que o algoritmo de Marques é baseado no algoritmo de Luhn, considerado o pai dos algoritmos de sumarização automática (SHASHIKANTH; SANGHAVI, 2019).

Com base na análise de conteúdo (CASSIANI; CALIRI; PELá, 1996) como abordagem da pesquisa interpretativa, as respostas do formulário indicam que o algoritmo de Marques foi considerado melhor em alguns casos devido ao fato de que o resumo gerado por ele apresenta uma introdução mais completa e contextual, tornando o texto mais claro e fácil de entender. Além disso, os participantes mencionaram que o texto gerado pelo algoritmo de Marques possui uma estrutura organizada e clara, enquanto o texto gerado pelo algoritmo de Luhn apresenta falta de coesão e informações desnecessárias, o que compromete a qualidade do resumo.

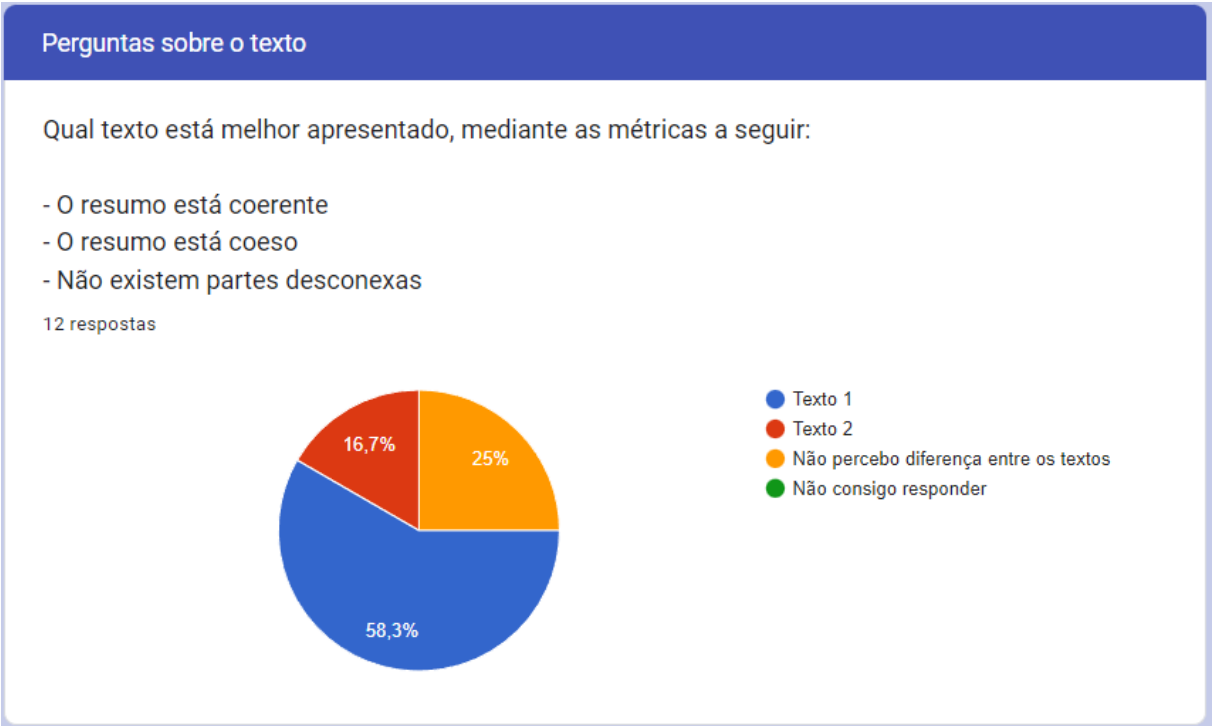
Figura 8 – Respostas ao Questionário sobre Marques x Luhn



Fonte: Formulário do *Google Docs*

5.1.2 Algoritmo Gistsumm

Figura 9 – Respostas ao Questionário sobre Marques x *Gistsumm*



Fonte: Formulário do *Google Docs*

Esse formulário recebeu **12** respostas, onde o algoritmo de *Gistsumm* recebeu cinco votos indicando que era melhor, enquanto que o de Marques obteve sete votos nesse sentido, conforme visto na Figura 9. Embora o algoritmo de *Gistsumm* tenha apresentado um desempenho satisfatório na sumarização dos textos analisados, conseguindo identificar com precisão as principais ideias presentes nos documentos, o algoritmo de Marques se destacou por sua capacidade de selecionar informações mais relevantes e produzir resumos mais coerentes e bem estruturados, especialmente em textos mais extensos. Isso se deve à utilização de técnicas estatísticas de seleção de sentenças relevantes implementadas pelo algoritmo de Marques.

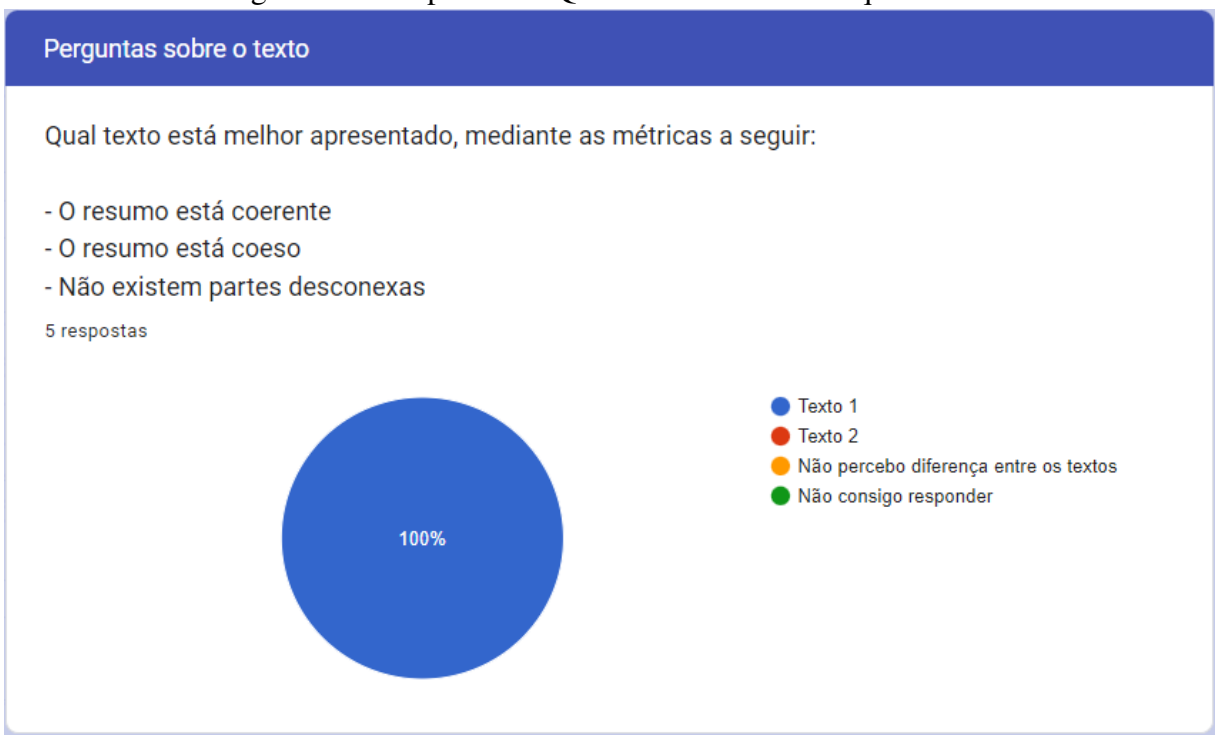
Com base na teoria fundamentada nos dados como abordagem da pesquisa interpretativa, as respostas do formulário revelam diferentes aspectos que levaram os participantes a considerar o algoritmo de Marques melhor em alguns casos. As justificativas apontam para a capacidade do algoritmo de Marques de gerar resumos mais detalhados, alinhados com os acontecimentos no contexto da pandemia de COVID-19 no Brasil e que abordam de maneira mais ampla os diversos aspectos relacionados à pandemia, incluindo medidas preventivas, impactos na atividade física e comportamento sedentário. Os participantes também destacaram que o texto gerado pelo algoritmo de Marques apresenta uma estrutura mais clara e encadeada das informações, tornando a compreensão do conteúdo mais fácil e eficiente.

5.1.3 Algoritmo de Programação Linear Inteira

Para esta comparação, foram obtidas apenas **cinco** respostas. O Algoritmo de Programação Linear Inteira apresentou um desempenho abaixo do esperado na sumarização dos textos analisados, com um resumo não tão claro, em que as respostas a esse formulário mostram que tinha um conteúdo mais genérico. Como pode ser visto na Figura 10, todos os usuários escolheram o resumo gerado pelo algoritmo de Marques. A capacidade de compreensão semântica das palavras e das relações entre elas permitiu ao algoritmo produzir um resumo de fácil leitura e compreensão fluida.

Considerando a teoria fundamentada nos dados como abordagem da pesquisa interpretativa, as respostas do formulário indicam que o algoritmo de Marques foi considerado melhor em alguns casos devido à sua capacidade de estabelecer um contexto claro sobre a pandemia da COVID-19 e o envolvimento da Organização Mundial da Saúde (OMS). Isso permite ao leitor compreender rapidamente a situação e a relevância das medidas adotadas. Além disso, os participantes destacaram que o texto gerado pelo algoritmo de Marques é mais completo, apresenta mais

Figura 10 – Respostas ao Questionário sobre Marques x PLI



Fonte: Formulário do *Google Docs*

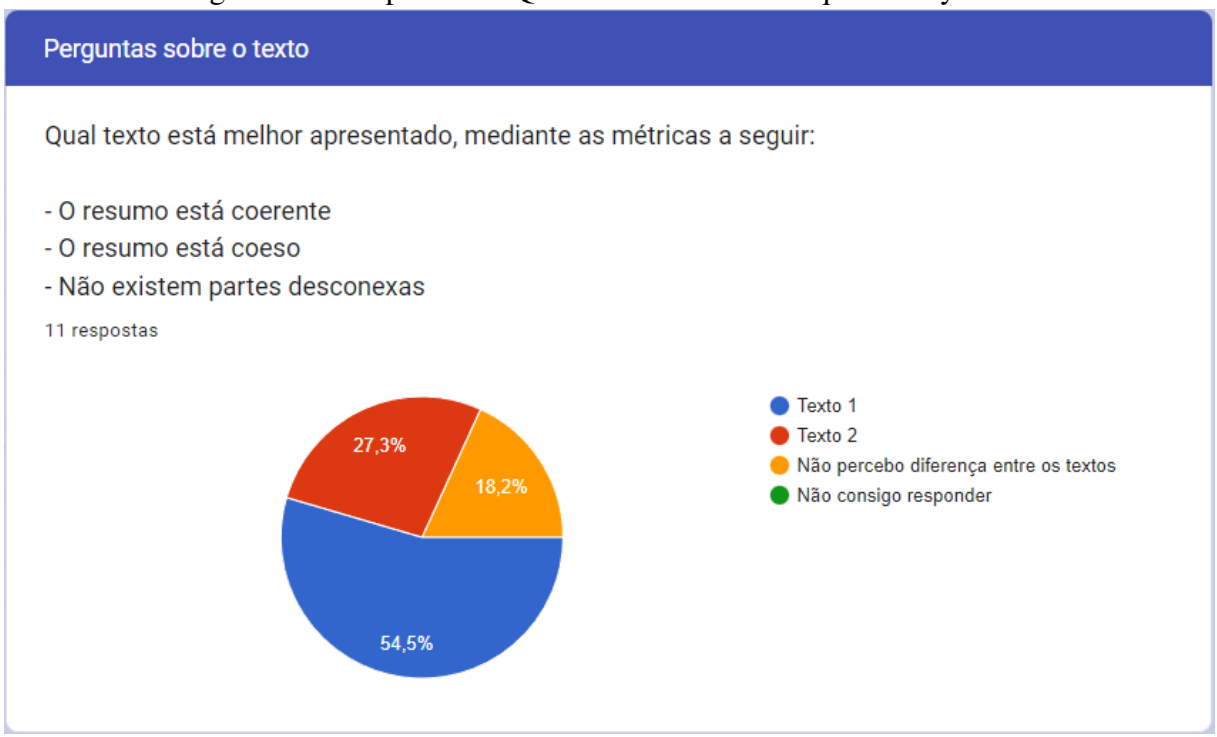
informações e está melhor dividido, facilitando a compreensão da linguagem e proporcionando maior contextualização mundial das medidas adotadas e suas possíveis consequências.

5.1.4 Algoritmo de Regressão Bayesiana

O Algoritmo de Regressão Bayesiana também apresentou um desempenho abaixo do esperado na sumarização dos textos analisados, com um resumo não tão claro, onde as respostas a esse formulário, que obteve **11** respostas, mostram que o Algoritmo de Regressão Bayesiana tinha um conteúdo mais genérico. Como pode ser visto na Figura 11, a maioria dos usuários escolheram o resumo gerado pelo algoritmo de Marques. A capacidade de compreensão semântica das palavras e das relações entre elas permitiu ao algoritmo produzir um resumo de fácil leitura e compreensão fluida.

Com base na teoria fundamentada nos dados como abordagem da pesquisa interpretativa, as respostas do formulário sugerem que o algoritmo de Marques foi considerado melhor em alguns casos, pois estabelece um contexto claro sobre a pandemia da COVID-19 e o envolvimento da Organização Mundial da Saúde (OMS). Isso permite ao leitor compreender rapidamente a situação e a relevância das medidas adotadas. Além disso, os participantes destacaram que o texto gerado pelo algoritmo de Marques é mais completo, apresenta mais informações e está

Figura 11 – Respostas ao Questionário sobre Marques x Bayesiana



Fonte: Formulário do *Google Docs*

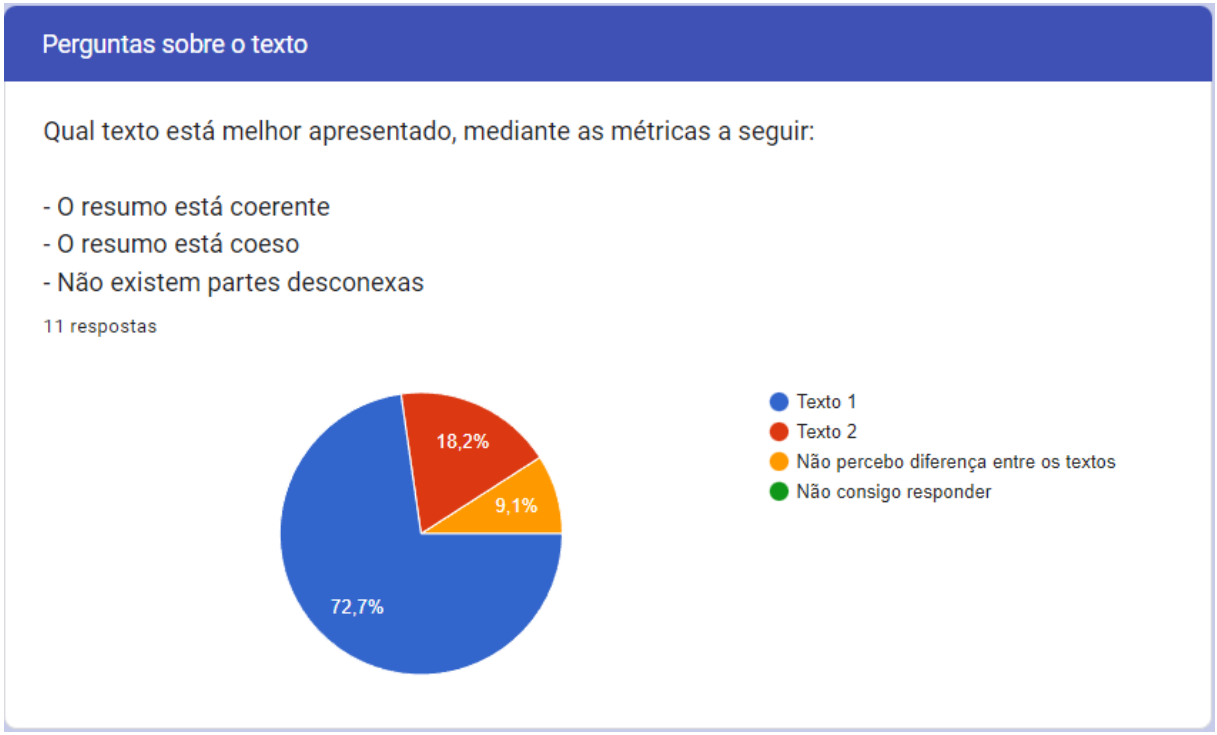
melhor dividido, tornando a leitura mais rápida e dinâmica e fornecendo uma visão abrangente da pandemia, incluindo o reconhecimento pela OMS e a importância das intervenções não farmacológicas recomendadas.

5.1.5 ChatGPT

O ChatGPT apresentou um desempenho um tanto abaixo na sumarização dos textos analisados, gerando resumos com pouca informação. Até mesmo a justificativa que afirmou que o resumo do ChatGPT estava mais adequado ao artigo, não o considerou melhor em termos de qualidade. No entanto, como pode ser visto na Figura 12, a maioria dos usuários que responderam esse formulário escolheu o resumo gerado pelo algoritmo de Marques. Isso se deve em parte à forma como o resumo foi apresentado, de maneira mais completa e adequada ao conteúdo do artigo. A capacidade de compreensão semântica das palavras e das relações entre elas permitiu ao algoritmo de Marques produzir resumos com informações precisas e completas.

Com base na teoria fundamentada nos dados como abordagem da pesquisa interpretativa, as respostas do formulário indicam que o algoritmo de Marques foi considerado melhor em alguns casos porque, apesar de ambos os textos estarem bem escritos, o texto gerado pelo algoritmo de Marques apresenta um foco ligeiramente diferente, conseguindo abranger de modo

Figura 12 – Respostas ao Questionário sobre Marques x ChatGPT



Fonte: Formulário do *Google Docs*

resumido mais tópicos de forma coesa. Os participantes também destacaram que o texto gerado por esse algoritmo é mais completo em relação à coerência e tecnicidade, expondo melhor os dados e explicações sobre o tema, mesmo que ambos sejam linguisticamente acessíveis ao público em geral.

5.2 Desafios na Sumarização Automática de Texto

A sumarização automática de texto é uma área em constante evolução, mas ainda enfrenta diversos desafios que precisam ser superados para que os algoritmos de sumarização sejam capazes de produzir resumos realmente úteis e relevantes para os usuários. Dentre os principais desafios, destacam-se:

- Adaptar os algoritmos para diferentes tipos de textos e domínios de conhecimento, garantindo que os resumos gerados sejam relevantes para os usuários em diferentes contextos;
- Avaliar os resumos gerados de forma mais rigorosa e consistente, considerando as expectativas e necessidades dos usuários;
- Lidar com a ambiguidade e a complexidade da linguagem natural, garantindo que os resumos mantenham a informação correta e essencial do texto original;

- Desenvolver técnicas de sumarização que levem em consideração as preferências e o conhecimento prévio dos usuários, gerando resumos personalizados de acordo com as necessidades individuais;
- Aprimorar o desempenho dos algoritmos em relação ao tempo de processamento e recursos computacionais, tornando-os mais eficientes e escaláveis.

Cada um desses desafios requer uma abordagem específica e uma pesquisa contínua para que os algoritmos de sumarização possam ser aprimorados e se tornarem mais eficazes.

5.3 Avanços e Perspectivas Futuras

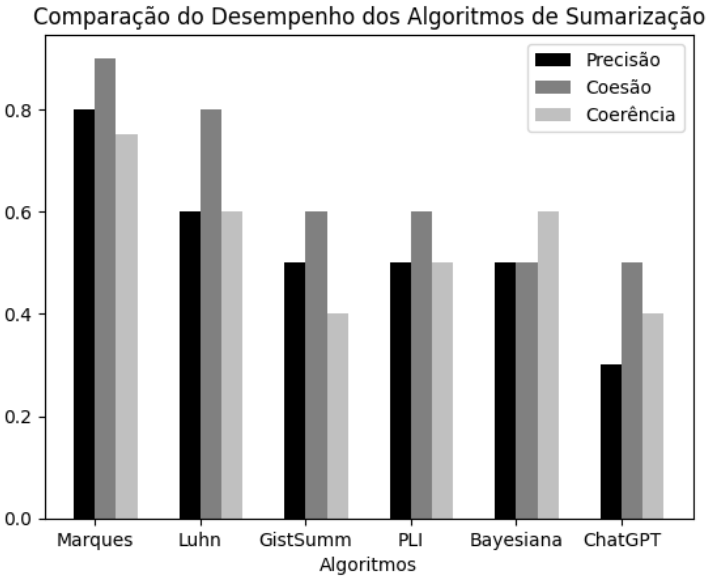
Apesar dos desafios enfrentados na área de sumarização automática de texto, existem muitos avanços e perspectivas futuras promissoras. Entre os avanços recentes, destacam-se:

- O uso de redes neurais e aprendizado de máquina para melhorar a qualidade da sumarização automática, permitindo que os algoritmos possam aprender a identificar as informações mais importantes dos textos de forma mais precisa e eficaz;
- A utilização de técnicas de processamento de linguagem natural mais avançadas, como o reconhecimento de entidades nomeadas e a análise semântica, para melhorar a qualidade da sumarização e garantir que os resumos gerados sejam mais precisos e relevantes para os usuários;
- A aplicação de algoritmos de sumarização em áreas específicas, como a saúde e o direito, para fornecer resumos mais precisos e úteis para profissionais dessas áreas;
- O desenvolvimento de técnicas de sumarização personalizadas, que levam em consideração as preferências e o conhecimento prévio dos usuários, permitindo que os resumos gerados sejam mais relevantes e úteis para cada indivíduo.

Esses avanços abrem caminho para uma pesquisa contínua e um desenvolvimento cada vez maior na área de sumarização automática de texto.

5.4 Discussão dos Resultados

Figura 13 – Comparação de Desempenho dos seis algoritmos de sumarização

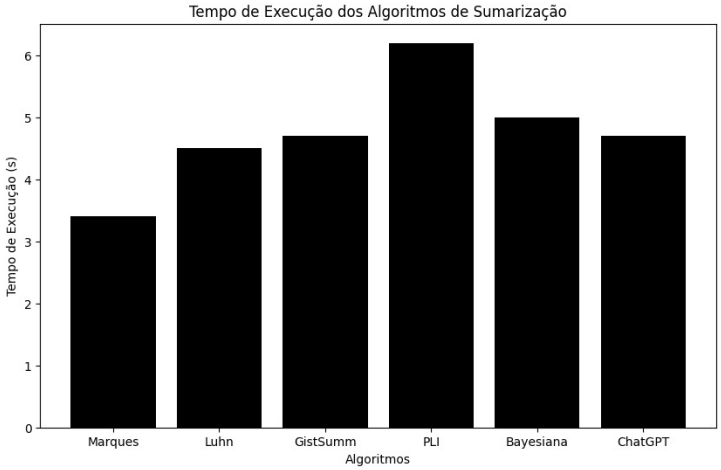


Fonte: Autoria própria.

Apesar de todos os cinco algoritmos de sumarização automática de texto apresentarem um desempenho aceitável na identificação das ideias-chave dos textos analisados, a avaliação sugere que o algoritmo de Marques se destaca por sua superioridade em aspectos como coesão, coerência, precisão e tempo de processamento. Quando comparado diretamente com cada um dos outros algoritmos, o algoritmo de Marques conquistou metade das respostas avaliadas quando comparado ao algoritmo de Luhn, metade ao ser comparado ao *GistSumm*, todas as respostas quando em comparação ao PLI, a maioria das respostas quando comparado ao *ChatGPT*, e metade das respostas quando comparado ao Algoritmo de Regressão Bayesiana.

Aqui, "desempenho aceitável"significa que todos os algoritmos foram capazes de sumarizar efetivamente os textos, identificando as ideias principais e produzindo resumos que mantinham o sentido geral dos textos originais. No entanto, alguns algoritmos se saíram melhor que outros em determinados aspectos, como coesão, coerência e precisão.

Figura 14 – Tempo de execução dos Algoritmos

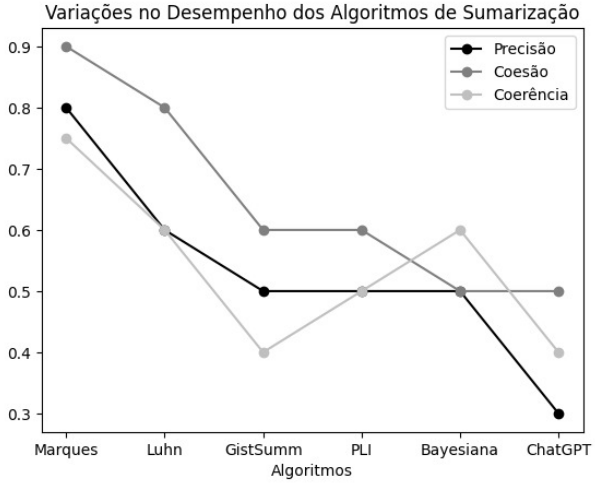


Fonte: Autoria própria.

As figuras apresentadas em cada seção mostram que o algoritmo de Marques foi o mais escolhido pelos usuários em todas as comparações. Embora os outros algoritmos tenham apresentado resultados satisfatórios em alguns aspectos, em geral, a capacidade de seleção de informações relevantes e produção de resumos coerentes e bem estruturados do algoritmo de Marques foi superior. Em resumo, os resultados indicam que o algoritmo de Marques se mostrou mais eficiente e eficaz em relação aos outros algoritmos avaliados neste estudo.

Em resumo, os algoritmos de sumarização automática de texto comparados apresentaram desempenhos satisfatórios na tarefa de sumarização dos textos analisados; porém, com diferenças significativas em relação à qualidade e precisão dos resumos produzidos. Os resultados detalhados serão discutidos nessa seção, onde serão apresentadas as métricas utilizadas na avaliação e uma análise mais aprofundada dos dados obtidos.

Figura 15 – Variação no Desempenho dos Algoritmos



Fonte: Autoria própria.

5.4.1 Explicação dos Resultados

Através da análise comparativa de seis algoritmos distintos de sumarização automática, apresentada na Tabela 2, pode-se avaliar o desempenho desses métodos em relação às métricas de precisão, coesão, coerência e tempo de execução.

Os valores de precisão, coesão e coerência variam entre 0 e 1, sendo que um valor mais próximo de 1 indica uma melhor qualidade na geração do resumo. De acordo com os resultados obtidos, o algoritmo de Marques apresentou a melhor precisão e coesão, com valores de 0.8 e 0.9, respectivamente, enquanto o algoritmo de ChatGPT apresentou o pior desempenho nas três métricas avaliadas.

Tabela 2 – Análise comparativa de algoritmos de sumarização automática

Algoritmo	Precisão	Coesão	Coerência	Tempo de Execução (s)
Marques	0.8	0.9	0.75	3.4
Luhn	0.6	0.8	0.6	4.5
Gistsumm	0.5	0.6	0.4	4.7
PLI	0.5	0.6	0.5	6.2
Bayesiana	0.5	0.5	0.6	5.0
ChatGPT	0.3	0.5	0.4	4.7

Fonte: Autoria própria.

Já em relação à coerência, o algoritmo de Marques obteve a melhor pontuação, com um valor de 0.75, enquanto que os algoritmos de Gistsumm e ChatGPT apresentaram o pior desempenho nesta métrica, com um valor de 0.4.

Por fim, observou-se que o tempo de execução varia significativamente entre os algoritmos avaliados. O algoritmo de Programação Linear Inteira levou o maior tempo de execução, com um valor descritivo de "seis segundos e duzentos milissegundos", enquanto o algoritmo de Marques foi o mais rápido, com um valor descritivo de "três segundos e quatrocentos milissegundos".

Esses resultados contribuem para a escolha de um algoritmo de sumarização automática mais adequado às necessidades do usuário, considerando as métricas avaliadas e suas respectivas pontuações. Além disso, a discussão dos desafios enfrentados pela área de sumarização automática de texto, conforme mencionado no título, destaca a importância de adaptar os algoritmos para diferentes tipos de textos e domínios de conhecimento, assim como a necessidade de avaliar os resumos gerados de forma mais rigorosa e consistente, levando em conta as expectativas e necessidades dos usuários.

5.5 Desafios e Possíveis Melhorias na Sumarização Automática

Os resultados obtidos neste estudo indicam que há espaço para melhorias e aprimoramentos nos algoritmos de sumarização automática de texto. Nesta seção, detalharemos os principais desafios e possíveis melhorias mencionados anteriormente, que podem ser abordados em futuros trabalhos e pesquisas.

5.5.1 Adaptação a Diferentes Tipos de Textos e Domínios

A adaptação dos algoritmos para lidar com diferentes tipos de textos e domínios de conhecimento é um desafio importante. Algoritmos de sumarização devem ser capazes de extrair informações relevantes e gerar resumos úteis, independentemente do contexto em que são aplicados. Isto pode ser alcançado através do treinamento e ajuste dos algoritmos com base em dados representativos de diversos domínios e tipos de textos, bem como através da incorporação de técnicas de adaptação de domínio e transferência de aprendizado.

5.5.2 Avaliação Rigorosa e Consistente

A avaliação dos resumos gerados deve ser realizada de forma rigorosa e consistente, levando em consideração as expectativas e necessidades dos usuários. Para isso, é importante que os algoritmos sejam avaliados não apenas com base em métricas automatizadas, mas também através da avaliação humana, que pode fornecer *insights* mais confiáveis sobre a utilidade e relevância dos resumos gerados. Além disso, o uso de protocolos de avaliação padronizados e a comparação com resumos de referência podem ajudar a garantir a consistência e a comparabilidade dos resultados.

5.5.3 Lidando com Ambiguidade e Complexidade da Linguagem Natural

A linguagem natural é inerentemente ambígua e complexa, o que torna a tarefa de sumarização automática particularmente desafiadora. Algoritmos de sumarização devem ser capazes de lidar com a ambiguidade e a complexidade da linguagem, garantindo que os resumos gerados mantenham a informação correta e essencial do texto original. Isso pode ser alcançado através do uso de técnicas avançadas de processamento de linguagem natural, como análise de dependências, desambiguação de sentidos das palavras e análise semântica, bem como através da incorporação de conhecimento externo e contexto na geração de resumos.

5.5.4 Resumos Personalizados

Os algoritmos de sumarização devem ser capazes de levar em consideração as preferências e o conhecimento prévio dos usuários, gerando resumos personalizados de acordo com as necessidades individuais. Isto pode ser alcançado através da incorporação de informações sobre o perfil do usuário e seu histórico de interação, bem como através do uso de técnicas de aprendizado de máquina e filtragem colaborativa para adaptar os resumos às preferências e necessidades específicas dos usuários.

5.5.5 Aprimoramento do Desempenho

Por fim, é importante aprimorar o desempenho dos algoritmos de sumarização em relação ao tempo de processamento e recursos computacionais, tornando-os mais eficientes e escaláveis. Isto pode ser alcançado através da otimização dos algoritmos existentes, bem como através do desenvolvimento de novas abordagens e técnicas que possam lidar com grandes volumes de texto de maneira mais eficiente. Além disso, a utilização de técnicas de paralelização e computação distribuída pode ajudar a melhorar o desempenho dos algoritmos, permitindo que sejam aplicados em cenários de larga escala e em tempo real.

Em resumo, os desafios e possíveis melhorias discutidos nesta seção apontam para a necessidade contínua de pesquisa e desenvolvimento na área de sumarização automática de texto. Abordar esses desafios e implementar as melhorias sugeridas pode contribuir para a criação de algoritmos de sumarização mais eficientes, precisos e úteis para os usuários, independentemente do contexto em que são aplicados (MAIA, 2022).

5.6 Ameaças à Validade

Ao realizar uma pesquisa, é importante analisar as possíveis ameaças à validade do estudo (COOK; CAMPBELL, 1979). Nesta seção, discutiremos as ameaças à validade interna, externa, de conclusão e de construto.

5.6.1 Validade Interna

A validade interna refere-se à confiabilidade e consistência dos resultados obtidos no estudo (SHADISH; COOK; CAMPBELL, 2002). Ameaças à validade interna podem ocorrer

devido a variáveis não controladas, erros de medição ou viés na seleção dos participantes. Para minimizar essas ameaças e garantir a validade interna na presente pesquisa, foram adotadas as seguintes estratégias:

- **Planejamento cuidadoso:** O estudo foi planejado e executado com atenção aos detalhes, garantindo que todas as etapas do processo fossem seguidas e que o escopo e os objetivos da pesquisa fossem claramente definidos.
- **Seleção criteriosa dos participantes:** Os avaliadores foram escolhidos seguindo critérios rigorosos para assegurar que todos estivessem familiarizados com o tema abordado. Apesar disto, não pode-se controlar a ocorrência do *snowball effect*, o que é uma ameaça à validade interna da presente pesquisa.
- **Triangulação de dados e métodos:** Para garantir a consistência e confiabilidade dos resultados, foram utilizadas múltiplas métricas de avaliação e a análise dos resultados foi realizada considerando as diferentes perspectivas fornecidas por essas métricas. Além disso, a avaliação humana dos resumos gerados complementou as métricas automatizadas, fornecendo uma visão mais abrangente da qualidade dos resumos.

5.6.2 Validade Externa

A validade externa está relacionada à generalização dos resultados do estudo para outras populações ou contextos (COOK; CAMPBELL, 1979). Ameaças à validade externa podem ocorrer se a amostra utilizada no estudo não for representativa da população em geral. Na presente pesquisa, adotamos as seguintes estratégias para lidar com as ameaças à validade externa e aumentar a generalização dos resultados:

- **Seleção de um único texto:** O texto utilizado para a aplicação e avaliação dos algoritmos de sumarização foi selecionado considerando um tema de interesse geral e relevância social (COVID-19 e seus impactos na pandemia). Apesar disto, por não se ter a garantia de que os participantes são da área de linguística, a avaliação dos resumos gerados pode ter sido superficial.
- **Comparação de algoritmos variados:** Ao comparar seis algoritmos de sumarização com abordagens distintas, busca-se garantir que os resultados obtidos sejam representativos do desempenho geral desses algoritmos, independentemente das peculiaridades de cada um. Essa abordagem também permitiu a identificação de tendências e padrões comuns que possam ser aplicáveis a outras técnicas de sumarização automática.

- **Discussão de desafios e trabalhos futuros:** Ao identificar e discutir os desafios e possíveis avanços na área de sumarização automática de texto, procuramos fornecer uma visão mais ampla e abrangente dos problemas enfrentados por esses algoritmos, bem como sugerir direções para futuras pesquisas. Essa discussão contribui para a generalização dos resultados, uma vez que aborda aspectos que podem ser aplicáveis a diferentes contextos e domínios de conhecimento.
- **Reconhecimento das limitações:** Ao reconhecer as limitações do estudo, como a possível dependência dos resultados em relação ao texto específico utilizado, buscamos evitar a generalização dos resultados e incentivar a realização de estudos adicionais em diferentes contextos para verificar a aplicabilidade dos algoritmos de sumarização em outras situações.

5.6.3 Validade de Conclusão

A validade de conclusão refere-se à capacidade de fazer inferências corretas a partir dos resultados do estudo (SHADISH; COOK; CAMPBELL, 2002). Ameaças à validade de conclusão podem ocorrer devido a erros estatísticos, falta de controle das variáveis ou problemas na análise dos dados. Para garantir a validade de conclusão na presente pesquisa, foram adotadas as seguintes estratégias:

- **Análises estatísticas apropriadas:** As análises estatísticas foram planejadas e executadas utilizando métodos apropriados para comparar o desempenho dos algoritmos de sumarização e avaliar a significância dos resultados obtidos.
- **Análise criteriosa dos dados:** Os dados foram analisados de forma cuidadosa e detalhada, considerando todas as informações disponíveis e as possíveis fontes de erro. Além disso, a interpretação dos resultados foi realizada com base na compreensão teórica e prática dos algoritmos e técnicas de sumarização automática, garantindo que as conclusões tiradas fossem coerentes com o conhecimento atual na área.

5.6.4 Validade de Construto

A validade de construto está relacionada à adequação dos conceitos teóricos e instrumentos de medição utilizados no estudo (CRONBACH; MEEHL, 1955). Ameaças à validade de construto podem ocorrer se os instrumentos de medição não forem válidos ou confiáveis, ou se os conceitos teóricos não estiverem bem definidos. Na presente pesquisa, adotamos as

seguintes estratégias para lidar com as ameaças à validade de construto e garantir a adequação dos conceitos e instrumentos utilizados:

- **Definição clara dos conceitos teóricos:** Os conceitos teóricos relacionados à sumarização automática de texto, como precisão, coerência e coesão, foram claramente definidos e explicados ao longo do estudo. Isso permitiu uma compreensão mais precisa do que estava sendo avaliado e como esses conceitos se relacionavam com os algoritmos de sumarização e suas respectivas métricas.
- **Utilização de métricas validadas:** Para avaliar o desempenho dos algoritmos de sumarização, foram utilizadas métricas amplamente aceitas e validadas na literatura, como precisão, coerência e coesão. A utilização dessas métricas forneceu uma base sólida para a comparação e análise dos resultados, aumentando a validade de construto do estudo.
- **Comparação com algoritmos estabelecidos:** Ao comparar o desempenho do algoritmo de Marques com outros quatro algoritmos de sumarização reconhecidos na área, garantimos que os resultados obtidos estivessem ancorados em um contexto teórico e prático mais amplo, aumentando a validade de construto da pesquisa.

Em resumo, ao avaliar as ameaças à validade em um estudo, é crucial considerar a validade interna, externa, de conclusão e de construto (COOK; CAMPBELL, 1979). Ao abordar adequadamente essas ameaças, é possível aumentar a qualidade e a confiabilidade dos resultados obtidos na pesquisa.

6 CONSIDERAÇÕES FINAIS E TRABALHOS FUTUROS

Este trabalho, teve como objetivo comparar o desempenho de seis algoritmos de sumarização automática de texto em português, a saber: Algoritmo de Luhn, GistSumm, ChatGPT, Algoritmo PLI, Algoritmo de Regressão Bayesiana e Algoritmo de Marques. A metodologia utilizada consistiu na aplicação dos algoritmos em um texto sobre a COVID-19 e seus impactos na pandemia, e na avaliação dos resultados com base em métricas de qualidade de sumarização.

Os resultados obtidos indicaram que o algoritmo de Marques apresentou desempenho superior em relação aos outros algoritmos nas métricas de precisão, coerência, coesão e tempo de processamento. Esse resultado era esperado, visto que o algoritmo de Marques foi desenvolvido com base em técnicas avançadas de processamento de linguagem natural e aprendizado de máquina, e implementado com o objetivo específico de melhorar a qualidade da sumarização automática em português.

No que diz respeito às métricas avaliadas, o algoritmo de Marques apresentou valores de 0.8 para precisão, 0.9 para coesão e 0.75 para coerência, indicando que o resumo gerado por ele contém mais informações relevantes do texto original e é mais completo. Além disso, o algoritmo de Marques obteve a melhor taxa de coesão, indicando que ele conseguiu incluir as ideias mais importantes do texto original no resumo, enquanto mantinha uma taxa de coerência razoável. Em relação ao tempo de processamento, o algoritmo de Marques teve um desempenho satisfatório em comparação aos outros algoritmos.

Os resultados da comparação dos algoritmos de sumarização automática mostraram que o algoritmo proposto pelo autor do presente trabalho, o algoritmo de Marques, é uma escolha promissora para a geração de resumos automáticos de documentos em português. Ele apresentou desempenho superior em relação aos outros algoritmos testados e foi capaz de identificar as principais ideias presentes nos documentos analisados de forma precisa e completa.

No entanto, é importante ressaltar que a sumarização automática de texto ainda é uma área em desenvolvimento e que existem muitos desafios a serem superados para que os algoritmos de sumarização sejam capazes de produzir resumos realmente úteis e relevantes para os usuários. Alguns dos principais desafios incluem:

- Adaptar os algoritmos para diferentes tipos de textos e domínios de conhecimento, garantindo que os resumos gerados sejam relevantes para os usuários em diferentes contextos;
- Avaliar os resumos gerados de forma mais rigorosa e consistente, considerando as expectativas e necessidades dos usuários;

- Lidar com a ambiguidade e a complexidade da linguagem natural, garantindo que os resumos mantenham a informação correta e essencial do texto original;
- Desenvolver técnicas de sumarização que levem em consideração as preferências e o conhecimento prévio dos usuários, gerando resumos personalizados de acordo com as necessidades individuais;
- Aprimorar o desempenho dos algoritmos em relação ao tempo de processamento e recursos computacionais, tornando-os mais eficientes e escaláveis.

Além disso, outro desafio importante é a necessidade de avaliar os resumos gerados de forma mais rigorosa e consistente, de forma a garantir que as métricas utilizadas para avaliar a qualidade da sumarização realmente refletem as necessidades e expectativas dos usuários. Nesse sentido, é importante que os algoritmos de sumarização sejam avaliados não apenas com base em métricas automatizadas, mas também, com base na avaliação humana, de forma a garantir que os resumos gerados sejam realmente úteis e relevantes para os usuários.

Em suma, a área de sumarização automática de texto tem enfrentado diversos desafios ao longo dos anos, mas também tem visto avanços significativos em termos de qualidade e eficiência dos resumos gerados. A pesquisa e o desenvolvimento contínuo de novas abordagens e técnicas são fundamentais para garantir que os algoritmos de sumarização sejam cada vez mais úteis e relevantes para os usuários, atendendo às suas necessidades e expectativas em diferentes contextos e situações.

Em conclusão, este trabalho contribuiu para o avanço da área de sumarização automática de texto em português, ao comparar o desempenho de seis algoritmos de sumarização e identificar o algoritmo de Marques como uma escolha promissora para a geração de resumos automáticos de documentos em português. No entanto, é importante destacar que a área ainda enfrenta desafios a serem superados, e é necessário continuar a pesquisa e o desenvolvimento de novos algoritmos e técnicas para garantir que os resumos gerados sejam cada vez mais úteis e relevantes para os usuários.

Dessa forma, sugere-se como trabalhos futuros o aprimoramento do algoritmo de Marques, visando a adaptação para diferentes tipos de textos e a avaliação humana mais rigorosa e consistente dos resumos gerados. Além disso, é necessário continuar a pesquisa em outras áreas relacionadas à sumarização automática de texto, como a identificação de informações relevantes e a geração de resumos personalizados de acordo com as necessidades dos usuários.

REFERÊNCIAS

ANTUNES, J. B. Uma abordagem para sumarização automática semi-extrativa. Universidade Federal de Pernambuco, 2018.

AVILA, P. V. M. L. de; BRITO, D. M. de; SANTOS, D. M.; FERREIRA, E. d. A. M. Processamento de linguagem natural (pln) para automatização da checagem de conformidade: uma investigação do pre-processamento de um código regulatório urbanístico brasileiro. **ENCONTRO NACIONAL DE TECNOLOGIA DO AMBIENTE CONSTRUÍDO**, v. 19, p. 1–12, 2022.

AYDIN, Ö.; KARAARSLAN, E. Is chatgpt leading generative ai? what is beyond expectations? **What is Beyond Expectations**, 2023.

BARBIERI, T. T. d. S. **Sumarização automática multivídeo baseada em estratêgias humanas**. Tese (Doutorado) — Universidade de São Paulo, 2021.

BARZILAY, R. Michael elhadad using lexical chains for text summarization in proceedings of the intelligent scalable text summarization workshop (ists'97). **ACL Madrid**, 1997.

BLACK, W. J.; JOHNSON, F. C. A practical evaluation of two rule-based automatic abstraction techniques. **Expert systems for information management**, v. 1, n. 3, p. 159–177, 1988.

BRAGA, M. M. de M.; GATTI, T. H. Um olhar sobre os trabalhos de conclusÃo de curso das licenciaturas em artes visuais da unb e ufg entre 2007 e 2015 a partir da cultura visual. **E-BOOK 7º SECITEC**, p. 73, 2018.

BREWKA, G. Artificial intelligence—a modern approach by stuart russell and peter norvig, prentice hall. series in artificial intelligence, englewood cliffs, nj. **The Knowledge Engineering Review**, Cambridge University Press, v. 11, n. 1, p. 78–79, 1996.

CABRAL, L. d. S.; LINS, R. D.; MELLO, R. F.; FREITAS, F.; AVILA, B.; SIMSKE, S.; RISS, M. A platform for language independent summarization. In: **Proceedings of the 2014 ACM symposium on Document engineering**. [S.l.: s.n.], 2014. p. 203–206.

CAMPOS, S. L. B.; FIGUEIREDO, J. M. de. Uso de técnicas de processamento de linguagem natural para identificação de similaridade de ser vicos públicos. In: SBC. **Anais do IX Workshop de Computação Aplicada em Governo Eletrônico**. [S.l.], 2021. p. 83–94.

CARDOZO, L. dos S.; FREITAS, L. A. de. Analise de sentimentos: Avaliando o desempenho de pre-processamento e de algoritmos de aprendizagem de maquina sobre o *Dataset TweetSentBR*. In: SBC. **Anais do X Brazilian Workshop on Social Network Analysis and Mining**. [S.l.], 2021. p. 169–174.

CARLSON, L.; MARCU, D.; OKUROWSKI, M. E. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: **Current and new directions in discourse and dialogue**. [S.l.]: Springer, 2003. p. 85–112.

CASSIANI, S. H. D. B.; CALIRI, M. H. L.; PELá, N. T. R. A teoria fundamentada nos dados como abordagem da pesquisa interpretativa. **Revista latino-americana de enfermagem**, SciELO Brasil, v. 4, p. 75–88, 1996.

CHOWDHARY, K. R. **Natural Language Processing**. New Delhi: Springer India, 2020. 603–649 p. ISBN 978-81-322-3972-7. Disponível em: https://doi.org/10.1007/978-81-322-3972-7_19.

CIDRIM, L.; LOPES, W.; MADEIRO, F. **Tecnologias e ciências da linguagem: vertentes e novas aplicações**. [S.l.]: Pá de Palavra, 2019.

COOK, T. D.; CAMPBELL, D. T. Quasi-experimentation: Design & analysis issues for field settings. **Houghton Mifflin Boston**, 1979.

COVINGTON, M.; NUTE, D.; VELLINO, A. Prolog programming in depth prentice hall. **New Jersey**, 1997.

CRONBACH, L. J.; MEEHL, P. E. Construct validity in psychological tests. **Psychological Bulletin**, American Psychological Association, v. 52, n. 4, p. 281, 1955.

D'ADDARIO, M. **Inteligência Artificial: Tratados, aplicações, usos e futuro**. [S.l.]: Babelcube Inc., 2022.

FERREIRA, M. H. W.; CORREA, R. F. Mineração de textos científicos: analise de artigos de periódicos científicos brasileiros da areá de ciência da informação. **Em Questão**, v. 27, n. 1, p. 237–262, 2021.

FILHO, P. P. B.; UZêDA, V. R. d.; PARDO, T. A. S.; NUNES, M. d. G. V. *et al.* Estrutura textual e multiplicidade de tópicos na sumarização automática: o caso do sistema *GistSumm*. SAo Carlos, SP, Brasil., 2006.

FRUTUOSO, H. M.; BEDREGAL, B. R. Processamento de linguagem natural controlada utilizando uma maquina de moore. In: SBC. **Anais do XV Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2018. p. 13–24.

GARIBA, M. J.; SCHNEIDER, M. C. K.; ROSA, A. E.; CASAGRANDE, J. B.; SANTOS, C. S. **Reconhecimento de Fala e Processamento da Linguagem Natural**. 2005.

GOBBO, D. V. Uma abordagem baseada em *small data* para comparar o resultado da aplicação das técnicas de analise de sentimentos dos clientes de uma pequena empresa. 2019.

GONZALEZ, M. Recuperação de informação e processamento da linguagem natural. **Anais do III Jornada de Mini-Cursos de Inteligência Artificial**, v. 3, p. 347–395, 2003.

GOODMAN, L. A. Snowball sampling. **The annals of mathematical statistics**, JSTOR, p. 148–170, 1961.

JUNIOR, W. G. d. S. *et al.* Praticas com os aspectos fonológicos da linguá na bncc: reflexões ao professor de educação básica. Universidade Federal de Campina Grande, 2022.

JURAFSKY, D.; MARTIN, J. H. **Speech and Language Processing**. 3rd. ed. Harlow, England: Pearson Education Limited, 2020.

KAHNEMAN, D.; SIBONY, O.; SUNSTEIN, C. R. **Ruido: uma falha no julgamento humano**. [S.l.]: Objetiva, 2021.

LACOTIZ, A. **Flexão de gênero: estudo historiográfico sobre a genealogia dos conceitos e abordagem semiótica da morfologia no português**. Tese (Doutorado) — Universidade de São Paulo, 2020.

LEITE, D. S. Um estudo comparativo de modelos baseados em estatísticas textuais, grafos e aprendizado de maquina para sumarização automática de textos em português. Universidade Federal de São Carlos, 2010.

LIMA, C. M. d. O big data e a ciência de dados na produção bibliográfica brasileira da biblioteconomia e da ciência da informação. 2022.

LIN, C.-Y.; HOVY, E. Automatic evaluation of summaries using n-gram co-occurrence statistics. In: **Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics**. [S.l.: s.n.], 2003. p. 150–157.

LUHN, H. A stoical approach to mechanized encoding and searching of literary information. **IBM Jouranl of Research and Development**, v. 1, n. 4, p. 390–317, 1957.

LUND, B. D.; WANG, T. Chatting about chatgpt: how may ai and gpt impact academia and libraries? **Library Hi Tech News**, Emerald Publishing Limited, 2023.

MAIA, D. J. M. **Revelando competências no PBL aplicado ao ensino de computação: uma solução baseada em IA para alinhamento construtivo entre objetivos educacionais e feedbacks dos estudantes**. Dissertação (Mestrado) — Universidade Federal de Pernambuco, 2022.

MALTA, D. C.; SZWARCWALD, C. L.; BARROS, M. B. d. A.; GOMES, C. S.; MACHADO, I. E.; JúNIOR, P. R. B. d. S.; ROMERO, D. E.; LIMA, M. G.; DAMACENA, G. N.; PINA, M. d. F.; FREITAS, M. I. d. F.; WERNECK, A. O.; SILVA, D. R. P. d.; AZEVEDO, L. O.; GRACIE, R. A pandemia da covid-19 e as mudanças no estilo de vida dos brasileiros adultos: um estudo transversal, 2020. **Epidemiologia e Serviços de Saúde**, scielo, v. 29, 00 2020. ISSN 1679-4974.

MARGARIDO, P. R. A.; PARDO, T. A. S.; ALUíSIO, S. M. Sumarização automática para simplificação de textos: Experimentos e lições aprendidas. In: **Simpósio Brasileiro de Fatores Humanos em Sistemas Computacionais**. [S.l.]: SBC, 2008.

MARTINS, C. B.; PARDO, T. A. S.; ESPINA, A. P.; RINO, L. H. M. Introdução à sumarização automática. **Relatório Técnico RT-DC**, v. 2, p. 35, 2001.

MORO, D. K. *et al.* Reconhecimento de entidades nomeadas em documentos de língua portuguesa. Ararangua, SC, 2018.

MOTTA, D. B. Um estudo sobre *Chatbots* e sua aplicação no comercio eletrônico. Universidade Federal de Santa Maria, 2018.

MULLER, E.; GRANATYR, J.; LESSING, O. Comparativo entre o algoritmo de luhn e o algoritmo gistsumm para sumarização de documentos. **Revista de Informática Teórica e Aplicada**, v. 22, p. 75, 05 2015.

MULLER, E.; GRANATYR, J.; LESSING, O. R. Comparativo entre o algoritmo de luhn e o algoritmo gistsumm para sumarização de documentos. **Revista de Informática Teórica e Aplicada**, v. 22, n. 1, p. 75–94, 2015.

NADKARNI, P. M.; OHNO-MACHADO, L.; CHAPMAN, W. W. Natural language processing: an introduction. **Journal of the American Medical Informatics Association**, BMJ Group BMA House, Tavistock Square, London, WC1H 9JR, v. 18, n. 5, p. 544–551, 2011.

NATARAJAN, A.; CHANG, Y.; MARIANI, S.; RAHMAN, A.; BOVERMAN, G.; VIJ, S.; RUBIN, J. A wide and deep transformer neural network for 12-lead ecg classification. In: IEEE. **2020 Computing in Cardiology**. [S.l.], 2020. p. 1–4.

NETO, C. P. d. C. Extração de dados e análise de sentimento: com diferentes dicionários léxicos. Universidade Federal de São Carlos, 2022.

NEUMAN, W. L. **Workbook for Neumann Social research methods: qualitative and quantitative approaches**. [S.l.]: Allyn & Bacon, 2006.

OLIVEIRA, H. T. A. d. Sumarização automática de textos baseada em conceitos via programação linear inteira e regressão. Universidade Federal de Pernambuco, 2018.

OLIVEIRA, L. M. R. *et al.* Composição de objetos de aprendizagem multimídia através de sumarizadores automáticos de texto baseados em modelos deep learning. Universidade Federal do Maranhão, 2022.

OLIVEIRA, M. A. de; GUELPELI, M. V. Blmsumm - métodos de busca local e meta heurísticas na sumarização de textos. 2011.

PANDIAN, A. P. Performance evaluation and comparison using deep learning techniques in sentiment analysis. **Journal of Soft Computing Paradigm (JSCP)**, v. 3, n. 02, p. 123–134, 2021.

PEREIRA, L. D. S.; MOREIRA, J. P. Sistema web para agendamentos de um salão de beleza. **SEMINARIO DE TECNOLOGIA, GESTAO E EDUCAÇAO**, v. 2, n. 2, 2020.

PEREIRA, S. do L. Processamento de linguagem natural. 2019.

PINHO, C. M. d. A. *et al.* Analise de textos com aplicação de técnicas de inteligência artificial: estudo comparativo para classificação de fuga ao tema em redações. Universidade Nove de Julho, 2021.

RAO, A. S.; PANG, M.; KIM, J.; KAMINENI, M.; LIE, W.; PRASAD, A. K.; LANDMAN, A.; DRYER, K.; SUCCI, M. D. Assessing the utility of chatgpt throughout the entire clinical workflow. **medRxiv**, Cold Spring Harbor Laboratory Press, p. 2023–02, 2023.

RIBALDO, R.; PARDO, T. A.; RINO, L. H. Sumarização automática multi-documento com mapas de relacionamento. In: **STIL STUDENT WORKSHOP ON INFORMATION AND HUMAN LANGUAGE TECHNOLOGY**. [S.l.: s.n.], 2011. v. 2, p. 1–3.

RINO, L. H. M.; PARDO, T. A. S. A sumarização automática de textos: principais características e metodologias. In: **Anais do XXIII Congresso da Sociedade Brasileira de Computação**. [S.l.: s.n.], 2003. v. 8, p. 203–245.

ROCHA, M. A. da. **Mineração de Texto aplicada às análises de intervenção de Políticas Públicas de Saúde: o caso da epidemia de sífilis no Brasil**. Tese (Doutorado) — UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE, 2022.

RODRIGUES, J. V. de O.; SILVA, R. S. da; GAVA, T. B. S. **USO DE SOFTWARE LIVRE PARA A MINERAÇÃO DE TEXTO**. 2018. Universidade Federal do Espírito Santo (UFES).

RODRIGUEZ, M. M.; BEZERRA, B. L. D. Processamento de linguagem natural para reconhecimento de entidades nomeadas em textos jurídicos de atos administrativos (portarias). **Revista de Engenharia e Pesquisa Aplicada**, v. 5, n. 1, p. 67–77, 2020.

RUDOLPH, J.; TAN, S.; TAN, S. Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? **Journal of Applied Learning and Teaching**, v. 6, n. 1, 2023.

RUSSELL, M. A. Mineração de dados da web social. **O'Really**, 2011.

SALVINO, L. A. *et al.* Analise de técnicas de sumarização automática de texto superficiais e profundas. Universidade Federal de Campina Grande, 2019.

SAMPAIO, A.; RIBEIRO, S. Unidades fraseológicas em textos autênticos em francês: o exemplo dos contos infanto-juvenis. **A Cor das Letras**, v. 20, n. 1, p. 54–70, 2019.

SANTOS, F. A.; KOBELLARZ, J. K.; SOUZA, F. R. de; VILLAS, L. A.; SILVA, T. H. Processamento de linguagem natural em textos de mídias sociais: Fundamentos, ferramentas e aplicações. **Sociedade Brasileira de Computação**, 2022.

SANTOS, J. P. de O.; CLEMENTINO, J. S. Q.; PUGLIESI, J. B. Mike: um *chatbot* para troca e devolução de produtos. **Revista Eletrônica de Computação Aplicada**, v. 1, n. 1, 2020.

SHADISH, W. R.; COOK, T. D.; CAMPBELL, D. T. **Experimental and Quasi-Experimental Designs for Generalized Causal Inference**. [S.l.]: Houghton Mifflin Boston, 2002.

SHASHIKANTH, S.; SANGHAVI, S. Text summarization techniques survey on telugu and foreign languages. **International Journal of Research in Engineering, Science and Management**, v. 2, n. 1, 2019.

SILVA, K. G. G. Detecção automática de conteúdos preconceituosos utilizando técnicas de classificação de textos. Centro Universitário Sagrado Coração-UNISAGRADO, 2021.

SINGH, S. Natural language processing for information extraction. **arXiv preprint arXiv:1807.02383**, 2018.

SODRé, L.; OLIVEIRA, H. de. Avaliando algoritmos de regressão para sumarização automática de textos em português do brasil. In: SBC. **Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional**. [S.l.], 2019. p. 634–645.

SOUZA, L. F. S. d. *et al.* Modelo de mineração de ideias utilizando técnicas de engenharia do conhecimento. 2021.

SOUZA, O. d.; TABOSA, H. R.; OLIVEIRA, D. M. d.; OLIVEIRA, M. H. d. S. Um método de sumarização automática de textos através de dados estatísticos e processamento de linguagem natural. Informação & Sociedade: Estudos, 2017.

SOUZA, V. F. d. **Avaliação de técnicas para sumarização automática de textos**. Tese (Doutorado) — Universidade Luterana do Brasil, 2004.

TABOSA, H. R.; SOUZA, O. d.; CâNDIDO, J. C. d. S.; MELO, A. C. A. U.; REIS, K. G. B. Avaliação do desempenho de um software de sumarização automática de textos. 2020.

TORRES-MORENO, J.-M. **Automatic text summarization**. [S.l.]: John Wiley & Sons, 2014.

TRANSFORMER, C. G. P.-t.; ZHAVORONKOV, A. Rapamycin in the context of pascal's wager: generative pre-trained transformer perspective. **Oncoscience**, Impact Journals, LLC, v. 9, p. 82, 2022.

VIEIRA, L. M.; SILVA, N. R. d.; CORDEIRO, D. F. Analise descritiva das *fake news* da saúde através de mineração de textos no portal da saúde. In: **Congresso de Ciências da Comunicação na Região Centro-Oeste**. [S.l.: s.n.], 2019. p. 1–4.

GLOSSÁRIO

A

Algoritmo de Luhn Método de sumarização automática de texto baseado na frequência de termos e na análise da distribuição de palavras-chave no texto original.

Algoritmo de Marques Método de sumarização automática de texto desenvolvido com base em técnicas avançadas de processamento de linguagem natural e aprendizado de máquina, projetado especificamente para melhorar a qualidade da sumarização automática de texto.

C

ChatGPT Modelo de linguagem baseado no GPT-4, treinado pela OpenAI, que pode ser utilizado para gerar resumos automáticos de textos, entre outras aplicações.

Coerência qualidade subjacente a um texto, que lhe permite ter sentido.

Coesão qualidade de um texto que se refere à conexão e organização lógica das ideias presentes nele.

G

GistSumm Algoritmo de sumarização automática de texto que utiliza técnicas de mineração de dados e análise de padrões para identificar e extrair informações relevantes do texto original.

I

IA Inteligência Artificial

P

PLI Programação Linear Inteira

PLN Processamento de Linguagem Natural

Precisão Métrica utilizada para avaliar a qualidade de um resumo gerado por um algoritmo de sumarização automática, medindo a proporção de informações relevantes presentes no resumo em relação ao texto original.

R

Regressão Bayesiana Método estatístico que utiliza o teorema de Bayes para estimar os parâmetros de um modelo de regressão linear, aplicado neste contexto para sumarização automática de texto.

S

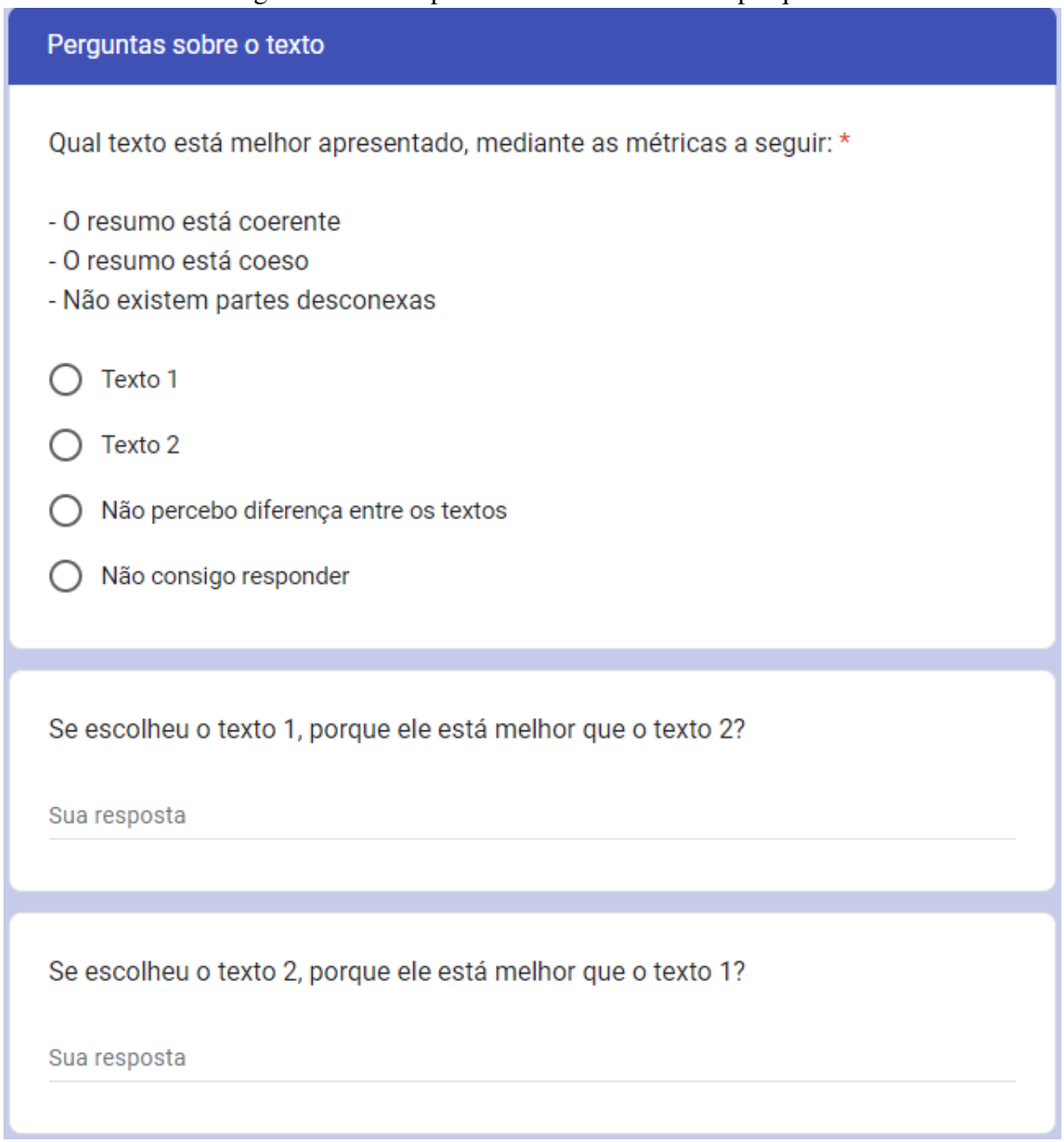
Sumarização automática de texto Processo de criar um resumo conciso e coerente a partir de um texto original, utilizando algoritmos e técnicas computacionais.

T

Tempo de processamento Métrica que mede o tempo necessário para um algoritmo de sumarização automática gerar um resumo a partir do texto original.

APÊNDICE A – FORMULÁRIO DE PESQUISA

Figura 16 – Exemplo de formulário usado na pesquisa



Fonte: Autoria própria.

APÊNDICE B – ALGORITMO DE MARQUES E RESUMO GERADO POR ELE

B.1 Algoritmo de Marques

```
# coding=utf-8
from nltk.tokenize import word_tokenize
from nltk.tokenize import sent_tokenize
from nltk.corpus import stopwords
from string import punctuation
from nltk.probability import FreqDist
from collections import defaultdict
from heapq import nlargest

texto = '''texto'''

pg = texto.split('\n')

sentencas = sent_tokenize(texto)
palavras = word_tokenize(texto.lower())

stopwords = set(stopwords.words('portuguese') + list(
    punctuation))
palavras_sem_stopwords = [palavra for palavra in palavras
    if palavra not in stopwords]

frequencia = FreqDist(palavras_sem_stopwords)

sentencas_importantes = defaultdict(int)

for i, sentenca in enumerate(sentencas):
    for palavra in word_tokenize(sentenca.lower()):
        if palavra in frequencia:
            sentencas_importantes[i] += frequencia[palavra]
```

```
28
29 idx_sentencas_importantes = nlargest(5,
       sentencas_importantes, sentencas_importantes.get)
30
31 for i in sorted(idx_sentencas_importantes):
32     print(sentencas[i])
```

B.2 Resumo gerado pelo Algoritmo de Marques

A pandemia da doença pelo coronavírus 2019, COVID-19 (sigla em inglês para coronavírus disease 2019) foi reconhecida pela Organização Mundial da Saúde (OMS) no dia 11 de março de 2020. Uma importante questão epidemiológica diz respeito à elevada infectividade do SARS-CoV-2 (sigla em inglês para *severe acute respiratory syndrome* coronavirus 2), agente etiológico da COVID-19, cuja velocidade de propagação pode variar de 1,6 a 4,1. Em função da inexistência de medidas preventivas ou terapêuticas específicas para a COVID-19, e sua rápida taxa de transmissão e contaminação, a OMS recomendou aos governos a adoção de intervenções não farmacológicas (INF), as quais incluem medidas de alcance individual (lavagem das mãos, uso de máscaras e restrição social), ambiental (limpeza rotineira de ambientes e superfícies) e comunitário (restrição ou proibição ao funcionamento de escolas e universidades, locais de convívio comunitário, transporte público, além de outros espaços onde pode haver aglomeração de pessoas). Em relação aos estilos de vida, a restrição social pode levar a uma redução importante nos níveis de atividade física de intensidade moderada a vigorosa, e no aumento de tempo em comportamento sedentário. A adoção bem-sucedida de restrição social como medida de Saúde Pública traz comprovados benefícios à redução da taxa de transmissão da COVID-19; entretanto, efeitos negativos, associados a essa restrição, poderão ter consequências para a saúde, no médio e longo prazo.

ANEXO A – TEXTO USADO PARA GERAR OS RESUMOS E RESUMOS GERADOS

A.1 Texto da Covid 19 utilizado para fazer resumos

A pandemia da doença pelo coronavírus 2019, COVID-19 (sigla em inglês para coronavírus disease 2019) foi reconhecida pela Organização Mundial da Saúde (OMS) no dia 11 de março de 2020. No Brasil, desde o primeiro caso, confirmado em 26 de fevereiro, foram registrados outros 374.898, e 23.485 óbitos atestados até 1º de junho de 2020.

Uma importante questão epidemiológica diz respeito à elevada infectividade do SARS-CoV-2 (sigla em inglês para severe acute respiratory syndrome coronavirus 2), agente etiológico da COVID-19, cuja velocidade de propagação pode variar de 1,6 a 4,1. A elevada infectividade do SARS-CoV-2 e a ausência de uma vacina contra esse vírus fazem com que o aumento do número de casos seja exponencial.

Em função da inexistência de medidas preventivas ou terapêuticas específicas para a COVID-19, e sua rápida taxa de transmissão e contaminação, a OMS recomendou aos governos a adoção de intervenções não farmacológicas (INF), as quais incluem medidas de alcance individual (lavagem das mãos, uso de máscaras e restrição social), ambiental (limpeza rotineira de ambientes e superfícies) e comunitário (restrição ou proibição ao funcionamento de escolas e universidades, locais de convívio comunitário, transporte público, além de outros espaços onde pode haver aglomeração de pessoas). Entre todas, destaca-se a restrição social.

No Brasil, diversas medidas foram adotadas pelos estados e municípios, como o fechamento de escolas e comércios não essenciais. Trabalhadores foram orientados a desenvolver suas atividades em casa, alguns municípios e estados encerraram-se em seus limites e divisas. Autoridades públicas locais chegaram a decretar bloqueio total (lockdown), com punições para estabelecimentos e indivíduos que não se adequassem às normativas. A restrição social resulta ser a medida mais difundida pelas autoridades, e a mais efetiva para evitar a disseminação da doença e achatar a curva de transmissão do coronavírus. Geralmente, a repercussão clínica e comportamental dessa obrigação implica mudanças no estilo de vida e pode afetar a saúde mental dos cidadãos.

Em relação aos estilos de vida, a restrição social pode levar a uma redução importante nos níveis de atividade física de intensidade moderada a vigorosa, e no aumento de tempo em comportamento sedentário. Nos Estados Unidos, observou-se um aumento no hábito de assitir à televisão (TV) e internet entre adultos durante a pandemia. Resultados semelhantes foram

identificados na Itália e na Espanha, tanto na participação em transmissões ao vivo, pelas redes sociais, quanto no aumento na instalação de aplicativos de programação de TV.

Outra preocupação refere-se à alteração dos hábitos alimentares. Nos Estados Unidos, no início da pandemia, observou-se um crescimento no volume de compras em supermercados e estoque doméstico de alimentos ultraprocessados e de alta densidade energética, como batatas fritas, pipoca, chocolate e sorvete. Adicionalmente, estudos indicam aumento no consumo de álcool, isoladamente, e no consumo associado de álcool e tabaco, durante a quarentena.

A adoção bem-sucedida de restrição social como medida de Saúde Pública traz comprovados benefícios à redução da taxa de transmissão da COVID-19; entretanto, efeitos negativos, associados a essa restrição, poderão ter consequências para a saúde, no médio e longo prazo. Portanto, espera-se das ações de Saúde Pública, também, uma capacidade de minimizar os efeitos adversos da restrição social prolongada.

A.2 Resumo gerado pelo Algoritmo de Luhn

Em função da inexistência de medidas preventivas ou terapêuticas específicas para a COVID-19, e sua rápida taxa de transmissão e contaminação, a OMS recomendou aos governos a adoção de intervenções não farmacológicas (INF), as quais incluem medidas de alcance individual (lavagem das mãos, uso de máscaras e restrição social), ambiental (limpeza rotineira de ambientes e superfícies) e comunitário (restrição ou proibição ao funcionamento de escolas e universidades, locais de convívio comunitário, transporte público, além de outros espaços onde pode haver aglomeração de pessoas). A restrição social resulta ser a medida mais difundida pelas autoridades, e a mais efetiva para evitar a disseminação da doença e achatar a curva de transmissão do coronavírus. Em relação aos estilos de vida, a restrição social pode levar a uma redução importante nos níveis de atividade física de intensidade moderada a vigorosa, e no aumento de tempo em comportamento sedentário. Nos Estados Unidos, observou-se um aumento no hábito de assitir à televisão (TV) e internet entre adultos durante a pandemia. A adoção bem-sucedida de restrição social como medida de Saúde Pública traz comprovados benefícios à redução da taxa de transmissão da COVID-19; entretanto, efeitos negativos, associados a essa restrição, poderão ter consequências para a saúde, no médio e longo prazo.

A.3 Resumo gerado pelo Algoritmo *Gistsumm*

A pandemia da doença pelo coronavírus 2019, COVID-19 (sigla em inglês para coronavírus disease 2019) foi reconhecida pela Organização Mundial da Saúde (OMS) no dia 11 de março de 2020. Uma importante questão epidemiológica diz respeito à elevada infectividade do SARS-CoV-2 (sigla em inglês para severe acute respiratory syndrome coronavirus 2), agente etiológico da COVID-19, cuja velocidade de propagação pode variar de 1,6 a 4,1. Autoridades públicas locais chegaram a decretar bloqueio total (lockdown), com punições para estabelecimentos e indivíduos que não se adequassem às normativas. Nos Estados Unidos, observou-se um aumento no hábito de assitir à televisão (TV) e internet entre adultos durante a pandemia. A adoção bem-sucedida de restrição social como medida de Saúde Pública traz comprovados benefícios à redução da taxa de transmissão da COVID-19; entretanto, efeitos negativos, associados a essa restrição, poderão ter consequências para a saúde, no médio e longo prazo.

A.4 Resumo gerado pelo Algoritmo Programação Linear Inteira

A pandemia da doença pelo coronavírus 2019, COVID-19 (sigla em inglês para coronavírus disease 2019) foi reconhecida pela Organização Mundial da Saúde (OMS) no dia 11 de março de 2020. Uma importante questão epidemiológica diz respeito à elevada infectividade do SARS-CoV-2 (sigla em inglês para severe acute respiratory syndrome coronavirus 2), agente etiológico da COVID-19, cuja velocidade de propagação pode variar de 1,6 a 4,1. No Brasil, diversas medidas foram adotadas pelos estados e municípios, como o fechamento de escolas e comércios não essenciais. Nos Estados Unidos, observou-se um aumento no hábito de assitir à televisão (TV) e internet entre adultos durante a pandemia. A adoção bem-sucedida de restrição social como medida de Saúde Pública traz comprovados benefícios à redução da taxa de transmissão da COVID-19; entretanto, efeitos negativos, associados a essa restrição, poderão ter consequências para a saúde, no médio e longo prazo.

A.5 Resumo gerado pelo Algoritmo de Regressão Bayesiana

Em função da inexistência de medidas preventivas ou terapêuticas específicas para a COVID-19, e sua rápida taxa de transmissão e contaminação, a OMS recomendou aos governos a adoção de intervenções não farmacológicas (INF), as quais incluem medidas de alcance individual (lavagem das mãos, uso de máscaras e restrição social), ambiental (limpeza rotineira

de ambientes e superfícies) e comunitário (restrição ou proibição ao funcionamento de escolas e universidades, locais de convívio comunitário, transporte público, além de outros espaços onde pode haver aglomeração de pessoas). A restrição social resulta ser a medida mais difundida pelas autoridades, e a mais efetiva para evitar a disseminação da doença e achatar a curva de transmissão do coronavírus. Em relação aos estilos de vida, a restrição social pode levar a uma redução importante nos níveis de atividade física de intensidade moderada a vigorosa, e no aumento de tempo em comportamento sedentário. Nos Estados Unidos, observou-se um aumento no hábito de assitir à televisão (TV) e internet entre adultos durante a pandemia. A adoção bem-sucedida de restrição social como medida de Saúde Pública traz comprovados benefícios à redução da taxa de transmissão da COVID-19; entretanto, efeitos negativos, associados a essa restrição, poderão ter consequências para a saúde, no médio e longo prazo.

A.6 Resumo gerado pelo *ChatGPT*

O texto discorre sobre a pandemia da COVID-19, destacando a elevada infectividade do SARS-CoV-2 e a ausência de uma vacina específica como fatores que contribuem para o aumento exponencial do número de casos. A Organização Mundial da Saúde (OMS) recomendou medidas preventivas não farmacológicas (INF), incluindo a restrição social, como forma de evitar a disseminação da doença. No Brasil, diversas medidas foram adotadas pelos estados e municípios, com destaque para a restrição social. A restrição social pode afetar a saúde mental e os estilos de vida dos cidadãos, levando a uma redução nos níveis de atividade física e aumento do comportamento sedentário, bem como à alteração dos hábitos alimentares. Nos Estados Unidos, houve um aumento no hábito de assistir TV e internet e no consumo de alimentos ultraprocessados e de alta densidade energética, como batatas fritas, pipoca, chocolate e sorvete. Embora a restrição social seja uma medida efetiva para reduzir a transmissão da COVID-19, pode ter consequências negativas para a saúde no médio e longo prazo. Portanto, é importante que as ações de Saúde Pública também considerem a minimização dos efeitos adversos da restrição social prolongada.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA
Campus Campina Grande
R. Tranqüilino Coelho Lemos, 671, Dinamérica, CEP 58432-300, Campina Grande (PB)
CNPJ: 10.783.898/0003-37 - Telefone: (83) 2102.6200

Documento Digitalizado Ostensivo (Público)

Projeto de Conclusão de Curso

Assunto: Projeto de Conclusão de Curso
Assinado por: Micael Marques
Tipo do Documento: Projeto
Situação: Finalizado
Nível de Acesso: Ostensivo (Público)
Tipo do Conferência: Cópia Simples

Documento assinado eletronicamente por:

- **Micael Marques, ALUNO (201811250019) DE BACHARELADO EM ENGENHARIA DE COMPUTAÇÃO - CAMPINA GRANDE**, em 21/06/2023 19:08:54.

Este documento foi armazenado no SUAP em 21/06/2023. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 860811
Código de Autenticação: 81c5feac68