Universidade Federal de São Carlos– UFSCar
Centro de Ciências Exatas e de Tecnologia– CCET
Departamento de Computação– DC
Programa de Pós-Graduação em Ciência da Computação– PPGCC

**Leonardo Capellaro**

# ToMAS: Sumarização Abstrativa Multinível Baseada em Tópicos usando LLMs

São Carlos
2025

**Leonardo Capellaro**

# ToMAS: Sumarização Abstrativa Multinível Baseada em Tópicos usando LLMs

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Centro de Ciências Exatas e de Tecnologia da Universidade Federal de São Carlos, como parte dos requisitos para a obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Aprendizado de Máquina e Processamento de Línguas Naturais

Orientador: Profa. Dra. Helena de Medeiros Caseli

São Carlos
2025

*Este trabalho é dedicado à minha família, namorada, amigos e a todos que me apoiaram e me deram suporte durante meu percurso.*

# Agradecimentos

Gostaria de expressar meus agradecimentos ao Departamento de Computação da UFSCar por me proporcionar a oportunidade e a infraestrutura necessárias para a realização do meu mestrado.

Além disso, quero estender meus agradecimentos à Profa. Dra. Helena de Medeiros Caseli, minha orientadora, pela orientação excepcional e pela paciência incansável que demonstrou ao longo de todo o processo de desenvolvimento da pesquisa. Suas contribuições e insights foram fundamentais para o sucesso deste trabalho.

Gostaria de agradecer também ao grupo Interfaces pela disponibilização do *corpus* utilizado neste trabalho e também ao integrante Ian Victor Rubini Ruiz, que nos ajudou na escolha dos tópicos a serem avaliados e forneceu uma descrição para cada um deles.

Agradeço também a todos os colegas, amigos e familiares que me apoiaram e encorajaram ao longo desta jornada. Suas palavras de incentivo e apoio emocional foram fundamentais para a minha motivação e determinação.

Este trabalho não teria sido possível sem o apoio dessas pessoas e instituições, e estou profundamente grato a todos.

*"Seu trabalho vai preencher uma parte grande da sua vida, e a única maneira de estar verdadeiramente satisfeito é fazer o que você acredita ser um ótimo trabalho. E a única maneira de fazer um ótimo trabalho é amar o que você faz."*

*(Steve Jobs)*

# Resumo

Recentemente, as discussões políticas no Brasil ganharam destaque, tornando-se um dos tópicos mais debatidos nas redes sociais. Fatores como a diversidade de temas, a ocorrência de eventos que atraem a atenção do público e o constante aumento no volume de mensagens tornaram desafiadora a tarefa de identificar de forma clara, concisa e objetiva os principais tópicos das postagens. Nesse contexto, este estudo propõe um novo método automático que explora o uso de técnicas automáticas de geração de tópicos e sumarização em múltiplos níveis utilizando modelos de linguagem de grande escala. O método proposto foi avaliado para gerar resumos (sumários) a partir de *tweets* sobre o domínio da política brasileira coletados durante as eleições presidenciais de 2022. Além do método de sumarização multinível, foi proposto também um novo método de avaliação de sumários baseado na estratégia de divisão e conquista para a aplicação da medida de avaliação automática BERTScore em textos extensos, o qual também agrega o tamanho da sentença gerada como um peso no valor da avaliação. Análises qualitativas e quantitativas indicam que a combinação dessas técnicas foi capaz de extrair e resumir os principais tópicos com sucesso, demonstrando um grande potencial para ser uma ferramenta informativa útil na avaliação de diferentes opiniões, questões e temas discutidos publicamente.

**Palavras-chave:** LLM. Sumarização. Geração de tópicos. Twitter. X. Política Brasileira.

# Abstract

Recently, political discussions in Brazil have gained prominence, becoming one of the most debated topics on social media. Factors such as the diversity of themes, the occurrence of events that attract public attention, and the constant increase in the volume of messages have made it challenging to identify in a clear, concise, and objective manner the main topics of the posts. In this context, this study proposes a new automated method that explores the use of automatic topic generation and multi-level summarization techniques using large-scale language models. The proposed method was evaluated to generate summaries from tweets about Brazilian politics collected during the 2022 presidential elections. In addition to the multi-level summarization method, a new summary evaluation method was also proposed, based on the divide and conquer strategy for applying the BERTScore automatic evaluation measure to lengthy texts, which also incorporates the generated sentence size as a weight in the evaluation score. Qualitative and quantitative analyses indicate that the combination of these techniques was able to successfully extract and summarize the main topics, demonstrating great potential to be a useful informative tool in the assessment of different opinions, issues, and topics publicly discussed.

**Keywords:** LLM. Summarization. Topic generation. Twitter. X. Brazilian Politics.

# Lista de ilustrações

# Lista de tabelas

# Lista de siglas

**BERT** Bidirectional Encoder representations from Transformers

**CBOW** Continuous Bag of Words

**CoT** chain-of-thought

**GPT** Generative Pre-trained Transformer

**HDBSCAN** Hierarchical Density-based spatial clustering of applications with noise

**LLM** Large Language Models

**MST** Minimum Spanning Tree

**NLM** Neural Language Models

**PCA** Principal Components Analysis

**PLM** Pre-trained Language Models

**PLN** Processamento de Linguagem Natural

**PTQ** Post-Training Quantization

**QAT** Quantization-Aware Training - QAT

**ROUGE** Recall-Oriented Understudy for Gisting Evaluation

**SLM** Statistical Language Models

**UMAP** Uniform Manifold Approximation and Projection

# Sumário

# Capítulo 1

# Introdução

O tema da política tem se tornado um dos principais assuntos discutidos em redes sociais na última década. De acordo com um levantamento feito pelo Twitter/X (TWITTER, 2022), do dia 01 de janeiro de 2022 até o dia 29 de setembro de 2022, período que precedeu as eleições presidenciais brasileiras realizadas em 02 de outubro de 2022, mais de 100 milhões de *tweets* sobre o tema haviam sido contabilizados, o que tornou essas eleições as mais *tweetadas* do mundo até agora. Durante o período oficial de campanha eleitoral, quase 45 milhões de *tweets* diretamente relacionados às eleições de 2022 foram contabilizados.

Uma das principais alavancas que levaram a esse engajamento é a polarização política. De acordo com o estudo proposto por Machado e Miskolci (2019), os conflitos políticos se devem a uma grande polarização entre vertentes políticas, no caso esquerda e direita. Os autores citam que esta não é uma ocorrência recente, tendo raízes na internet desde a década de 1990. Entretanto, com a consolidação do oligopólio de empresas de tecnologia, como o Facebook e o Twitter, ocorreu uma unificação de perfis de usuários, criando um ambiente propenso a incompreensões e conflitos. Os autores também dizem que o funcionamento das redes sociais propicia a criação de bolhas de opinião que se baseiam em consensos e se opõem a outras bolhas, expondo diferenças e levando a atritos. Como ponto de inflexão para a polarização, os autores citam as Jornadas de Junho de 2013[1], que marcaram uma virada e uma ascensão da polarização de opiniões nas redes sociais.

Diversos autores realizaram análises focadas em distinguir cada um dos grupos nas

[1] As Jornadas de Junho de 2013 no Brasil foram marcadas por manifestações em todo o país, inicialmente motivadas pelo aumento das tarifas de transporte público. O Movimento Passe Livre (MPL) liderou a luta, que transcendeu a questão tarifária, revelando problemas como a precarização do mercado de trabalho e o descontentamento com os gastos públicos, especialmente relacionados à Copa do Mundo de 2014.

redes sociais, como em Santos et al. (2023) que analisaram a disseminação de conteúdo hostil no Twitter entre julho e agosto de 2022, mais especificamente relacionado à credibilidade do sistema eleitoral brasileiro, e Interian e Rodrigues (2023) que realizaram um trabalho sobre a contribuição de grupos específicos para a polarização das redes sociais por meio da modularidade de redes[2].

Um marco recente no cenário político brasileiro que reverbera até os dias atuais foram os atos antidemocráticos de 8 de janeiro de 2023. Eles foram caracterizados pela invasão de prédios governamentais em Brasília, capital do país, por apoiadores do ex-presidente Jair Bolsonaro. As invasões ocorreram logo após a confirmação da vitória do presidente Luiz Inácio Lula da Silva. Esses atos antidemocráticos são considerados a maior ameaça à democracia brasileira desde a transição democrática na década de 1980. Os manifestantes superaram as forças de segurança, invadiram o Congresso, o Supremo Tribunal Federal e o Palácio do Planalto, vandalizando escritórios e destruindo obras de arte, enquanto documentavam e divulgavam suas ações nas redes sociais (RAMOS, 2023).

Garrossini et al. (2023) citam a manipulação das redes sociais como um ponto crucial para possibilitar os atos antidemocráticos, já que essas plataformas foram usadas como ferramentas poderosas para a propagação de ideologias extremistas que desempenharam um papel significativo na organização e mobilização dos ataques. A desinformação também desempenhou um papel fundamental, com a disseminação de *fake news* servindo como instrumento de caos político (SOUZA; LEAL, 2023). Dada a relevância dos atos antidemocráticos de 08/01/2023 e a grande quantidade de *tweets* relativos a esses atos, este foi o momento da história brasileira selecionado como estudo de caso neste trabalho. Como os *tweets* são compostos por textos, para processá-los foram usadas técnicas automáticas de processamento de linguagem natural e inteligência artificial.

Em especial, o presente trabalho utilizou os Large Language Models (LLMs), que possuem as chamadas "habilidades emergentes" (WEI et al., 2022b), adquirindo a capacidade de resolver tarefas complexas para as quais o modelo não foi explicitamente treinado, e, em muitos casos, demonstrando habilidades de resolução de tarefas que envolvem algum nível de raciocínio lógico (WEI et al., 2023a).

Com a ascensão dos modelos de linguagem de larga escala e com o crescente número de assuntos e tópicos abordados nas mensagens das redes sociais, a tarefa de obter quais são os tópicos mais influentes sendo abordados no momento e sumarizar o conteúdo das mensagens que se referem a cada tópico se torna atraente. Entretanto, essa tarefa apresenta diversos níveis de complexidade.

Quando abordamos a natureza dos *tweets*, destacam-se características distintivas em

[2] A modularidade é uma métrica que avalia a concentração de conexões dentro de grupos em relação a uma distribuição aleatória. Ela quantifica a estrutura organizacional da rede, destacando a tendência de nós se agruparem mais fortemente do que seria esperado ao acaso. Essa medida é fundamental para compreender como os elementos de uma rede estão organizados em comunidades ou grupos coesos, revelando padrões de interconexão que diferem de uma disposição totalmente aleatória.

relação a outros tipos de textos. A limitação de 280 caracteres por postagem é uma característica fundamental que impõe uma restrição significativa na quantidade de informações transmitidas em cada mensagem. Consequentemente, os *tweets* são notavelmente mais concisos e diretos, frequentemente recorrendo a abreviações e símbolos para economizar espaço e possuindo uma sintaxe única em determinadas palavras, como "RT"[3] (CASTELLUCCI et al., 2013). Além disso, uma diferença marcante reside na dinâmica dos tópicos abordados. O conteúdo no Twitter é extremamente dinâmico, tendendo a se adaptar rapidamente às notícias e eventos do momento, com opiniões em constante evolução e uma profunda diversidade de perspectivas que divergem entre grupos de usuários (SANKARANARAYANAN et al., 2009). Assim, além das distinções gramaticais, emerge a complexidade de inúmeras opiniões sobre um mesmo tópico, as quais frequentemente mudam ao longo do tempo, ao contrário do conteúdo estático encontrado em textos mais tradicionais, como artigos ou notícias.

Devido ao alto volume de *tweets* analisados neste trabalho, foram empregadas técnicas de sumarização automática para obter um resumo do que estava sendo discutido. Mani (2001) cita a sumarização como o processo cujo objetivo é produzir uma representação condensada de uma determinada entrada para consumo humano.

Existem diversas técnicas de sumarização aplicadas em textos na literatura, as quais podem ser resumidas em duas abordagens principais: sumarização extrativa e sumarização abstrativa. A sumarização extrativa possui a característica de gerar sumários compostos por sentenças dos textos-fonte. Assim, o conteúdo não passa por um processo de edição linguística, mas é selecionado a partir de determinados critérios, como o tamanho disponível do sumário e a importância do conteúdo, por exemplo. Por outro lado, a sumarização abstrativa é caracterizada por gerar sumários cujo conteúdo, seja parcial ou integral, é submetido a alguma operação linguística de reescrita, como a paráfrase (SOUZA; CARDOSO; PAIXãO, 2024).

Alguns pesquisadores empregaram redes neurais recorrentes (LUCKY; SUHARTONO, 2021) e redes neurais convolucionais (NARAYAN; COHEN; LAPATA, 2019) para realizar sumarizações abstrativas. Entretanto, é importante destacar que esses estudos lidaram com textos contínuos e abordaram um único tópico por vez. Quando se trata de aplicar a sumarização abstrativa a *tweets*, surgem desafios adicionais devido à natureza não linear e concisa desses textos. Além disso, os *tweets* frequentemente incorporam diferentes estilos de escrita e expressam perspectivas diversas de cada autor em relação ao tema tratado.

Trabalhos mais recentes demonstraram o grande potencial do uso de LLMs em tarefas de sumarização (GOYAL; LI; DURRETT, 2023; SHEN et al., 2023). Foi observado um desempenho similar entre o sumário gerado por um LLM comparado a sumários gerados por humanos, apesar de que, quando comparado detalhadamente, o sumário gerado pelo

[3] *Retweet* é a ação de compartilhar, na rede social *Twitter* (X), uma publicação (*tweet*) de outro usuário para que seus próprios seguidores também a vejam. Na prática, funciona como repassar um conteúdo, mantendo a autoria original.

modelo de linguagem se mostrou um pouco mais extrativo do que o gerado por humanos, ou seja, com menos parafraseamentos (ZHANG et al., 2023; SHEN et al., 2023).

Para tratar os desafios da sumarização de *tweets*, o presente trabalho propõe um *pipeline* de processamento visando obter a sumarização dos principais tópicos abordados em *tweets*. A sumarização proposta neste trabalho é feita através de um processo multinível, onde os *tweets* são organizados em blocos e sumarizados recursivamente, até que haja somente um único sumário no final do processo. Esse pode ser considerado um ponto de inovação deste trabalho, visto que, apesar de existirem processos de divisão e conquista na sumarização, como o *Map Reduce*[4] do *Langchain*, não foram encontrados trabalhos na literatura que realizaram o uso dele no contexto de *tweets*. A este método de sumarização abstrativa multinível deu-se o nome de ToMAS (**To**pic-based **M**ultilevel **A**bstractive **S**umarization). Dentre as etapas macro desse *pipeline*, estão: (i) a geração de representações vetoriais das palavras dos *tweets* por meio de *embeddings*; (ii) a redução de dimensionalidade espacial dessas representações; (iii) a aplicação de algoritmos de agrupamento dos textos similares; e (iv) a aplicação de *large language models* (LLMs) para a tarefa de sumarização abstrativa desses textos.

## 1.1 Objetivos e Hipóteses

Neste contexto, o presente trabalho visa gerar automaticamente sumários contendo os principais assuntos sendo abordados em um grupo de *tweets*. Embora a estratégia desenvolvida neste trabalho foque especificamente em *tweets* no domínio da política brasileira, entendemos que ela pode facilmente ser estendida para outros domínios e outras plataformas de redes sociais com características semelhantes.

Devido à natureza mutável e diversa dos *tweets*, havendo opiniões distintas e diversas linhas de conversa tratando de um mesmo assunto, foi adotada como **hipótese** que a realização de uma extração de tópicos e uma subsequente sumarização destes tópicos utilizando LLMs seriam métodos eficazes para obter de forma clara e objetiva quais são os principais temas sendo tratados sobre um determinado conjunto de *tweets*, permitindo a geração de *insights* e uma possível identificação de *fake news* de forma rápida. Esta busca pela identificação dos temas principais sendo tratados nos *tweets* é a principal motivação para o uso da extração de tópicos e a posterior sumarização de cada tópico ao invés de sumarizar o conjunto inicial diretamente.

Assim, este projeto visa responder à seguinte questão de pesquisa:

**QP1** Utilizar um *pipeline* que engloba extração de tópicos e sumarização abstrativa usando LLM é uma estratégia efetiva para a identificação dos principais assuntos tratados de forma clara e objetiva em *tweets*?

[4] <https://js.langchain.com/v0.1/docs/modules/chains/document/map_reduce/>

Para responder à QP1, foi utilizada uma abordagem que consiste em realizar uma extração de tópicos dos *tweets* antes de fazer a sumarização. Após a realização de diversos pré-processamentos para remover ruídos dos *tweets*, é aplicado um algoritmo de extração de tópicos, que em seus passos gera as representações vetoriais de cada *tweet* por meio do mecanismo de *embeddings* do BERT. Em seguida, o algoritmo de agrupamento hierárquico HDBSCAN (CAMPELLO; MOULAVI; SANDER, 2013) é aplicado para separar os *tweets* em diferentes grupos com base em sua similaridade. Por último, o algoritmo de TF-IDF é aplicado para obter as palavras mais representativas de cada grupo, gerando assim os tópicos e rotulando os *tweets* pertencentes ao grupo gerado como sendo integrantes do tópico extraído. Isso resulta na organização de *tweets* que compartilham uma escrita semelhante e abordam um mesmo assunto em grupos iguais. Posteriormente, são selecionados os vinte tópicos mais populosos de cada uma das bases de *tweets* para a realização da sumarização abstrativa, o que se assemelha a uma tarefa de sumarização multidocumento.

Além da QP1, outra questão de pesquisa relacionada à forma de avaliação dos sumários também foi proposta. Uma vez que avaliar manualmente sumários gerados a partir de grupos de mais de mil *tweets* é uma tarefa complexa, a seguinte QP2 foi definida:

**QP2** Utilizar um método de avaliação automático baseado no BERTScore e na estratégia de divisão e conquista, que leva em consideração o sumário gerado, seu comprimento e o texto original como referência é uma forma efetiva de avaliar sumários automáticos gerados por modelos de linguagem?

## 1.2 Organização da monografia

Este texto está organizado como segue. No Capítulo 2 são descritos os principais conceitos teóricos sobre agrupamento hierárquico, Transformers e LLMs. Em seguida, o Capítulo 3 traz um apanhado geral do uso de agrupamentos para sumarização e o uso de LLMs em tarefas de sumarização abstrativa. O Capítulo 4 apresenta a proposta para o *pipeline* de sumarização multinível denominado ToMAS. No Capítulo 5 são discutidos resultados de experimentos. Por fim, o Capítulo 6 finaliza o trabalho com uma conclusão geral sobre os resultados e a performance do método proposto.

# Capítulo 2

# Fundamentação Teórica

Neste capítulo são abordados os conceitos teóricos que fundamentaram este trabalho, como conceitos de sumarização abstrativa (Seção 2.1), representação vetorial de sentenças (Seção 2.1.1), agrupamento e extração de tópicos (Seção 2.2), LLMs (Seção 2.3) e a estratégia adotada para a proposta do modelo de sumarização abstrativa multinível que é a de "dividir para conquistar" (Seção 2.4).

## 2.1 Sumarização abstrativa

A sumarização automática é o processo que visa condensar um conjunto de dados textuais automaticamente, tendo por objetivo criar um subconjunto, neste caso chamado sumário, que contempla os pedaços mais importantes de informação do texto original (CAO, 2022). Diversos autores dividem os métodos de sumarização em dois tipos principais: a sumarização extrativa, que busca sentenças do texto original e faz uma concatenação das mais representativas com base na probabilidade; e a sumarização abstrativa (CAO, 2022; HOŠOVSKÝ et al., 2022; LIU; HEALEY, 2023), utilizada neste trabalho.

A sumarização abstrativa é um processo que envolve a reescrita de um texto original com o intuito de produzir um resumo conciso e coerente. Este processo não se limita somente a extrair frases ou partes do texto original, mas sim a entender o conteúdo todo e reescrevê-lo de uma forma mais resumida (GUPTA; GUPTA, 2019). De acordo com Lin e Ng (2019), tradicionalmente o processo de sumarização abstrativa pode ser dividido em três etapas principais: extração de informações, seleção de conteúdo e realização de superfície.

Na extração de informações, o objetivo é extrair as informações mais importantes do texto original. Isso pode envolver a extração de frases nominais e verbais, juntamente com

suas informações contextuais. Alguns métodos de sumarização abstrativa também podem usar a extração baseada em consultas, que visa extrair conteúdos importantes usando consultas geradas automaticamente e filtrar conteúdos que têm baixa probabilidade de serem incluídos no resumo.

A seleção de conteúdo é a próxima etapa, onde um subconjunto das frases candidatas extraídas na etapa anterior é selecionado para inclusão no resumo final. Isso geralmente é feito com base em restrições de comprimento. Alguns métodos usam abordagens heurísticas para selecionar as frases mais frequentemente mencionadas, enquanto outros usam a Programação Linear Inteira[1] (ILP) para otimizar uma função objetivo sujeita a um conjunto de restrições lineares.

A realização de superfície é a etapa final, onde as frases selecionadas na etapa de seleção de conteúdo são combinadas usando regras gramaticais/sintáticas para gerar um resumo. Isso pode ser feito usando um gerador de linguagem natural existente.

As abordagens para a sumarização abstrativa, conforme citado por Gupta e Gupta (2019), incluem:

- a abordagem baseada em estrutura, que identifica informações importantes e usa modelos e regras para criar resumos;
- a abordagem baseada em árvore, que organiza sentenças semelhantes em uma estrutura de árvore;
- a abordagem baseada em modelo, que extrai trechos de texto usando palavras-chave e os insere em modelos predefinidos;
- a abordagem baseada em ontologia, que usa uma estrutura de conhecimento para extrair informações e criar resumos personalizados;
- a abordagem baseada em grafo, que representa o texto como um grafo e visa remover informações redundantes;
- a abordagem baseada em regras, que usa regras e categorias para identificar informações importantes e criar o resumo; e
- a abordagem neural, que faz parte do campo de *deep learning* e utiliza múltiplas camadas de processamento não linear para extrair características do texto.

A abordagem neural para resumos de texto utiliza um modelo codificador-decodificador. O codificador é responsável por processar as sequências de entrada, enquanto o decodificador gera as sequências de saída. O codificador converte palavras em representações

[1] Programação Linear Inteira (PLI) é uma extensão da Programação Linear (PL) em que as variáveis de decisão são restritas a valores inteiros. Assim, além das restrições lineares presentes na PL, as soluções devem ser números inteiros.

vetoriais, capturando o contexto do texto. Geralmente, as palavras são representadas usando técnicas como incorporação de palavras (*word embeddings*) ou modelo de saco de palavras (*bag-of-words*). Por outro lado, o decodificador ajuda a determinar a próxima palavra no resumo, com base nas palavras anteriores. Quando tanto a entrada quanto a saída são sequências de texto, como na sumarização de texto, o problema é frequentemente referido como Seq2Seq (sequência para sequência). Modelos de codificador-decodificador são particularmente úteis para resolver esses tipos de problemas, pois podem capturar a estrutura de sequência.

A tarefa em foco neste projeto, que é a de gerar um sumário a partir de *tweets* que abordam um mesmo assunto, remete ao conceito de sumarização multi-documento. Essa sumarização consiste na criação de um sumário conciso e pequeno que contém informações relevantes a partir de um conjunto de documentos relatando o assunto (WOLHANDLER et al., 2022). Ao contrário de uma sumarização de documento único, no qual o texto a ser sumarizado é sequencial e corrido, a sumarização de múltiplos textos visa combinar e reunir as informações espalhadas por diversos textos distintos. Trabalhos como o de Wolhandler et al. (2022) e o de Liu e Healey (2023) citam a etapa de clusterização como crucial para a geração de sumários multi-documentos, utilizando-a para agrupar textos que se referem a um mesmo aspecto e posteriormente utilizando modelos de linguagem para realizar a sumarização abstrativa destes grupos, como modelos de LLM.

Quando tratamos da fonte dos dados utilizados na sumarização, os sumários podem ser classificados em mono ou multidocumento, conforme citado por Souza, Cardoso e Paixão (2024). Os sumários monodocumento são gerados a partir de um único texto que serviu de base para a sumarização, enquanto o processo de gerar sumários a partir de dois ou mais textos que dissertam sobre um mesmo assunto é chamado de sumarização multidocumento. Neste trabalho são gerados sumários de diversos *tweets* tratando sobre um mesmo tópico, portanto, a sumarização é considerada multidocumento.

Com relação à avaliação de sumários, diversos trabalhos (CAO, 2022; HOŠOVSKÝ et al., 2022; LIN; NG, 2019) utilizaram a métrica proposta por Lin (2004), a métrica Rouge. O Recall-Oriented Understudy for Gisting Evaluation (ROUGE) é uma métrica comumente usada para avaliar sistemas de sumarização automática. Ela calcula o grau de sobreposição de palavras ($n$-gramas) entre o resumo gerado e o texto original de referência.

Embora a Rouge tenha sido amplamente utilizado pela comunidade por anos, no caso da sumarização abstrativa ela não é a mais indicada uma vez que sumários abstrativos são gerados com palavras novas que em geral não se sobrepõe ao texto original. Por isso, o método Rouge Score vem sendo substituído por métodos vetoriais de comparação de textos. O BERTScore (ZHANG et al., 2020a) é um destes métodos. Diferentemente de outras métricas tradicionalmente aplicadas para avaliar resumos que se baseiam na correspondência de $n$-gramas entre o texto de referência e um candidato, o BERTScore avalia a similaridade semântica entre o texto gerado e o texto de referência, calculando

a similaridade de cosseno entre as *embeddings* dos *tokens* desses textos. Isso permite que o BERTScore supere as limitações de métodos baseados em $n$-gramas, incapazes de reconhecer paráfrases semânticas e de captar dependências distantes e mudanças na ordem das palavras.

Para entender como medidas de similaridade semântica funcionam, a Seção 2.1.1 traz uma breve descrição sobre representação vetorial de sentenças.

### 2.1.1 Representação vetorial de uma sentença

Como mencionado anteriormente, tarefas de Processamento de Linguagem Natural (PLN) demandam a criação de representações vetoriais de palavras ou até de sentenças completas, pois os algoritmos computacionais não conseguem processar palavras e símbolos (SENO et al., 2023). Isso remete a uma série de desafios teóricos e computacionais. Por exemplo, um modelo de significado de palavra deve ser capaz de identificar palavras com significados semelhantes, antônimos e conotações positivas. A representação de uma palavra como apenas uma sequência de letras ou um índice em uma lista de vocabulário se torna insuficiente neste cenário (JURAFSKY; MARTIN, 2009).

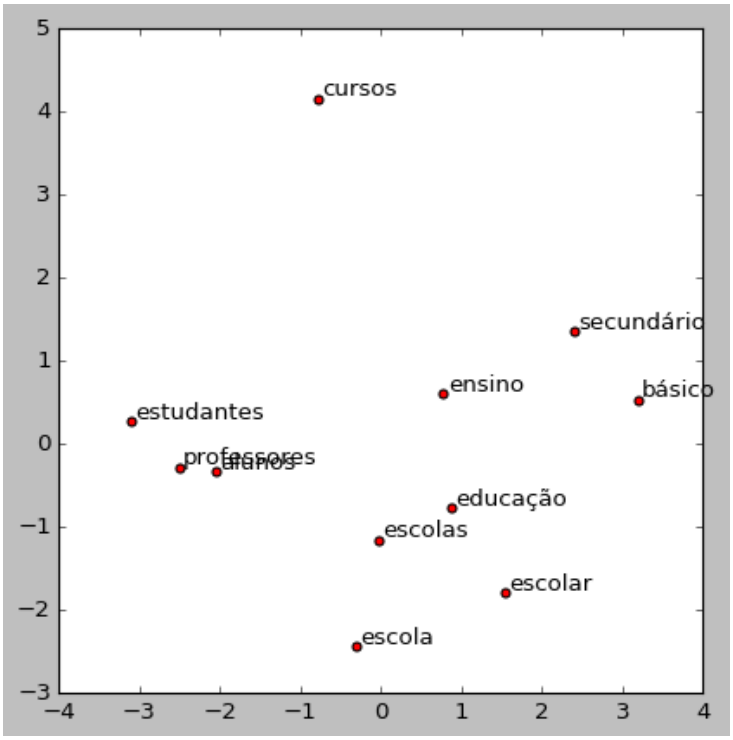
Dentre os desafios de representação, encontram-se as diferenças de significado para uma mesma palavra, como no caso de "manga" em português, que pode se referir à fruta ou à manga da camiseta. Isso remete ao conceito de lema, que é a forma principal das palavras, e à ideia de que os lemas podem ter diferentes significados, chamados de sentidos.Além disso, existe o conceito de sinonímia, que se refere à relação entre palavras que têm sentidos idênticos ou quase idênticos (JURAFSKY; MARTIN, 2009).

De acordo com Seno et al. (2023), a semântica distribucional tem sido a principal abordagem de representação do significado lexical de uma palavra. Utilizando essa abordagem, as palavras são representadas por meio de vetores de valores reais que codificam o significado da palavra de acordo com sua distribuição no texto. Esses vetores são chamados de vetores semânticos ou *embeddings*. A semântica distribucional se apoia na hipótese distribucional, inicialmente proposta por Firth (1935) e Harris (1954), que diz que palavras que possuem contexto linguístico semelhante tendem a ter significados similares.

Para exemplificar o espaço vetorial de uma palavra, Seno et al. (2023) utilizaram o modelo GloVe (PENNINGTON; SOCHER; MANNING, 2014) para representar vetorialmente a palavra "ensino". Na Figura 1 é possível notar que as palavras que compartilham um mesmo sentido semântico, como "estudantes", "ensino", "educação", estão próximas entre si no espaço vetorial.

Dado o contexto acima, as *embeddings* são representações numéricas contínuas de palavras ou unidades linguísticas em um espaço vetorial. Ao contrário de representações esparsas que usam contagens, as *embeddings* usam vetores densos com valores reais. Esses vetores têm dimensões menores, geralmente variando de 50 a 1.000, e suas dimensões não possuem uma interpretação clara (JURAFSKY; MARTIN, 2009).

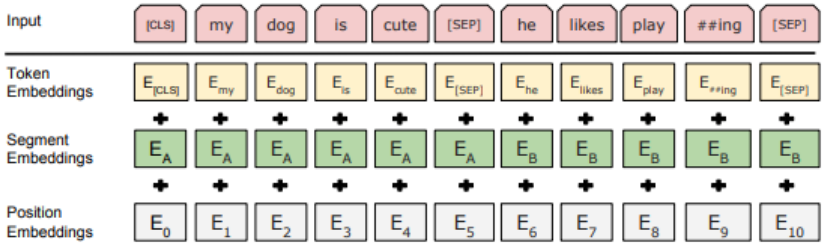
Figura 1 – Representação vetorial da palavra “ensino” utilizando o modelo GloVe




Fonte: Seno et al. (2023)

Alguns trabalhos passaram a abordar o uso de *embeddings* para uma sentença completa ao invés de *embeddings* de uma palavra única (KIROS et al., 2015; LOGESWARAN; LEE, 2018; DEVLIN et al., 2019). Na Figura 2 é possível ver uma representação de como o BERT trabalha com a geração de *embeddings* de uma sentença. O vetor de *embeddings* possui três componentes principais, onde o vetor final será uma soma das três componentes de cada palavra. O *token embedding* se refere a representação vetorial da palavra na sentença. O *segment embedding* se refere a qual segmento da sentença a palavra se refere. No exemplo, o trecho “my dog is cute” é considerado o segmento “A” e o trecho “he likes play” é considerado o segmento “B”. Por último, há a componente de posição, onde cada palavra na sentença está associada a uma posição distinta.

Figura 2 – Embeddings de sentença - BERT




Fonte: Devlin et al. (2019)

Por se tratarem de representações vetoriais, é possível medir a semelhança entre as

*embeddings* das sentenças por meio da similaridade do cosseno. O cosseno de dois vetores é calculado através do produto escalar, que é a soma dos produtos dos componentes correspondentes entre eles. Entretanto, por conta do produto escalar favorecer vetores mais longos nos cálculos, Jurafsky e Martin (2009) propõem normalizar o produto escalar dividindo-o pelos comprimentos dos vetores envolvidos. Isso resulta na métrica de similaridade do cosseno, que varia de -1 (para vetores apontando em direções opostas) a 1 (para vetores apontando na mesma direção), mas para valores não negativos, a métrica de cosseno varia de 0 a 1.

Assim, o cosseno de dois vetores $v$ e $w$ é obtido a partir do produto escalar entre eles:

$$\text{cosseno}(v, w) = \frac{v \cdot w}{\|v\|\|w\|} \tag{1}$$

O produto escalar (v.w) é a soma dos produtos dos componentes correspondentes dos dois vetores:

$$v \cdot w = \sum_{i=1}^{N} v_i w_i \tag{2}$$

A métrica de cosseno é normalizada para lidar com o problema de favorecer vetores mais longos:

$$\text{cosseno}(v, w) = \frac{\sum_{i=1}^{N} v_i w_i}{\sqrt{\sum_{i=1}^{N} v_i^2}\sqrt{\sum_{i=1}^{N} w_i^2}} \tag{3}$$

No presente trabalho, foram geradas *embeddings* de cada *tweet* utilizando o modelo BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), que é um modelo BERT treinado em português do Brasil utilizando o *corpus* brWaC (FILHO et al., 2018) contendo 2,68 bilhões de *tokens* provindos de 3,53 milhões de documentos em português do Brasil de domínios variados.

### 2.1.2 Avaliação da Sumarização usando BERTScore

Para calcular o BERTScore, primeiramente separamos os textos a serem avaliados em sentenças de referência e sentenças candidatas. As sentenças candidatas são as sentenças geradas pelo processo de sumarização automática, enquanto as sentenças de referência, em geral, são sentenças de sumários gerados por humanos. Entretanto, neste trabalho não foi possível obter as sentenças de referência da forma tradicional, ou seja, geradas por humanos. Isso ocorreu por conta da quantidade muito grande de *tweets* sumarizados, o que tornaria o processo de sumarização feito por um humano complexo e, em determinados casos, inviável. Dessa forma, neste trabalho adotou-se a estratégia alternativa de utilizar como sentenças de referência o conjunto de *tweets* original utilizado como *input* no ToMAS, conforme exemplificado na seção 4.2.4.1. Ambas as sentenças são tokenizadas e convertidas em *embeddings* usando o modelo BERT. A similaridade de *tokens* é então

computada usando a similaridade de cosseno entre os vetores de *embeddings*. Esse processo é realizado para cada *token* na sentença de referência em relação a todos os *tokens* na sentença gerada e vice-versa, permitindo uma correspondência flexível que considera o contexto e a semântica das palavras. O cálculo do BERTScore envolve os conceitos de precisão, recall e F1-Score. Os valores de BERTScore variam de 0 a 1, com valores próximos de 1 indicando maior similaridade entre o texto de referência e o texto gerado.

Para formalizar o cálculo do BERTScore, seja $x = [x_1, x_2, \ldots, x_k]$ a sentença de referência, e $\hat{x} = [\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_l]$ a sentença gerada pelo modelo. Ambas as sentenças são tokenizadas, e cada *token* é convertido em um vetor de *embeddings* que são geradas utilizando o modelo BERT, resultando nas sequências de vetores $[x_1, x_2, \ldots, x_k]$ e $[\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_l]$.

No passo seguinte é então calculada a similaridade entre os *tokens* usando a similaridade do cosseno entre os vetores de *embeddings* gerados anteriormente. A similaridade do cosseno entre um *token* da referência $x_i$ e um *token* gerado $\hat{x}_j$ é dada pela equação 4.

$$\text{sim}(x_i, \hat{x}_j) = \frac{x_i^\top \hat{x}_j}{|x_i||\hat{x}_j|}, \tag{4}$$

onde $x_i^\top \hat{x}_j$ representa o produto interno entre os vetores $x_i$ e $\hat{x}_j$, e $|x_i|$ e $|\hat{x}_j|$ são as normas dos vetores de *embedding* dos *tokens* $x_i$ e $\hat{x}_j$, respectivamente.

Como os vetores de *embedding* gerados pelo BERT são pré-normalizados, o cálculo da similaridade pode ser simplificado para o produto interno visto na equação 5.

$$\text{sim}(x_i, \hat{x}_j) = x_i^\top \hat{x}_j. \tag{5}$$

Para calcular o *recall* ($R_{\text{BERT}}$), cada *token* da sentença de referência $x_i$ é associado ao *token* da sentença gerada $\hat{x}_j$ que maximiza a similaridade, como pode ser visto na equação 6.

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \text{sim}(x_i, \hat{x}_j), \tag{6}$$

onde $|x|$ é o número de *tokens* na sentença de referência.

De forma análoga, para calcular a precisão ($P_{\text{BERT}}$), cada *token* da sentença gerada $\hat{x}_j$ é associado ao *token* da sentença de referência $x_i$ que maximiza a similaridade, de acordo com a equação 7.

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}j \in \hat{x}} \max_{x_i \in x} \text{sim}(x_i, \hat{x}_j), \tag{7}$$

onde $|\hat{x}|$ é o número de *tokens* na sentença gerada.

Por fim, o *F1-Score* ($F_{\text{BERT}}$) é calculado como a média harmônica entre a precisão e o *recall*, como na equação 8.

$$F_{\text{BERT}} = 2 \times \frac{P_{\text{BERT}} \times R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}. \tag{8}$$

Os valores de BERTScore variam de 0 a 1, sendo que valores próximos de 1 indicam maior similaridade entre o texto gerado e o texto de referência.

Neste trabalho, como não há resumos de referência gerados por humanos para os *tweets* avaliados, o conjunto de *tweets* originais dos tópicos resumidos foi usado como texto de referência. Embora este não seja o método de avaliação ideal, essa abordagem ainda pode fornecer uma métrica útil de similaridade semântica entre os resumos gerados e o texto original, o que permite avaliar a preservação do contexto nos resumos e obter uma avaliação relativa entre os resumos gerados por diferentes modelos. O modelo BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020) foi utilizado neste trabalho aliado a uma nova técnica de divisão e conquista proposta com o intuito de avaliar grandes textos e contornar a limitação de 512 *tokens* de entrada do modelo. Também foi proposto um novo cálculo de *score*, no qual o tamanho da sentença gerada é inserido como um peso no valor final. O detalhamento destes cálculos pode ser visto na Seção 4.2.4.

## 2.2 Agrupamento e Extração de Tópicos

Com o crescente e contínuo aumento na geração de dados, formas de organização e de geração de conhecimento a partir deles se tornam cada vez mais necessárias. Com o intuito de encontrar padrões, organizar os dados e prepará-los para uma análise ou aplicação em modelos, surgiram as técnicas de agrupamento, que fornecem um meio de explorar e verificar as estruturas existentes nos dados, buscando organizá-los de acordo com sua similaridade (JAIN; DUBES, 1988).

Nesta seção serão abordados os conceitos empregados no BERTopic (GROOTENDORST, 2020), método escolhido para extração de tópicos neste trabalho, que combina técnicas de geração de *embeddings* de palavras (Seção 2.1.1), redução de dimensionalidade (Seção 2.2.1), agrupamento hierárquico com base na densidade das palavras (Seção 2.2.2) e, finalmente, gera tópicos com base na importância das palavras usando uma variação da técnica TF-IDF, chamada c-TF-IDF (Seção 2.2.3).

### 2.2.1 Redução de dimensionalidade

No contexto de aprendizado de máquina e ciência de dados, é comum encontrar dados de alta dimensionalidade. Sinais de fala, imagens digitais e representações vetoriais de texto são exemplos de dados com essa característica. Frequentemente, para se tornar eficaz trabalhar com esses tipos de dados, é necessário reduzir sua dimensionalidade. A redução de dimensionalidade é a transformação de dados de alta dimensão em uma representação significativa e de dimensão reduzida (MAATEN et al., 2009). Essa nova representação idealmente deve refletir a verdadeira estrutura subjacente dos dados, conhecida como dimensionalidade intrínseca. A dimensionalidade intrínseca de um conjunto de dados representa o número mínimo de parâmetros necessários para descrever suas propriedades

observadas (FUKUNAGA, 2013). Tradicionalmente, a redução de dimensionalidade era realizada por meio de técnicas lineares como o Principal Components Analysis (PCA). Apesar dessas técnicas terem um bom resultado em cenários artificiais, elas não conseguiam lidar adequadamente com dados não lineares (FUKUNAGA, 2013).

Buscando resolver esses problemas, surgiu o algoritmo Uniform Manifold Approximation and Projection (UMAP), que é uma técnica não linear de redução de dimensionalidade. Esse algoritmo tem o objetivo de simplificar dados complexos, preservando suas estruturas essenciais e tornando-os mais fáceis de visualizar e analisar. O UMAP foi desenvolvido para funcionar bem com dados em que as distâncias entre pontos são importantes. Ele se destaca por possuir vantagens em termos de desempenho e escalabilidade comparado a algoritmos concorrentes, como o t-SNE, especialmente em grandes conjuntos de dados. Além disso, o UMAP pode ser usado como uma técnica geral de redução de dimensionalidade em aplicações de aprendizado de máquina (MCINNES; HEALY; MELVILLE, 2020).

Conforme McInnes, Healy e Melville (2020) explicam, o algoritmo UMAP pode ser entendido como uma técnica de redução de dimensionalidade que opera com grafos ponderados. Ele faz parte da classe de algoritmos de aprendizado baseados em grafos de k-vizinhos, como Laplacian Eigenmaps, Isomap e t-SNE. Sua abordagem pode ser dividida em duas fases principais, a construção do grafo ponderado e o layout do grafo.

Na primeira fase, a construção do grafo ponderado, é construído um grafo de k-vizinhos ponderado com base no conjunto de dados de entrada X, composto por pontos $\{x_1, x_2, \ldots x_n\}$. Dentre os passos desta etapa, estão:

- Definição dos vizinhos mais próximos: Para cada ponto $x_i$ no conjunto de dados, é calculado o conjunto de seus $k$ vizinhos mais próximos. Isso é feito utilizando uma métrica de distancia $d$ que mede a diferença entre os pontos.

- Cálculo de $\rho_i$ e $\sigma_i$: Para cada ponto $x_i$ são definidos dois valores, $\rho_i$ e $\sigma_i$. O $\rho_i$ é o valor mínimo da distância entre $x_i$ e seus $k$ vizinhos mais próximos, garantindo que $x_i$ esteja conectado a pelo menos um outro ponto com uma aresta de peso 1. O $\sigma_i$ é calculado para normalizar as distâncias entre os vizinhos de $x_i$.

$$\rho_i = \min\{d(x_i, x_{ij}) \,|\, 1 \leq j \leq k, d(x_i, x_{ij}) > 0\} \tag{9}$$

$$\sigma_i = \frac{\log(2)}{\sum_{j=1}^{k} \exp\left(-\max(0, d(x_i, x_{ij}) - \rho_i)/\sigma_i\right)} \tag{10}$$

- Construção do grafo ponderado: Posteriormente, um grafo direcionado ponderado é definido como $G\acute{\imath} = (V, E, w)$. Os vértices $V$ representam os pontos do conjunto de dados $X$, e as arestas $E$ são determinadas pelas conexões entre os pontos, sendo ponderadas com base nos valores de $\rho_i$ e $\sigma_i$ calculados na etapa anterior.

$$w((x_i, x_{ij})) = \exp\left(-\max(0, d(x_i, x_{ij}) - \rho_i)/\sigma_i\right) \tag{11}$$

O segundo passo do algoritmo UMAP é chamado Layout do Grafo. Nesta fase, o UMAP utiliza um algoritmo de layout de grafo direcionado por forças em um espaço de baixa dimensionalidade. Esse algoritmo aplica forças atraentes e repulsivas entre os vértices e arestas do grafo para posicionar os pontos em uma representação de baixa dimensão. Existem três etapas principais envolvidas neste processo, que são:

- Cálculo das forças atraentes entre os vértices $i$ e $j$: As forças atraentes entre dois vértices $i$ e $j$ são determinadas pela equação:

$$F_{\text{atrativa}}(i, j) = -\frac{2abk(y_i - y_j)}{2(b-1)(1 + k||y_i - y_j||^2)^{\frac{b-1}{2}}} \tag{12}$$

  Onde $a$ e $b$ são hiperparâmetros, $y_i$ e $y_j$ são as coordenadas dos vértices $i$ e $j$ no espaço de baixa dimensionalidade e $k$ é uma constante.

- Forças repulsivas: As forças repulsivas são calculadas através de amostragem devido a limitações computacionais. Quando uma força atraente é aplicada a uma aresta, um dos vértices da aresta é repelido por meio da amostragem.

- Inicialização do algoritmo: Embora o algoritmo possa ser iniciado aleatoriamente, a inicialização é frequentemente feita usando um layout espectral para obter convergência mais rápida e maior estabilidade dentro do algoritmo.

Neste trabalho, o UMAP é utilizado na redução da dimensionalidade das *embeddings* geradas pelo BERT de 1024 para 512 dimensões. Isso proporciona um menor custo computacional no processo de agrupamento, preservando as estruturas locais e globais dos dados.

### 2.2.2 Agrupamento Hierárquico

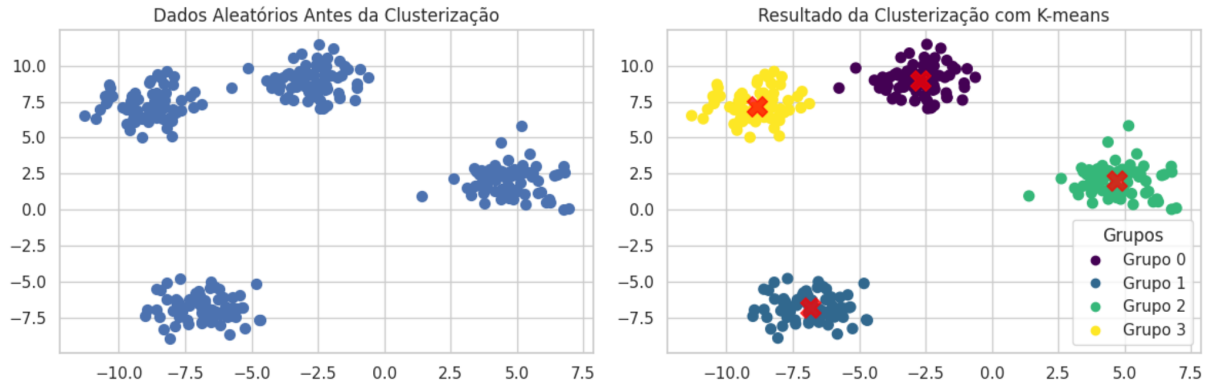
O agrupamento, ou *clustering* em inglês, é uma técnica de modelagem de dados. Nessa técnica, os dados de um determinado conjunto são divididos em grupos, de tal modo que elementos pertencentes a um mesmo grupo possuem uma maior semelhança entre si comparados aos elementos pertencentes a qualquer outro grupo dado um critério de similaridade (LACHI; ROCHA, 2005). Ester et al. (1996) cita que há dois principais tipos de algoritmos: os algoritmos de particionamento, como o k-means, e os algoritmos hierárquicos baseados em densidade.

Os algoritmos de particionamento dividem um conjunto de dados em k *clusters*, onde k é um parâmetro de entrada, requerendo conhecimento prévio do domínio. Eles usam uma abordagem de duas etapas, primeiro determinando representantes para os *clusters* e,

em seguida, atribuindo objetos a *clusters* com representantes próximos. Já os algoritmos hierárquicos criam uma decomposição hierárquica representada por um dendrograma. O agrupamento baseado em densidade é um paradigma de agrupamento popular. Ele busca utilizar métodos de aprendizado não supervisionado que detectam grupos distintos nos dados, com base na concepção de que um *cluster* em um espaço de dados é uma região contígua de alta densidade de pontos, separada de outros *clusters* semelhantes por regiões contíguas de baixa densidade de pontos. Os pontos de dados nas regiões separadoras de baixa densidade de pontos são geralmente considerados como *outliers*. (SANDER, 2010).

Por se tratar de uma técnica que busca gerar grupos sem possuir rótulos prévios dos dados, o agrupamento, em geral, é considerado uma técnica de aprendizado não supervisionado (SINAGA; YANG, 2020). Dessa forma, o algoritmo busca encontrar padrões através de medidas de dissimilaridade entre elementos dos dados e os separa de acordo com suas características. No exemplo da Figura 3 é possível ver o resultado de um agrupamento simples realizado utilizando o algoritmo *k-means*, que visa obter os n-vizinhos mais próximos dos elementos dos dados. No exemplo, é possível ver um conjunto de 300 elementos de dados com 4 centros principais. No gráfico à esquerda é possível observar os dados antes do agrupamento, e no gráfico à direita após o agrupamento. É possível ver que foram gerados quatro grupos (*clusters*) distintos após a aplicação do algoritmo.

Figura 3 – Representação gráfica de uma clusterização simples realizada por k-means




Fonte: Autor

Métodos anteriores ao método de agrupamento hierárquico apresentavam várias limitações. Alguns métodos, como o DBSCAN (ESTER et al., 1996), forneciam apenas uma rotulagem plana dos objetos de dados com base em um limite global de densidade. O uso de um único limite de densidade muitas vezes não é capaz de caracterizar adequadamente conjuntos de dados comuns com *clusters* de densidades muito diferentes.

Campello, Moulavi e Sander (2013) propuseram o Hierarchical Density-based spatial clustering of applications with noise (HDBSCAN) (CAMPELLO; MOULAVI; SANDER, 2013), que é um método de agrupamento hierárquico que gera uma hierarquia de agrupamento baseada em densidade. A partir desta hierarquia, uma hierarquia simplificada composta apenas pelos *clusters* mais significativos é extraída. Os autores também pro-

puseram uma nova medida de estabilidade de *cluster* para o propósito de extrair um conjunto de *clusters* significativos de diferentes níveis de uma árvore de *cluster* simplificada. O algoritmo original do HDBSCAN tem como único parâmetro de entrada um valor para *mpts*[2], que é um fator de suavização em estimativas de densidade cujo comportamento é bem compreendido. Entretanto, em diversas implementações este parâmetro é chamado de *min_samples* e é acompanhado de diversos outros parâmetros auxiliares, como o *min_cluster_size*, que especifica o tamanho mínimo necessário para que um *cluster* seja considerado válido; o *metric*, que especifica a métrica de distância a ser utilizada; o *cluster_selection_method*, que define como os *clusters* finais serão recortados da árvore condensada, podendo assumir os valores "*eom*" (*excess of mass*), onde o algoritmo escolhe automaticamente o ponto ótimo em cada ramo ou "*leaf*", onde são selecionados apenas os "nós folha", produzindo mais *clusters*, porém menores. Há diversos outros parâmetros que podem variar para cada implementação. Vale ressaltar que, quando não especificado, o valor de *min_samples* é o mesmo do *min_cluster_size*. Na Seção 4.2.2 são citados os valores de cada parâmetro utilizado neste trabalho.

No algoritmo 1, extraído do trabalho de Campello, Moulavi e Sander (2013) é possível compreender o funcionamento base do HDBSCAN.

O primeiro passo do algoritmo consiste em calcular a distância central (*core distance*) em relação a *mpts* para todos os objetos de dados em "X". A distância central é definida como a distância do objeto em questão até o seu vizinho mais próximo dentro do alcance especificado por *mpts*. É nessa etapa que são identificados os núcleos de cada grupo. Para cada objeto em "X", a sua distância central é calculada com relação ao valor *mpts*. Na Figura 4, gerada por McInnes John Healy (2016), é possível enxergar graficamente o cálculo da distância central de um ponto com relação aos outros objetos.

O segundo passo envolve o cálculo de uma árvore geradora mínima (Minimum Spanning Tree (MST)) do "Gmpts", que é o Grafo de Alcançabilidade Mútua. Esse grafo é construído considerando os objetos de dados como vértices, e os pesos das arestas representam a distância de alcançabilidade mútua entre os pares de objetos.

No terceiro passo, para obter a MST estendida (MSText), uma "auto-aresta" é adicionada para cada vértice, onde o peso da aresta é definido pela distância central do objeto correspondente. Essas auto-arestas refletem a densidade do próprio objeto e são consideradas ao remover arestas posteriormente. Na Figura 5 é possível enxergar um exemplo

[2] O *mpts* é um fator de suavização de uma estimativa de densidade realizada pelo algoritmo. A função desse parâmetro é ajustar a suavização da estimativa de densidade, afetando indiretamente a sensibilidade do algoritmo à densidade dos *clusters*. Este parâmetro influencia a forma como o algoritmo identifica e agrupa os pontos de dados com base em sua densidade no espaço.

[4] Uma MST de um grafo é um subconjunto de arestas que conecta todos os vértices do grafo sem formar ciclos e tem a menor soma de pesos possível

[5] Um grafo de alcançabilidade mútua é uma representação gráfica em que os nós representam os objetos, e as arestas são ponderadas pelas distâncias de alcançabilidade mútua entre essos objetos. A medida de alcançabilidade mútua indica a facilidade com que um ponto pode ser alcançado a partir de outro, considerando a densidade local

**Algoritmo 1** Passos principais do HDBSCAN

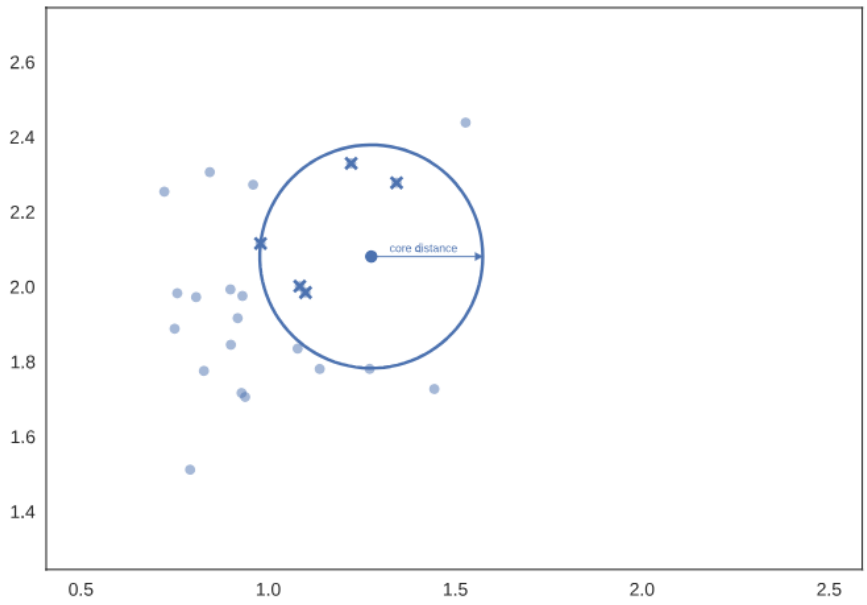
1. Calcular a distância central em relação a mpts para todos os objetos de dados em um conjunto de dados X.
2. Calcular uma MST[4] (Minimum Spanning Tree) de GMPTS[5] (Grafo de Alcançabilidade Mútua).
3. Estender a MST para obter MSText (Extended Minimum Spanning Tree), adicionando para cada vértice uma "aresta própria" com a distância central do objeto correspondente como peso.
4. Extrair a hierarquia HDBSCAN como um dendrograma da Minimum Spanning Tree estendida:
   4.1 Para a raiz da árvore, atribuir a todos os objetos o mesmo rótulo (único "*cluster*").
   4.2 Iterativamente, remover todas as arestas de MSText em ordem decrescente de pesos (em caso de empates, as arestas devem ser removidas simultaneamente):
      4.2.1 Antes de cada remoção, definir o valor da escala do dendrograma do nível hierárquico atual como o peso da(s) aresta(s) a ser(em) removida(s).
      4.2.2 Após cada remoção, atribuir rótulos ao(s) componente(s) conectado(s) que contém(êm) o(s) vértice(s) final(is) da(s) aresta(s) removida(s), para obter o próximo nível hierárquico: atribuir um novo rótulo de *cluster* a um componente se ele ainda tiver pelo menos uma aresta, caso contrário, atribuir-lhe um rótulo nulo ("ruído").

gerado através do HDBSCan por McInnes John Healy (2016).

No quarto passo, a hierarquia do HDBSCAN é extraída como um dendrograma a partir da MSText. Inicialmente, no nível mais alto da árvore, todos os objetos são atribuídos ao mesmo rótulo, formando um único grupo. Em seguida, de maneira iterativa, as arestas são removidas da MSText em ordem decrescente de pesos. Antes de cada remoção, o valor da escala do dendrograma no nível hierárquico atual é definido como o peso das arestas a serem removidas. Após cada remoção, rótulos são atribuídos aos componentes conectados que contêm os vértices finais das arestas removidas. Assim, é criado o próximo nível hierárquico, onde um novo rótulo de grupo é atribuído a um componente se ele ainda tiver pelo menos uma aresta; caso contrário, é atribuído o rótulo nulo, indicando um *outlier*. Na Figura 6 é possível enxergar de forma gráfica o dendograma gerado a partir da MSText.

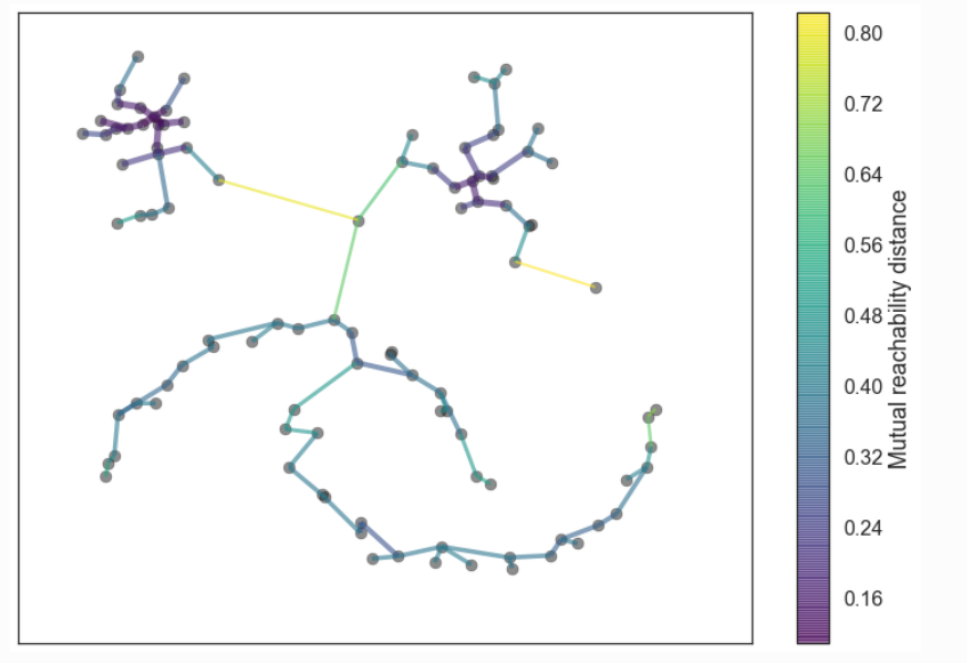
Esse processo permite criar uma hierarquia de grupos baseada na densidade dos dados, onde diferentes níveis da hierarquia correspondem a diferentes valores de raio $\varepsilon$. Através do uso das distâncias centrais e do grafo de alcançabilidade mútua, o algoritmo identifica os núcleos de cada grupo e os *outliers* de maneira eficiente, permitindo uma análise hierárquica dos dados. Neste projeto, o HDBSCan é utilizado pelo BERTopic para o agrupamento dos *tweets* relacionados a cada uma das bases utilizadas no trabalho, buscando gerar grupos com elementos similares entre si.

Figura 4 – Cálculo da distância central - HDBSCan




Fonte: McInnes John Healy (2016)

Figura 5 – Minimun Spanning Tree - HDBSCan

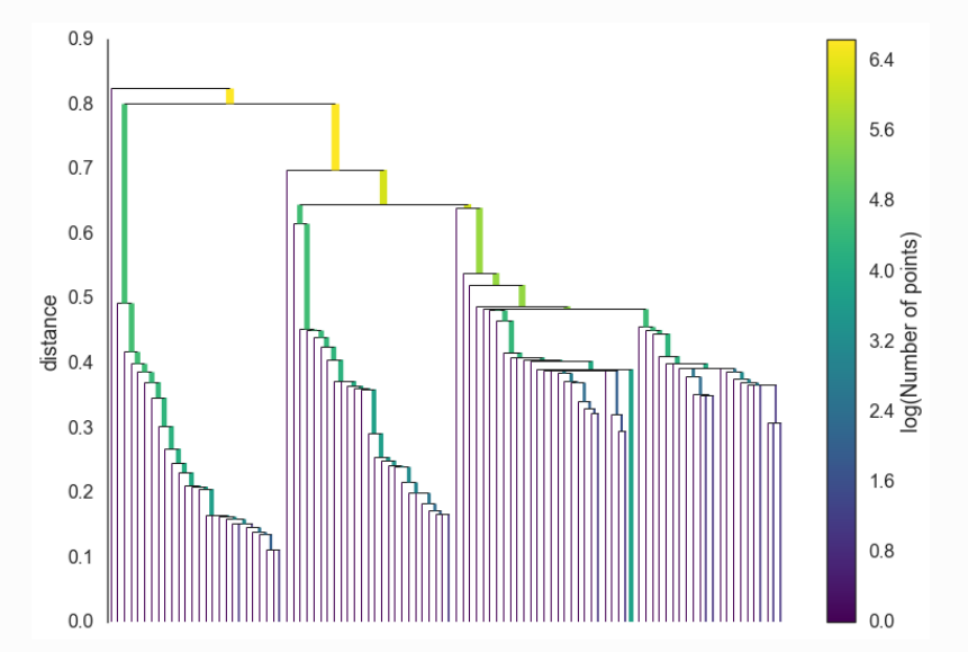



Fonte: McInnes John Healy (2016)

### 2.2.3 c-TF-IDF

O BERTopic utiliza, em sua etapa final, uma variação do TF-IDF intitulada por Grootendorst (2020) de c-TF-IDF. Esta é uma variação da técnica tradicional TF-IDF adaptada para a modelagem de tópicos em agrupamentos de documentos. Nesta abordagem, após o agrupamento dos documentos de texto em *clusters* utilizando o HDBSCan,

Figura 6 – Dendograma - HDBSCan




Fonte: McInnes John Healy (2016)

cada *cluster* gerado representa um tópico distinto. Para gerar a representação de cada tópico, busca-se identificar quais palavras são mais relevantes dentro de cada *cluster*.

Na técnica TF-IDF original (JONES, 1972), a importância de uma palavra é calculada em um documento individual, considerando a frequência do termo no documento (TF) e a frequência inversa do termo no conjunto de documentos (IDF), determinando, desta forma, a importância relativa da palavra no documento. Entretanto, ao trabalhar com *clusters* de textos, é necessário adaptar essa técnica para capturar a importância das palavras em um conjunto de textos que compõem um tópico.

Dessa forma, no c-*TF-IDF*, todos os documentos pertencentes a um mesmo *cluster* são concatenados, formando um único documento que representa o *cluster*. A partir disso, calcula-se a frequência do termo no *cluster* ($tf_{t,c}$) e a frequência do termo em todos os *clusters* ($tf_t$). A importância da palavra no *cluster* é então determinada pela equação 13, Onde $tf_{t,c}$ é a frequência do termo $t$ no *cluster* $c$, $A$ é o número médio de palavras por *cluster* e $tf_t$ é a frequência do termo $t$ em todos os *clusters*.

$$W_{t,c} = tf_{t,c} \cdot \log\left(1 + \frac{A}{tf_t}\right) \tag{13}$$

Essa abordagem permite captar a importância das palavras em *clusters* ao invés de documentos individuais, permitindo uma geração de representações de tópicos mais coerentes e significativas. Assim, o c-*TF-IDF* captura as características únicas de cada *cluster*, gerando os termos que melhor representam o tópico avaliado.

Por se tratar de uma técnica que busca identificar os termos mais relevantes em um conjunto de documentos agrupados, o c-*TF-IDF* contribui para a melhoria na qualidade

das representações dos tópicos gerados pelo BERTopic.

Neste trabalho, o HDBSCan foi utilizado como método de agrupamento por ser a biblioteca padrão utilizada dentro do BERTopic, além de ser um método de fácil configuração por possuir uma quantidade menor de parâmetros ajustáveis e estes parâmetros possuírem conceitos mais simplificados. Por ser um algoritmo baseado em densidade, ele possui uma boa capacidade de lidar com diferentes densidades de dados e detectar ruídos de forma natural, que são características importantes se tratando de *embeddings de tweets*, que possuem diferentes distribuições, formas de escrita e tamanho.

## 2.3 LLMs

Os LLMs se referem aos modelos de linguagem de Transformers que contêm centenas de bilhões de parâmetros, e são treinados com quantidades massivas de dados de texto (SHANAHAN, 2023). Estes modelos de linguagem mostraram uma grande capacidade de entender linguagem natural e resolver tarefas complexas através da geração de textos.

Nesta seção serão abordados os principais tópicos de LLM pertinentes a este trabalho, como as habilidades emergentes de um LLM, os conceitos de Instruction Tuning e como funcionam os modelos quantizados para redução de tamanho e complexidade computacional.

### 2.3.1 Habilidades Emergentes

Um dos temas mais emergentes em inteligência artificial nos últimos anos é o de modelagem de linguagem, mais especificamente os modelos generativos. Os modelos de linguagem têm recebido uma grande atenção na literatura, e de acordo com Zhao et al. (2023) podem ser agrupados em quatro principais estágios de evolução: os Statistical Language Models (SLM), os Neural Language Models (NLM), os Pre-trained Language Models (PLM) e os Large Language Models (LLM).

Os modelos de linguagem estatísticos surgiram na década de 1990 utilizando métodos de aprendizado estatístico. Nestes modelos, são utilizadas cadeias de Markov para prever a próxima palavra dado um contexto recente (GAO; LIN, 2004). Em geral, são utilizados com janelas de contexto de tamanho fixo $n$, e por isso ficaram conhecidos como modelos de linguagem *n-gram* (BROWN et al., 1992). Entretanto, devido à simplicidade deste tipo de modelagem, modelos deste tipo não eram capazes de tratar estruturas gramaticais profundas e não possuíam bom desempenho em agrupar palavras de um vocabulário para formar frases completas (ROSENFELD, 2000).

O segundo marco com relação aos modelos de linguagem foi o surgimento dos modelos de linguagem neurais (NLM). Estes modelos buscavam calcular a probabilidade de sequências de palavras utilizando redes neurais, como as redes neurais recorrentes (RNN). Houve

importantes trabalhos que contribuíram para este campo, como o trabalho de (BENGIO et al., 2003) que introduziu a ideia de representação distribuída de palavras e criou uma função de previsão de palavras baseada em contexto. Posteriormente foi apresentado por Mikolov et al. (2013b) o word2vec, uma rede neural rasa que também seguia o conceito de representação distribuída de palavras, mas que possuía melhorias em eficiência de treinamento, escalabilidade e qualidade das representações, utilizando duas principais arquiteturas novas, o Continuous Bag of Words (CBOW) e o *Skip-gram* (MIKOLOV et al., 2013a).

Durante a segunda metade da década de 2010, surgiram estudos que propuseram o conceito de modelos de linguagem pré-treinados (PLM). Um dos modelos pioneiros foi o modelo desenvolvido por Peters et al. (2018), o ELMo, que trouxe a ideia de capturar representações contextuais de palavras por meio de redes neurais bidirecionais, ao invés de utilizar o aprendizado de representações fixas das palavras. Logo depois surgiu o modelo Bidirectional Encoder representations from Transformers (BERT), baseado na arquitetura Transformer altamente paralelizável e utilizando modelos de linguagem bidirecionais com tarefas de pré-treinamento especialmente projetadas em grandes *corpora* não rotulados em larga escala. Essas representações de palavras contextuais pré-treinadas são muito eficazes como recursos linguísticos de uso geral, o que elevou significativamente o desempenho das tarefas de PLN. Este estudo inspirou uma grande quantidade de trabalhos subsequentes, estabelecendo os paradigmas de pré-treinamento e *fine-tuning* (DEVLIN et al., 2019). Esse estudo foi precursor para o surgimento dos LLMs, inspirando os primeiros modelos Generative Pre-trained Transformer (GPT), como o GPT-2 (RADFORD et al., 2019).

Mais recentemente, no início da década de 2020, pesquisadores observaram que expandir tanto o tamanho do modelo quanto o tamanho dos dados de entrada em PLMs frequentemente resultava em melhorias na capacidade do modelo para executar tarefas *downstream* (KAPLAN et al., 2020). Essas tarefas são aquelas para as quais o modelo de linguagem pré-treinado é aplicado após a sua fase de aprendizado geral. Vários estudos exploraram o aumento de parâmetros em PLMs para bilhões, exemplificado pelo GPT-3, que possui 175 bilhões de parâmetros, e o PaLM, com seus 540 bilhões de parâmetros, em comparação com os 330 milhões de parâmetros do BERT (BROWN et al., 2020; CHOWDHERY et al., 2022).

Wei et al. (2022c) descrevem as habilidades emergentes como "as habilidades que não estão presentes em modelos pequenos, mas surgem em modelos de larga escala", o que é uma das características mais notáveis que distinguem os LLMs dos modelos de linguagem menores. Dentre as principais características notadas nas habilidades emergentes, está o aumento significativo do desempenho do modelo em uma série de tarefas que necessitam de conhecimento de contexto e raciocínio. Há três tipos principais de habilidades emergentes para LLMs:

- ❏ Aprendizado no contexto (*In-context Learning*): Assumindo que o modelo de lingua-

gem tenha sido alimentado com instruções de linguagem natural e demonstrações de tarefas, ele é capaz de gerar uma saída para os dados fornecidos como entrada completando a sequência de palavras do texto fornecido, sem a necessidade de realizar treinamentos adicionais. Essa habilidade foi formalmente introduzida por Brown et al. (2020) no modelo GPT-3.

- Seguimento de instruções (*Instruction following*): Através da realização de *fine-tuning* com conjuntos de dados multitarefa, os LLMs podem seguir instruções para realizar tarefas nunca vistas anteriormente sem o uso de exemplos específicos. De acordo com experimentos realizados por Wei et al. (2022a), o LaMDA-PT ajustado com instruções passou a superar significativamente modelos não ajustados em tarefas inéditas quando o tamanho do modelo atingiu 68B, mas o comportamento não se repetiu para modelos menores que 8B de parâmetros.

- Raciocínio passo a passo (*Step-by-step reasoning*): Em modelos de linguagem pequenos, geralmente são encontradas dificuldades para resolver tarefas complexas que envolvam múltiplas etapas de raciocínios, como problemas matemáticos. Entretanto, com a estratégia de prompt chain-of-thought (CoT), os LLMs são capazes de realizar etapas intermediárias de raciocínio para chegar à resposta final (WEI et al., 2023b). De acordo com alguns autores, essa habilidade pode ter surgido por meio de treinamento do modelo com códigos computacionais (FU YAO; PENG; KHOT, 2022) (WEI et al., 2023b).

Neste trabalho, foram aproveitadas algumas habilidades emergentes dos LLMs nas tarefas de sumarização. O fato de não ter sido realizado nenhum tipo de *instruction tuning* para a sumarização mostra a capacidade de seguimento de instruções dos modelos.

### 2.3.2 Modelos quantizados

Os LLMs costumam ter um número muito alto de parâmetros. Isso acaba tornando a inferência desses modelos muito custosa computacionalmente, consumindo uma quantidade consideravelmente alta de memória e recursos de processamento. Pensando em resolver esses problemas, foi criado o método de quantização, que busca permitir que modelos de alta complexidade possam ser utilizados em máquinas com recursos limitados, além de reduzir a latência de inferência dos resultados (ZHAO et al., 2023).

A técnica de quantização tem por conceito converter números de pontos flutuantes para números inteiros. Como os LLMs possuem grandes quantidades de parâmetros representados como números de pontos flutuantes, essa técnica se torna bastante eficaz para reduzir a complexidade e o tamanho desses modelos. Para realizar a quantização, busca-se mapear os números de pontos flutuantes, que geralmente são utilizados no armazenamento dos parâmetros (pesos) dos modelos e suas ativações das camadas ocultas, para números

inteiros. Geralmente são utilizadas quantizações de 8 bits (INT8) (GHOLAMI et al., 2021).

A quantização envolve o uso de uma função que mapeia um número de ponto flutuante $x$ em um valor quantizado $x_q$. Essa função, em geral, inclui três principais fatores:

- ❏ Fator de escala: São estabelecidos parâmetros $\alpha$ e $\beta$ que irão determinar uma faixa de valores da rede que serão mantidos após a quantização.
- ❏ Fator de ponto zero: Esse fator determina se a quantização é simétrica, ou seja, se ela será realizada de forma equilibrada em relação a zero. Essa simetria é útil quando se deseja representar tanto positivos quanto negativos de forma precisa.
- ❏ Arredondamento: A função de quantização envolve o arredondamento do valor quantizado para o valor inteiro mais próximo.

Na equação 14, $x_q$ é o valor quantizado, $x$ é o valor original em ponto flutuante, $S$ é o fator de escala, $Z$ é o fator de ponto zero e $R(\mathring{u})$ é a operação de arredondamento.

$$x_q = R\left(\frac{x}{S}\right) - Z \tag{14}$$

Como processo reverso, há a desquantização, que recupera o valor original de um valor quantizado seguindo a equação 15. Essa etapa é necessária durante a inferência para garantir que os resultados obtidos continuem com uma precisão adequada.

$$x \approx S * (xq + Z) \tag{15}$$

O erro de uma quantização é dado pela diferença entre o valor original em ponto flutuante e o valor recuperado após a desquantização.

Existem duas abordagens principais de quantização para LLMs. A primeira, é a *Quantization-Aware Training - QAT (QAT)*. Nessa abordagem, o modelo é treinado novamente já se tendo estabelecido que ele será quantizado posteriormente. Essa abordagem busca que o modelo seja treinado para funcionar bem com números inteiros em vez de números de ponto flutuante. Entretanto, o fato desta abordagem requerer um treinamento adicional a torna computacionalmente mais custosa. A segunda abordagem é a *Post-Training Quantization (PTQ)*. Essa abordagem envolve a aplicação de quantização após o treinamento original do modelo, sem a necessidade de um retreinamento. Essa abordagem é mais eficiente computacionalmente, já que não requer treinamento adicional, porém pode não gerar resultados tão satisfatórios devido à arquitetura dos modelos ser originalmente pensada para trabalhar com pontos flutuantes.

Existem diversos métodos de aplicação do PTQ, cada um com suas particularidades. Um dos métodos mais comuns é a decomposição em precisão mista, que se mostrou eficaz em LLMs com um grande número de parâmetros. Nesses modelos, foi observado que valores extremamente altos podem aparecer nas camadas de ativações ocultas, um fenômeno

que é chamado de emergência de *outliers*. Esses *outliers* se encontram principalmente em dimensões de características específicas em camadas do modelo Transformer. Para lidar com isso, Dettmers et al. (2022) propôs uma abordagem chamada LLM.int8(), que busca quantizar separadamente as dimensões de características com *outliers* e as dimensões restantes, utilizando números de ponto flutuante de 16 bits e inteiros de 8 bits para recuperar esses *outliers* com uma precisão mais alta.

Por fim, a quantização camada a camada é uma abordagem que busca encontrar pesos quantizados ótimos que minimizem uma perda de reconstrução camada a camada na rede. O GPTQ, proposto por Frantar et al. (2023), aplica esta abordagem e é capaz de quantizar modelos muito grandes em precisão de 3 ou 4 bits.

Quanto aos métodos QAT, o principal utilizado é a quantização aprimorada por *fine-tuning* eficiente. Quando se busca realizar uma quantização diretamente no pós-treinamento utilizando valores de baixa precisão, como o INT4, é comum acontecer uma grande degradação no desempenho do modelo. Para resolver esse problema, o método QLoRA (DETTMERS et al., 2023) incorpora pequenos adaptadores ajustáveis com precisão de 16 bits nos modelos quantizados. Isso permite um ajuste fino eficiente e de alta precisão do modelo. O QLoRA busca combinar as vantagens da técnica LoRA com métodos de quantização. Experimentos realizados por Dettmers et al. (2023) mostraram que modelos quantizados com uma precisão de 4 bits conseguiram obter resultados similares a modelos de 16-bit utilizando este método.

Para o presente trabalho, entre os modelos exemplificados na Seção 4.1.4 foram utilizados o Llama 2 13B (TOUVRON et al., 2023) e o Bode (GARCIA et al., 2024) com a quantização PTQ de 8-bit. Isso proporciona um menor custo computacional no processamento do modelo e gera uma perda mínima de qualidade comparado ao modelo original.

## 2.4 Divisão e Conquista

A divisão e conquista, como definida no livro *Introduction to Algorithms* de Cormen et al. (2022), é uma técnica poderosa para projetar e desenvolver algoritmos eficientes. Esta abordagem busca realizar a decomposição de um problema original complexo em subproblemas menores, que são instâncias do mesmo problema original. O processo segue dois casos principais: o caso base e o caso recursivo.

No caso base do algoritmo, o problema é suficientemente pequeno para ser resolvido diretamente, sem necessidade de mais uma recursão para ser concluído. Já no caso recursivo, o problema é dividido em um ou mais subproblemas menores. Cada subproblema é então resolvido recursivamente, e as soluções obtidas são combinadas para formar a solução final do problema original.

Nos algoritmos que buscam utilizar a técnica de divisão e conquista, frequentemente

são utilizadas técnicas de recorrências. Uma recorrência é uma equação que expressa uma função em termos de seus valores em argumentos menores, permitindo expressar matematicamente o tempo de execução de algoritmos recursivos. Cormen et al. (2022) exemplificaram o cálculo através de um algoritmo de multiplicação de matrizes que utiliza divisão e conquista, mostrando que é possível estabelecer uma recorrência que descreve como o tempo total depende do tamanho das matrizes e do número de subproblemas gerados.

No algoritmo utilizado como exemplo, duas matrizes de tamanho $n \times n$ originais são divididas em submatrizes de tamanho $\frac{n}{2} \times \frac{n}{2}$. O tempo de execução $T(n)$ desse algoritmo é descrito pela recorrência da equação 16. Esta equação descreve a multiplicação de matrizes via divisão e conquista ao partir duas matrizes $n \times n$ em quatro submatrizes de dimensão $\frac{n}{2} \times \frac{n}{2}$, gerando oito subproblemas de multiplicação de menor tamanho (cada um $\frac{n}{2} \times \frac{n}{2}$). Depois, ao combinar as respostas dos subproblemas, é necessária a realização de operações de soma com custo proporcional a $\Theta(n^2)$. Assim, o termo $8T\left(\frac{n}{2}\right)$ se refere às oito multiplicações recursivas e $\Theta(n^2)$ representa o custo adicional de reunir os resultados em cada nível da recursão.

$$T(n) = 8T\left(\frac{n}{2}\right) + \Theta(n^2) \tag{16}$$

Para resolver essas recorrências e compreender o comportamento assintótico dos algoritmos, são utilizados métodos como a substituição, árvores de recursão e o método mestre. Essas ferramentas permitem determinar limites superiores e inferiores para o tempo de execução, gerando informações sobre a eficiência do algoritmo. Assim, a divisão e conquista permite que problemas complexos sejam solucionados de maneira eficiente ao serem quebrados em partes menores. Essa estratégia não só simplifica a resolução de problemas complexos, como também facilita a análise do desempenho dos algoritmos utilizados.

Nos LLMs atuais, frequentemente torna-se necessário contornar uma limitação comum e conhecida destes modelos: o número máximo de *tokens* de entrada que o modelo aceita para realizar o processamento. Essa limitação acaba se tornando um problema quando a quantidade de textos que é necessário realizar o processamento ultrapassa esse número de *tokens*, impedindo o uso do modelo para a finalidade desejada. Com o intuito de contornar essa limitação, diversos autores lançaram mão da técnica de divisão e conquista para conseguir realizar processamentos em grandes quantidades de texto utilizando LLMs. Alguns *frameworks* destinados ao uso destes modelos, como o Langchain[3], implementaram variações da técnica de *Map Reduce* para quebrar um documento em várias partes, processar cada parte individualmente e gerar um resultado único no final. Nesta seção será explicada a origem do algoritmo *Map Reduce* e as semelhanças e diferenças com a estratégia de divisão e conquista abordada na solução proposta neste trabalho.

[3] https://www.langchain.com/

### 2.4.1 MapReduce

O MapReduce é um modelo de programação proposto por Dean e Ghemawat (2008) que busca facilitar o processamento e a geração de grandes conjuntos de dados em ambientes de computação distribuída. O modelo proposto se baseia em duas funções principais: a função *Map* e a função *Reduce*. A função *Map* aplica uma operação em cada par de entrada chave/valor, produzindo um conjunto intermediário de pares chave/valor. Na sequência, a função *Reduce* processa cada chave intermediária gerada na operação anterior, combinando todos os valores associados a ela para gerar o resultado final.

Matematicamente, a função *Map* pode ser definida seguindo a equação 17, onde $(k_1, v_1)$ representa os pares de entrada e $\{(k_2, v_2)\}$ é o conjunto de pares intermediários gerados. No fluxo do MapReduce, cada função Map recebe um par ($k1$, $v1$) como entrada e produz vários pares ($k2$, $v2$). Já a função Reduce agrupa todos os valores $v_2$ que têm a mesma chave $k_2$ em uma lista, gerando a tupla ($k_2$, $[v_2]$) e produz o par ($k2$, $v3$).

$$\text{Map} : (k_1, v_1) \rightarrow \{(k_2, v_2)\} \tag{17}$$

Já a função *Reduce* é definida matematicamente na equação 18, onde $[v_2]$ é a lista de valores associados à chave intermediária $k_2$, e $(k_2, v_3)$ é o par resultante após a agregação dos valores.

$$\text{Reduce} : (k_2, [v_2]) \rightarrow (k_2, v_3) \tag{18}$$

Em contextos de grande volumetria de dados, o Map Reduce pode ser aplicado em diversas tarefas, como contagem de frequências de palavras, construção de índices invertidos e processamento de logs. Por exemplo, ao realizar a contagem de ocorrências de cada palavra em um grande *corpus* textual, a função *Map* associaria um número "1" a cada ocorrência da palavra encontrada. Já a função *Reduce* então somaria todos os "1"s associados a cada palavra, resultando na contagem total de ocorrências dessa palavra.

Em cenários onde há grandes conjuntos de dados e diversos nós de processamento, o Map Reduce incorpora mecanismos para tolerância a falhas e otimização de desempenho. Quando uma falha é detectada em um nó de processamento, as tarefas afetadas são automaticamente reatribuídas a outros nós disponíveis. Para otimizar o uso de recursos, o sistema tenta executar as tarefas *Map* em máquinas que possuem os dados necessários localmente.

O *Map Reduce* é, portanto, um algoritmo de divisão e conquista. No âmbito de processamento de textos com LLM, como citado no início da seção, alguns *frameworks* buscaram replicar esse comportamento de mapear e reduzir para o âmbito de PLN, buscando quebrar um conjunto de multidocumentos, por exemplo, processá-los cada um individualmente, e depois gerar um resultado único processado.

Para este trabalho, foi desenvolvida uma técnica que se assemelha ao *Map Reduce* original em seu conceito de divisão e conquista, mas com algumas diferenças chave. Por se tratar da sumarização de *tweets*, que individualmente possuem um tamanho reduzido, não foi necessária a realização da etapa de divisão do problema, que foi substituída por um agrupamento dos *tweets* em blocos para a posterior sumarização e o seguimento da recursão até chegar no caso base, que se equivaleria a etapa *Reduce*. Porém, a etapa de *Reduce* do algoritmo foi utilizada como base da sumarização multinível proposta na Seção 4.2.3, onde os resultados dos sumários são reunidos e reprocessados até atingir um único valor final. O processamento também foi realizado em uma única GPU, o que não permitiu a paralelização da execução e cada bloco foi processado serializadamente. O processo é detalhado na Seção 4.2.3.

# Capítulo 3

# Trabalhos Relacionados

Diversos trabalhos recentes buscam abordar o tema de sumarização abstrativa, seja ela multi-documentos ou de textos únicos. Este Capítulo tem por objetivo entender os avanços no tema, quais as principais técnicas e desenvolvimentos apresentados na literatura e os resultados obtidos. Nas subseções deste capítulo foram explorados trabalhos que utilizaram sumarização multi-documentos com LLMs, modelagem de tópicos, sumarização utilizando métodos baseados em Map Reduce e avaliações de sumários utilizando BERTScore.

## 3.1 Sumarização multi-documentos com LLM

Zhang et al. (2023) exploraram o uso de LLMs em tarefas de sumarização automática abstrativa. Os autores realizaram uma avaliação sistemática de dez modelos diferentes em tarefas de sumarização de notícias de dois datasets distintos, o CNN/Daily Mail e o XsUM, avaliando configurações sem nenhum exemplo de sumário fornecido no prompt (*zero-shot*), configurações onde foram passados cinco exemplos no prompt (*five-shot*), sumários gerados em modelos com *fine-tuning* específico para sumarização comparados aos sumários de referência originais dos *datasets* e sumários gerados por humanos.

Foi identificado que o *tuning* de instruções desempenhou um papel fundamental na capacidade de sumarização *zero-shot*, ou seja, sem exemplos passados para treinamento. Os sumários de referência originais dos *datasets* utilizados, como o CNN/Daily Mail e XSum, apesar de terem sido gerados por humanos, não foram elaborados para terem uma alta qualidade, pois são basicamente *bullet points* aproveitados dos sites de notícias. Dessa forma, buscando resolver este problema e ter uma base confiável de comparação, os autores contrataram pessoas para escrever novos resumos humanos de melhor qualidade.

Feito isso, eles puderam demonstrar que as avaliações melhoram quando as referências são mais consistentes. Vale ressaltar que os resultados dos sumários gerados por humanos não foram reportados separadamente para cada *dataset* no trabalho.

Como mencionado anteriormente, para mensurar a qualidade dos modelos em tarefas de sumarização, foram utilizados os *datasets* CNN/DailyMail e XSum. Em cada conjunto, foram selecionados 100 artigos para avaliação manual, de forma que cada modelo gerou um resumo para cada artigo. Além disso, foram comparados tanto modelos de linguagem em diferentes configurações (*zero-shot, few-shot e fine-tuned*) quanto os resumos de referência originais presentes nos *datasets*, além dos sumários gerados por humanos.

A avaliação foi realizada com anotações humanas, considerando três dimensões principais, sendo elas:

- ❏ Fidedignidade: mede se as informações apresentadas no sumário estão corretas e condizem com o artigo.
- ❏ Coerência: avalia se o texto do sumário é fluido, organizado e não apresenta contradições internas.
- ❏ Relevância: verifica se o sumário captura os pontos mais importantes do artigo.

Cada sumário foi avaliado por três anotadores na escala Likert de 1 a 5 para os critérios de coerência e relevância e binária no caso da fidedignidade, cujas médias são apresentadas na Tabela 1. Foi observado que modelos treinados com *instruction tuning* tendem a produzir sumários mais consistentes e fiéis ao texto. Contudo, os sumários de referência padrão dos conjuntos de dados apresentaram qualidade abaixo do esperado em aspectos como coerência, destacando a necessidade de novas referências humanas de melhor qualidade.

Foi constatado também que o melhor modelo, o Davinci GPT-3, possui mais cópias diretas do texto original quando comparado a sumários gerados por humanos, entretanto a qualidade de ambos os sumários são equiparáveis e possuem um resultado similar.

Contudo, os autores argumentam que à medida que os LLMs se aproximam do desempenho humano, a avaliação humana passa a precisar de um número maior de amostras e medidas menos ruidosas. Também foi citado que a qualidade das referências de avaliação é um problema crucial, e que a avaliação humana é afetada pela subjetividade dos avaliadores. Assim, sugere-se que a sumarização pode ser melhor avaliada em aplicações práticas onde os valores do usuário estão melhor definidos.

Liu e Healey (2023) utilizaram o GPT-3 para realizar a sumarização abstrativa de uma coleção de documentos grandes. Os autores discutem as limitações existentes nos modelos atuais de LLMs com relação à quantidade de dados de entrada, o que impede de utilizar os documentos por inteiro como entrada. Com o intuito de superar essa limitação, os autores propuseram uma abordagem que envolve a compressão inteligente dos documentos antes

Tabela 1 – Resultados dos modelos de linguagem e modelos *fine-tuned*, além dos sumários de referência e dos sumários humanos.

| **Config. / Modelo** | **CNN/Daily Mail** | | | **XSum** | | |
|---|---|---|---|---|---|---|
| | **Fidedignidade** | **Coerência** | **Relevância** | **Fidedignidade** | **Coerência** | **Relevância** |
| **Zero-shot language models** | | | | | | |
| GPT-3 (350M) | 0.29 | 1.92 | 1.84 | 0.26 | 2.03 | 1.90 |
| GPT-3 (6.7B) | 0.29 | 1.77 | 1.93 | 0.77 | 3.16 | 3.39 |
| GPT-3 (175B) | 0.76 | 2.65 | 3.50 | 0.80 | 2.78 | 3.52 |
| Ada Instruct v1 (350M*) | 0.88 | 4.02 | 4.26 | 0.81 | 3.90 | 3.87 |
| Curie Instruct v1 (6.7B*) | 0.97 | 4.24 | 4.59 | 0.96 | 4.27 | 4.34 |
| Davinci Instruct v2 (175B*) | **0.99** | 4.15 | 4.60 | **0.97** | 4.41 | 4.28 |
| **Five-shot language models** | | | | | | |
| Anthropic-LM (52B) | 0.94 | 3.88 | 4.33 | 0.70 | 4.77 | 4.14 |
| Cohere XL (52.4B) | 0.99 | 3.42 | 4.48 | 0.63 | 4.79 | 4.00 |
| GLM (130B) | 0.94 | 3.69 | 4.24 | 0.74 | 4.72 | 4.12 |
| OPT (175B) | 0.96 | 3.64 | 4.33 | 0.67 | 4.80 | 4.01 |
| GPT-3 (350M) | 0.86 | 3.73 | 3.85 | – | – | – |
| GPT-3 (6.7B) | 0.97 | 3.87 | 4.17 | 0.75 | 4.19 | 3.36 |
| GPT-3 (175B) | 0.99 | 3.95 | 4.34 | 0.69 | 4.69 | 4.03 |
| Ada Instruct v1 (350M*) | 0.84 | 3.84 | 4.07 | 0.63 | 3.54 | 3.07 |
| Curie Instruct v1 (6.7B*) | 0.96 | 4.30 | 4.43 | 0.85 | 4.28 | 3.80 |
| Davinci Instruct v2 (175B*) | 0.98 | 4.13 | 4.49 | 0.77 | 4.83 | 4.33 |
| **Fine-tuned language models** | | | | | | |
| BRIO | 0.94 | 3.94 | 4.40 | 0.58 | 4.68 | 3.89 |
| PEGASUS | 0.97 | 3.93 | 4.38 | 0.57 | 4.73 | 3.85 |
| *Sumários de referência existentes* | 0.84 | 3.20 | 3.94 | 0.37 | 4.13 | 3.00 |
| *Sumários humanos* | 0.93 | **4.39** | 4.26 | 0.93 | **4.39** | 4.26 |

**Observações:** (*) indica versões *instruction-tuned*. Valores de *Sumários humanos* são médias consolidadas (não reportados separadamente para cada conjunto).

Fonte: Zhang et al. (2023)

da sumarização. Dentre as etapas, estão a extração de informações semanticamente significativas, a subdivisão dos documentos em *clusters* temáticos, a sumarização abstrativa de cada *cluster* e a agregação desses resumos para criar um resumo abstrato final.

Para a clusterização, os autores utilizaram o UMAP para a redução de dimensionalidade das representações dos textos e o HDBSCan para a clusterização com base na semelhança. Feito isto, foi realizada a extração de sentenças de tópicos. Nesta etapa, cada *cluster* obtido na etapa anterior é analisado para identificar palavras-chave relacionadas a conceitos. Isso é feito usando o algoritmo LDA (Latent Dirichlet Allocation), que transforma termos em matrizes conceito-documento. Essas palavras-chave são usadas para extrair frases dos documentos que contêm informações relevantes aos tópicos. Feito isto, é realizada a divisão semântica de frases. Como as frases extraídas podem ser muito longas, elas são divididas em partes menores chamadas de pedaços semânticos. Isso é feito usando *embeddings* de sentenças e uma matriz de similaridade para identificar os pontos de divisão entre as frases. Essa etapa ajuda a organizar as informações de maneira mais granular. Na etapa seguinte, cada pedaço semântico é resumido utilizando o modelo GPT. Os resumos de cada pedaço são posteriormente combinados para criar um resumo abstrato do cluster de documentos. Finalmente, é realizada uma análise de sentimento nos resumos para avaliar o tom emocional dos resumos gerados.

Tabela 2 – Resultados do sistema de sumarização de Liu e Healey (2023)

| System | CNN/Daily Mail | | | Gigaword | | |
|---|---|---|---|---|---|---|
| | ROGUE-1 | ROGUE-2 | ROGUE-L | ROGUE-1 | ROGUE-2 | ROGUE-L |
| BART | 44.2 | 21.3 | 40.9 | 39.1 | 20.1 | 36.4 |
| BRIO | 48.0 | 23.8 | 44.7 | — | — | — |
| PEGASUS | 44.2 | 21.5 | 41.1 | 39.1 | 19.9 | 36.2 |
| MoCa | 48.9 | 24.9 | 45.8 | 39.6 | 20.6 | 36.8 |
| Ours | 58.7 | 25.6 | 56.0 | 38.7 | 19.7 | 35.8 |

Fonte: Liu e Healey (2023)

Como resultados, o Rouge-Score obtido por esse sistema obteve um desempenho consideravelmente maior que seus concorrentes no dataset *CNN/Daily Mail*, enquanto no dataset *Gigaword* todos os modelos obtiveram um desempenho parecido, como pode ser visto na Tabela 2. Entretanto, este sistema apresenta a vantagem de ser escalável e obter vários níveis de detalhes dos documentos, permitindo ao usuário explorar o sentimento geral de toda a coleção de documentos, o sentimento de tópicos individuais dentro da coleção, ou até mesmo o sentimento de pedaços semânticos específicos ou frases, além de aproveitar a capacidade dos modelos mais atuais de LLM, no caso, o GPT-3.

Goyal, Li e Durrett (2023) estudaram o impacto dos LLMs, como o GPT-3, na sumarização de texto com foco na área de sumarização de notícias. Inicialmente eles compararam o desempenho do GPT-3 com modelos que tiveram o *fine-tuning* realizado em grandes corpus de sumarização. Para realizar os testes, os autores conduziram testes A/B para avaliar a preferência humana entre os resumos gerados pelo GPT-3 e pelos modelos com *fine-tuning*. Como resultado da avaliação humana, na maioria dos casos os humanos preferiram os resumos do GPT-3, mesmo quando o modelo é executado com um *prompt* que contém apenas a descrição da tarefa, sem nenhum treinamento adicional.

Como métricas quantitativas, os autores utilizaram as métricas ROUGE (LIN, 2004) e BERTScore (ZHANG et al., 2020b). Entretanto, os autores argumentam que essas métricas não são adequadas para avaliar os resumos gerados pelo GPT-3, por conta de beneficiarem resumos onde as palavras se sobrepõem ao texto original, e não conseguem captar a qualidade dos resumos gerados. Conforme visto na Tabela 3, na avaliação humana o GPT-3 obteve uma votação de melhor modelo por 58% dos avaliadores no dataset da CNN e em 57% dos avaliadores no dataset BBC. Já em relação ao Rouge Score, o GPT-3 obteve o pior resultado em todos os cenários avaliados quando comparado ao Pegasus, ao Brio, que é um modelo de sumarização abstrativa proposto por Liu et al. (2022) e ao T0, modelo proposto por Sanh et al. (2022), mostrando que, de fato, a métrica possui suas deficiências na avaliação quantitativa.

Deroy, Ghosh e Ghosh (2023) realizaram um estudo sobre a aplicação de sumarização abstrativa automática de julgamentos legais. Os autores citam que tradicionalmente a su-

Tabela 3 – Resultados do sistema de sumarização

| **Dataset** | **BRIO** | | **T0** | | **GPT3** | |
|---|---|---|---|---|---|---|
| | Best ↑ | Worst ↓ | Best ↑ | Worst ↓ | Best ↑ | Worst ↓ |
| **CNN** | 36 | 24 | 8 | 67 | 58 | 9 |
| **BBC** | 20 | 56 | 30 | 29 | 57 | 15 |

Fonte: Goyal, Li e Durrett (2023)

marização de documentos legais era feita majoritariamente de forma extrativa, mas, com a popularização da sumarização abstrativa, estes modelos estão sendo cada vez mais usados por gerarem resumos mais naturais e coerentes. Nesse contexto, os autores buscaram responder se modelos de sumarização abstrativa pré-treinados específicos para o domínio jurídico e LLMs gerais, como o ChatGPT, estão prontos para serem usados diretamente na geração automática de resumos de julgamentos. Para responder a essa pergunta, eles aplicaram esses modelos a julgamentos de tribunais indianos e avaliaram a qualidade dos resumos gerados. Para a avaliação quantitativa, foram utilizadas as métricas ROUGE (LIN, 2004), METEOR (BANERJEE; LAVIE, 2005) e BLEU (PAPINENI et al., 2002). Nos resultados, os modelos de domínio geral como o ChatGPT e o Davinci obtiveram os melhores resultados, entretanto sendo superados por modelos pré-treinados em sumarização jurídica, como o LegPegasus e em alguns casos por modelos extrativos, como o CaseSummarizer.

Porém, os autores citam alguns erros graves nos sumários gerados pelos modelos abstrativos. Entre os erros citados estão a fusão de duas sentenças, onde a primeira sentença fica incompleta. Também houve erros de geração de números incorretos, como anos e valores, além de haver alucinações em certos casos como a inserção de nomes de tribunais e estatutos dos EUA que não tinham relação com os documentos de entrada.

Soni e Wade (2023) realizaram um estudo para avaliar o desempenho do ChatGPT em tarefas de sumarização abstrativa. O estudo teve por objetivo avaliar o desempenho do modelo em tarefas de sumarização usando métricas automatizadas, utilizar revisores humanos para realizarem revisões cegas e construir classificadores de texto automatizados para distinguir entre os resumos gerados pelo ChatGPT e resumos reais. Para o estudo, foi preparado um conjunto de dados com 50 resumos gerados do dataset da CNN pelo modelo e uma análise desses resumos utilizando métricas automatizadas. Os autores não realizaram a comparação dos resultados com nenhum outro modelo conhecido, porém os ROUGE-L score foi de 0,20 e o ROUGE-1 foi de 0,30, valores que não são tão altos comparados ao teto de 1. Na avaliação humana, os revisores não conseguiram distinguir de forma consistente entre os resumos gerados pelo modelo e os resumos humanos, obtendo uma precisão de 49% e mostrando que o resumo gerado pelo ChatGPT possui uma qualidade comparável aos resumos gerados por humanos. A matriz de confusão da Figura 8 mostra a avaliação dos revisores dos resumos gerados e dos resumos originais feitos

por humanos. Os resultados mostram uma dificuldade de distinção entre os resumos, indicando a qualidade parecida de ambos.

Figura 7 – Matriz de confusão na avaliação humana do trabalho de Soni e Wade (2023)

| | | ***Truth*** | |
|---|---|---|---|
| | | Generated | Original |
| ***Reviewers*** | Generated | 33 | 29 |
| | Original | 20 | 15 |

Fonte: Soni e Wade (2023)

Fikri, Oflazer e Yanıkoğlu (2023) citam em seu artigo dois desafios principais com relação à tarefa de sumarização abstrativa: a forma de avaliar o desempenho da sumarização e qual é um bom objetivo de treinamento. Para buscar a resolução desses problemas, eles propuseram o uso de uma nova medida de avaliação baseada em similaridade semântica entre o texto original e o resumo gerado pelos modelos. A similaridade semântica resolve o problema de avaliação das métricas Rouge, que buscam a ocorrência de pares de n-gramas do resumo no texto original. Para obter as medidas de similaridade foi utilizado um modelo BERT treinado em dois conjuntos na língua turca: um de similaridade textual semântica e outro de inferência de linguagem natural turca.

Com relação à avaliação baseada em similaridade, os autores focaram em duas tarefas principais. A primeira foi similaridade textual semântica, que visa determinar o quão semelhante são dois textos. A segunda é a inferência de linguagem natural, que busca determinar se há uma implicação, contradição ou relação neutra entre as sentenças dadas. Na avaliação dos resultados, foram obtidas altas correlações entre as pontuações previstas pelos modelos e as pontuações de similaridade reais dos datasets de treinamento, que consistiam em dados de similaridade de sentenças. Além disso, foi realizada uma avaliação de correlação das similaridades calculadas pelo ROUGE, BERTScore e o modelo proposto, o BERTurk, além de avaliações humanas. Os resultados indicaram que a avaliação baseada em similaridade semântica do BERTurk se correlaciona melhor com as avaliações humanas do que as métricas tradicionais como a ROUGE.

Os autores também propuseram uma abordagem de aprendizado por reforço utilizando uma política de gradiente auto-crítico. Essa etapa envolveu o treinamento de um modelo de linguagem, no caso o mT5, que é uma variante multilíngue do T5. Nesse caso, o mT5 atua como um agente que interage com o ambiente para prever a próxima palavra na sequência e observa como recompensa a similaridade semântica do resumo gerado e do resumo original. A similaridade semântica é usada como um sinal de recompensa para treinar o modelo para gerar resumos que são semanticamente semelhantes ao texto original. Dentre os reforços utilizados, como o ROUGE, o BERTScore e o BERTurk para similaridade semântica, o reforço utilizando similaridade semântica se mostrou o melhor paradigma, com os sumários gerados por esse modelo recebendo as melhores avaliações

quantitativas e humanas.

Veen et al. (2023) realizaram um estudo sobre a aplicação de LLMs na sumarização de textos clínicos, com o intuito de investigar se modelos de linguagem podem superar especialistas humanos em tarefas de sumarização de documentos médicos. Os autores destacam que a documentação clínica é uma tarefa que consome muito tempo dos profissionais de saúde, contribuindo para a sobrecarga de trabalho e o *burnout* dos funcionários envolvidos.

No estudo, foram avaliados oito modelos de LLM, incluindo tanto modelos de código aberto quanto proprietários, como o GPT-3.5 e o GPT-4. Os modelos foram aplicados em seis conjuntos de dados e quatro tarefas distintas de sumarização clínica: relatórios radiológicos, perguntas de pacientes, notas de progresso e diálogos entre médico e paciente. Para adaptar os modelos às tarefas específicas, os autores utilizaram a aprendizagem em contexto (*In-Context Learning* - ICL) e métodos de quantização de modelos como o (*Quantized Low-Rank Adaptation* - QLoRA). A ICL envolve fornecer exemplos no *prompt* para orientar o modelo, enquanto a QLoRA é uma técnica de redução de tamanho do modelo que permite adaptar modelos grandes a recursos computacionais limitados, conforme explicado na Seção 2.3.2.

Os resultados quantitativos mostraram que a adaptação dos modelos melhorou significativamente o desempenho em comparação com o *zero-shot prompting* - método que faz a solicitação da tarefa via *prompt* sem fornecer nenhum exemplo. Dentre os modelos avaliados, o GPT-4 com ICL utilizando o máximo de exemplos permitidos pelo contexto obteve os melhores resultados em todas as tarefas. Além da avaliação quantitativa, os autores realizaram um estudo com leitores clínicos, envolvendo dez médicos que avaliaram os sumários gerados pelos modelos e por especialistas humanos em termos de completude, correção e concisão. Como resultado, os sumários gerados pelo GPT-4 foram preferidos aos sumários humanos na maioria dos casos, sendo considerados mais completos e corretos.

A Tabela 4 apresenta os resultados do estudo de leitura clínica, mostrando as médias das pontuações atribuídas pelos médicos aos sumários gerados pelo GPT-4 e pelos especialistas humanos em cada uma das três tarefas avaliadas. Os valores na tabela representam a diferença média das pontuações atribuídas ao GPT-4 em relação aos sumários humanos, em uma escala de -10 a 10, onde valores positivos indicam preferência pelo GPT-4. É possível observar que o GPT-4 superou os especialistas humanos em termos de completude e correção em todas as tarefas, sendo considerado igualmente conciso.

Tabela 4 – Resultados do estudo de leitura clínica de Veen et al. (2023)

| **Tarefa** | **Completude** | **Correção** | **Concisão** |
|---|---|---|---|
| Relatórios Radiológicos | 2,8 | 1,7 | 0,0 |
| Perguntas de Pacientes | 1,6 | 0,6 | 0,6 |
| Notas de Progresso | 2,6 | 0,4 | 0,6 |

Contudo, os autores ressaltam que tanto os modelos quanto os humanos enfrentaram problemas na tarefa de sumarização clínica. Foram identificados casos em que os modelos cometeram erros de alucinação incluindo informações incorretas nos sumários, assim como especialistas humanos omitiram informações importantes ou cometeram erros. Os autores também analisaram a correlação entre as métricas quantitativas de PLN e as avaliações dos médicos, concluindo que as métricas tradicionais, como Rouge, BLEU e BERTScore não capturam totalmente a qualidade dos sumários do ponto de vista clínico, destacando a importância de avaliações humanas na validação de modelos para aplicações clínicas.

Dessa forma, foi concluído que os LLMs, quando adaptados adequadamente, podem superar especialistas humanos em tarefas de sumarização de textos clínicos, representando uma oportunidade para reduzir a carga de trabalho dos profissionais de saúde e melhorar a eficiência na documentação. Os autores sugerem que a integração desses modelos nos fluxos de trabalho clínicos pode permitir que os médicos dediquem mais tempo ao atendimento direto ao paciente e aos aspectos humanos da medicina.

Neste trabalho, para a criação do *pipeline* de sumarização, foi utilizado o BERTopic, que emprega abordagens de redução de dimensionalidade com agrupamento utilizando UMAP e HDBSCan, assim como feito por Liu e Healey (2023), mas no contexto do domínio da política brasileira com *tweets* em português. Para a sumarização multi-documento, foram utilizados diversos modelos *open source*, como diferentes versões do Llama, Bode e Mistral, comparando-os quantitativamente utilizando a métrica BERTScore e qualitativamente através de uma pesquisa com voluntários humanos.

## 3.2 Sumarização de textos de redes sociais com LLM

Pereira, Nogueira e Lotufo (2024) propuseram um método de sumarização de textos de mídias sociais no contexto de gerenciamento de emergências utilizando LLMs. Os autores introduziram uma abordagem que combina algoritmos avançados de busca com LLMs para gerar resumos concisos e relevantes com base em consultas do usuário. Inicialmente, os autores utilizaram algoritmos como o BM25 (ROBERTSON; ZARAGOZA et al., 2009) e o reranqueador monoT5 (NOGUEIRA et al., 2020) para filtrar os documentos mais pertinentes, que em seguida foram resumidos utilizando os modelos GPT-3.5-turbo e GPT-4. Os resultados demonstraram que a integração do reranqueador monoT5 com o

GPT-3.5-turbo reduziu significativamente a redundância e aumentou a abrangência dos resumos.

Para realizar os experimentos, os autores utilizaram um conjunto de dados que abrangeu 18 eventos de crise significativos, como eventos de incêndios florestais, furacões e inundações, combinando informações de várias plataformas como Twitter, Facebook, Reddit e sites de notícias online. O objetivo da tarefa era analisar fluxos diários de dados e criar resumos com base em requisitos informacionais específicos, gerando fatos a partir dos itens de dados lançados em cada dia.

Os resultados dos experimentos podem ser vistos na Tabela 5, onde a configuração que utilizou o monoT5 com o GPT-3.5-turbo obteve o melhor desempenho em termos de BERTScore e ROUGE-2 quando comparada às referências do NIST. A inclusão do GPT-4 melhorou ainda mais o desempenho da sumarização, mostrando um maior alinhamento com o estilo e a qualidade esperados em referências da Wikipédia[1].

Tabela 5 – Resultados da avaliação automática dos resumos

| **Configuração** | **NIST** | | **Wikipédia** | |
|---|---|---|---|---|
| | BERTScore | ROUGE-2 | BERTScore | ROUGE-2 |
| BM25 + GPT-3.5-turbo | 0,642 | 0,318 | 0,481 | 0,032 |
| MonoT5 + GPT-3.5-turbo | **0,668** | **0,416** | 0,471 | 0,035 |
| MonoT5 + GPT-4 | 0,645 | 0,353 | **0,488** | **0,035** |

Os autores também realizaram uma avaliação manual para avaliar a redundância e a abrangência dos resumos. Conforme mostrado na Tabela 6, a configuração que utilizou o monoT5 com o GPT-3.5-turbo apresentou maior abrangência na captura de fatos relevantes, apesar de ter havido um aumento na redundância, indicando uma tendência do modelo a incluir informações repetitivas.

Tabela 6 – Resultados da avaliação manual de redundância e abrangência

| **Configuração** | **Redundância** | **Abrangência** |
|---|---|---|
| BM25 + GPT-3.5-turbo | 0,409 | 0,126 |
| MonoT5 + GPT-3.5-turbo | 0,630 | 0,201 |

A partir da pesquisa, concluiu-se que a combinação de algoritmos de recuperação aliados a LLMs aprimora a qualidade da sumarização automática no cenário de análises de eventos de crise. Foi destacado que os modelos de linguagem mais recentes são capazes de processar contextos significativamente maiores sem a necessidade de grandes quantidades de dados de treinamento, o que aumenta a eficácia da sumarização.

Singh et al. (2024) exploraram em seu trabalho o uso de LLMs na extração e sumarização de evidências de ideação suicida em conteúdo de mídias sociais. Os autores

[1] https://pt.wikipedia.org/

utilizaram os modelos Mixtral7Bx8 (JIANG et al., 2024) e Tulu-2-DPO-70B (IVISON et al., 2023), aplicando diferentes estratégias de *prompting* para aprimorar a eficácia na extração e sumarização do conteúdo relevante.

A metodologia proposta envolveu a aplicação de estratégias de *zero-shot* e *few-shot* learning, avaliando a eficácia das abordagens de *Chain-of-Thought* e *Direct Prompting*. Para isso, os autores elaboraram *prompts* detalhados que orientavam os modelos a identificar e extrair trechos de texto que indicassem ideação suicida, considerando seis aspectos cruciais: emoções, cognições, comportamento e motivação, suporte interpessoal e social, questões relacionadas à saúde mental e fatores de risco adicionais.

Na estratégia de *zero-shot prompting*, os modelos foram fornecidos apenas com instruções claras sobre a tarefa, sem exemplos adicionais. Já na abordagem de *few-shot prompting*, foram incorporados exemplos de contexto (demonstrações) para auxiliar no *in-context learning*. As estratégias de *prompting* avaliadas incluíram o *Direct Prompting*, que apresenta instruções diretas e explícitas, e o *Chain-of-Thought Prompting*, que induz os modelos a realizarem um processo de raciocínio passo a passo para lidar com tarefas complexas.

Os resultados demonstraram que, avaliando a tarefa de extração de evidências, ou seja, quantas evidências corretas de ideação suicida foram extraídas pelos modelos, o Tulu-2-DPO-70B se beneficiou significativamente da estratégia *few-shot* com *Chain-of-Thought Prompting*, alcançando o maior *recall* de 0,943 e uma pontuação F1 de 0,929, conforme apresentado na Tabela 7. Isso indica que a incorporação de um raciocínio estruturado e exemplos contextuais pode melhorar a capacidade de identificação dos LLMs. Para a avaliação, foram utilizadas métricas clássicas de classificação binária, em que o *recall* é a proporção de evidências relevantes corretamente extraídas, *precision* é a proporção de evidências extraídas que são realmente relevantes e a *f1-score* é a combinação harmônica de precisão e recall.

Tabela 7 – Desempenho dos modelos LLM na tarefa de extração de evidências segundo Singh et al. (2024)

| **Modelo** | **Estratégia** | **Recall** | **Precisão** | **F1-score** |
|---|---|---|---|---|
| Mixtral7bx8 | Zero-shot | 0,914 | 0,911 | 0,912 |
| Mixtral7bx8 | Direct Prompting | 0,914 | 0,907 | 0,910 |
| Tulu-2-DPO-70B | Chain-of-Thought | **0,943** | 0,916 | **0,929** |

Para a tarefa de sumarização de evidências, os autores adotaram uma abordagem *zero-shot*, explorando o impacto da inclusão de metainformações nos *prompts*, como sentimentos, emoções e rótulos de risco de suicídio atribuídos aos usuários. O modelo Mixtral7Bx8 alcançou uma alta pontuação de consistência média de 0,977, indicando uma capacidade aprimorada de gerar resumos precisos e coerentes alinhados com as evidências extraídas. A Tabela 8 apresenta os resultados obtidos nessa tarefa. Para esta metrificação, foram

avalidas a consistência média, que mede o alinhamento das informações resumidas com as evidências extraídas e a contradição máxima dos resumos gerados. Os autores não especificaram no artigo qual método específico foi utilizado para chegar nestes valores, porém presume-se que seja algo baseado em *embeddings* ou modelos como o BERTScore.

Tabela 8 – Desempenho dos modelos LLM na tarefa de sumarização de evidências segundo Singh et al. (2024)

| **Modelo** | **Metainformação** | **Consistência Média** | **Contradição Máxima** |
|---|---|---|---|
| Mixtral7bx8 | Não | 0,951 | 0,127 |
| Mixtral7bx8 | Sim | **0,977** | **0,079** |
| Tulu-2-DPO-70B | Sim | 0,966 | 0,107 |

Os autores concluíram que os LLMs possuem um potencial significativo na análise de saúde mental em mídias sociais, especialmente na identificação e sumarização de evidências de ideação suicida. O estudo destacou a eficácia dos LLMs em lidar com dados complexos de saúde mental e a importância da incorporação de metainformações para aprimorar as tarefas de sumarização.

## 3.3 Modelagem de Tópicos

Egger (2022) realizou uma comparação entre quatro técnicas de modelagem de tópicos aplicadas a postagens do X, com o objetivo de avaliar o desempenho dessas técnicas na análise de textos curtos e não estruturados das mídias sociais. Os modelos analisados foram o LDA (*Latent Dirichlet Allocation*) (BLEI; NG; JORDAN, 2003), o NMF (*Non-Negative Matrix Factorization*) (LEE; SEUNG, 1999), o Top2Vec (ANGELOV, 2020) e o BERTopic (GROOTENDORST, 2020). Para a realização do estudo, os autores coletaram 31.800 *tweets* únicos relacionados ao termo *covidtravel*, utilizando a ferramenta Phantombuster. O pré-processamento dos dados variou conforme o modelo a ser aplicado. Para o LDA e o NMF, foram realizadas etapas clássicas de processamento de linguagem natural (PLN), incluindo remoção de *stopwords*, tokenização, *stemming* e lematização. Já para o Top2Vec e o BERTopic, que utilizam abordagens baseadas em *embeddings*, os *tweets* originais foram mantidos sem pré-processamento, já que essas técnicas dependem da estrutura original do texto para gerar representações semânticas adequadas.

Na aplicação do LDA, os autores buscaram os melhores hiperparâmetros, realizando uma busca em grade para otimizar o número de tópicos ($K$) e os parâmetros *alpha* e *beta*, obtendo um modelo com 14 tópicos. No caso do NMF, utilizando a biblioteca Gensim, foram identificados 10 tópicos com base no maior *score* de coerência. Já os modelos Top2Vec e BERTopic não necessitam de uma definição prévia do número de tópicos, pois conseguem identificar automaticamente a quantidade adequada de tópicos com base nos dados de entrada.

Os resultados obtidos foram comparados qualitativamente, considerando a capacidade de cada modelo em identificar tópicos coerentes e relevantes. Conforme apresentado na Tabela 9, o BERTopic e o NMF demonstraram melhor desempenho na extração de tópicos significativos dos *tweets* analisados, superando o LDA e o Top2Vec. Os autores citam que o BERTopic se beneficiou do uso de *embeddings* pré-treinados e do algoritmo c-TF-IDF para representar os tópicos, enquanto o NMF se mostrou eficaz ao lidar com textos curtos devido à sua abordagem baseada em álgebra linear e no uso da ponderação TF-IDF.

Tabela 9 – Comparação das técnicas de modelagem de tópicos

| **Modelo** | **Desempenho** |
|---|---|
| LDA | Gerou tópicos universais e pouco relevantes, com sobreposição de temas e necessidade de definição prévia de hiperparâmetros. |
| NMF | Apresentou tópicos claros e distintos, alinhados com a interpretação humana, superando o LDA em geral. |
| Top2Vec | Identificou tópicos múltiplos e sobrepostos, com dificuldade em interpretar alguns resultados devido à mistura de conceitos. |
| BERTopic | Demonstrou melhor desempenho geral, com extração de tópicos específicos e relevantes, aproveitando a abordagem de *embeddings* e o algoritmo c-TF-IDF. |

Fonte: Egger (2022)

Os autores também ressaltaram que apesar de o LDA ser amplamente utilizado em pesquisas de ciências sociais, seu desempenho não foi satisfatório ao lidar com textos curtos e não estruturados. Já o NMF mostrou-se mais adequado para esse tipo de dado, não exigindo conhecimento prévio do domínio e sendo eficiente no processamento de textos curtos. Com relação ao Top2Vec e ao BERTopic, ambos utilizam abordagens baseadas em *embeddings*, o que permite capturar informações semânticas mais profundas dos textos. O BERTopic se destacou por sua versatilidade e estabilidade em diferentes domínios de trabalho, além de oferecer recursos como redução hierárquica de tópicos e suporte a múltiplos idiomas através da geração de *embeddings* com modelos pré-treinados para cada língua.

Ruocco Yuqian Zhuang (2024) desenvolveram uma plataforma que utiliza LLMs para conectar dados de mídias sociais a tópicos específicos de domínio, com aplicação na saúde mental de estudantes universitários. A plataforma propõe uma abordagem inovadora ao analisar grandes volumes de dados de texto não estruturados de mídias sociais, como o *Reddit*[2], identificando temas coerentes e partes dos dados relacionadas a um tema de interesse (TOI) definido pelo usuário.

No núcleo da plataforma é utilizado o BERTopic, que permitiu agrupar postagens de mídias sociais em coleções de palavras que representam tópicos distintos, capturando

[2] https://www.reddit.com/

assim os temas emergentes nas discussões online. Uma característica chave do estudo é a capacidade de mapear as palavras dos tópicos para sentenças completas, o que aprimora a capacidade de realizar operações de similaridade semântica em relação ao tema de interesse proposto. Para realizar essa similaridade semântica, os autores utilizaram o SBERT (REIMERS; GUREVYCH, 2019), um modelo baseado no BERT que produz *embeddings* de sentenças semanticamente significativas. Ao gerar as *embeddings* tanto das sentenças do *corpus* principal quanto do texto que define o TOI, o algoritmo compara essas *embeddings* e identifica as sentenças que estão semanticamente próximas ao TOI. Isso permite identificar partes relevantes do texto não estruturado que estejam relacionadas ao tema de interesse, mesmo que não contenham palavras-chave explícitas.

Dois estudos de caso focados na identificação de sinais relacionados à depressão entre estudantes universitários foram conduzidos pelos autores. No primeiro estudo, eles compararam dados do *Reddit* de uma universidade canadense em dois períodos distintos (2020 e 2022) para analisar como os temas relacionados à COVID-19 e à depressão evoluíram ao longo do tempo. No segundo, compararam dados de duas universidades diferentes para identificar diferenças regionais nos temas discutidos, utilizando um pré-processamento adicional para filtrar tópicos relacionados a habitação.

Como resultados, a plataforma foi capaz de identificar temas correlacionados ao TOI, como preocupações com a saúde mental durante a pandemia e questões relacionadas à habitação estudantil. A utilização do BERTopic permitiu agrupar as discussões em tópicos coerentes, enquanto o uso de LLMs – no caso o SBERT – possibilitou identificar sentenças relevantes ao TOI dentro de cada tópico. Foi destacado que a capacidade de mapear palavras para sentenças inteiras proporciona uma compreensão mais profunda dos temas emergentes e de sua relação com o TOI.

Gökçimen (2024) realizaram um estudo sobre o discurso sobre mudanças climáticas em mídias sociais e blogs utilizando uma análise de modelagem de tópicos. Os autores destacaram a importância de entender a percepção pública e a conscientização sobre as mudanças climáticas para o desenvolvimento de políticas eficazes que busquem reduzir seus efeitos. Para isso, eles coletaram um conjunto de dados composto por 10.000 documentos relacionados às mudanças climáticas provenientes de mídias sociais como X, Facebook e Instagram, além de plataformas de blogs como Medium, WordPress e Tumblr. Após a coleta, os dados passaram por etapas de pré-processamento, incluindo normalização de texto, remoção de números e pontuação, remoção de *stop words*, lematização e tokenização.

Os autores aplicaram dois métodos de modelagem de tópicos: O LDA e o BERTopic. O LDA sendo um método probabilístico que busca identificar tópicos com base na distribuição de palavras e o BERTopic combinando *embeddings* com técnicas de agrupamento para identificar tópicos de forma mais precisa. Para o BERTopic, os autores compararam diferentes técnicas de extração de palavras-chave e representação de texto, incluindo

OpenAI, Maximal Marginal Relevance (MMR) e KeyBERT. A performance dos modelos foi avaliada utilizando métricas de coerência de tópico NPMI – uma variação normalizada do PMI (CHURCH; HANKS, 1990), coerência de tópico (C_v) (RÖDER; BOTH; HINNEBURG, 2015) e diversidade de tópico[3].

Os resultados indicaram que o BERTopic com as extrações de palavras chaves e representações textuais da OpenAI obteve o melhor desempenho, apresentando maior coerência e diversidade de tópicos. Conforme mostrado na Tabela 10, o BERTopic com OpenAI superou os demais modelos em termos de métricas de desempenho.

Tabela 10 – Resultados das métricas de desempenho para os modelos

| **Modelos** | **Coerência de Tópico NPMI** | **Coerência de Tópico $C_v$** | **Diversidade de Tópico** |
|---|---|---|---|
| NMF | 0,0480 | 0,5164 | 0,7500 |
| LDA | 0,0510 | 0,5220 | 0,8100 |
| BERTopic com KeyBERT | 0,1063 | 0,6521 | 0,8300 |
| BERTopic com MMR | 0,1240 | 0,6690 | 0,8900 |
| BERTopic com OpenAI | 0,0978 | 0,7348 | 0,9100 |

Fonte: Gökçimen (2024)

Foram identificados os principais tópicos discutidos pelo público sobre mudanças climáticas, incluindo sustentabilidade, gestão de recursos, impacto ambiental, reciclagem e transição energética. A análise também mostrou que as mídias sociais e blogs desempenham um papel importante na formação da percepção pública sobre o tema. Também foram utilizadas técnicas de similaridade de sentenças para identificar semelhanças nos comentários e determinar a que categoria de tópico eles pertenciam. A similaridade foi medida utilizando a técnica de similaridade cosseno entre *embeddings* de sentenças, permitindo agrupar comentários semelhantes e entender melhor as preocupações e sentimentos do público.

## 3.4 Sumarização de documentos utilizando métodos baseados em Map Reduce

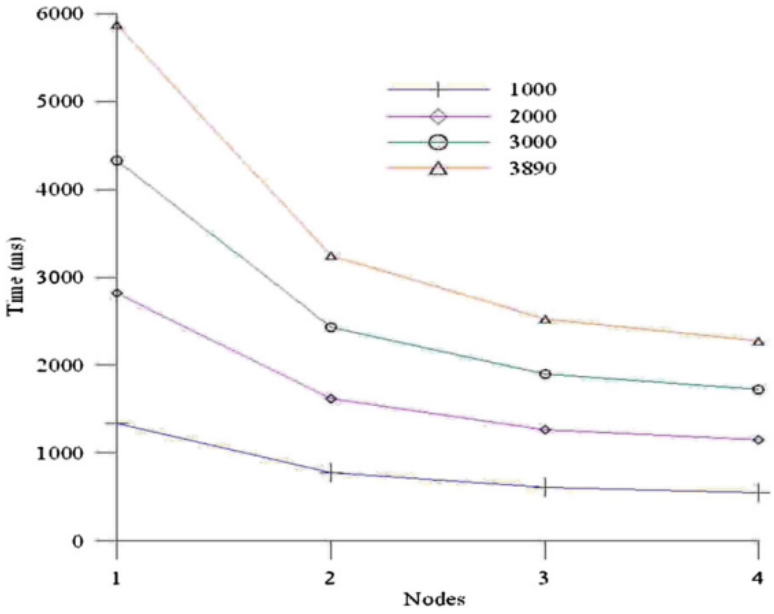
Nagwani (2015) propôs um *framework* para sumarização de grandes coleções de texto utilizando modelagem de tópicos e clusterização baseada na estrutura MapReduce. Os autores citam que a sumarização de grandes volumes de texto é um problema desafiador e demorado, especialmente quando se considera o cálculo de similaridade semântica no processo de sumarização extrativa. Com o objetivo de superar essas limitações, eles apresentaram uma abordagem que utiliza o *framework* MapReduce, uma tecnologia comumente utilizada para lidar com *Big Data*.

[3] <https://github.com/mind-Lab/octis>

O método proposto envolve a clusterização baseada em similaridade semântica e modelagem de tópicos usando LDA (*Latent Dirichlet Allocation*) para sumarizar grandes coleções de texto de forma extrativa. A tarefa de sumarização é realizada em quatro etapas principais. Inicialmente, ocorre a *clusterização* de documentos, onde os documentos são agrupados utilizando o algoritmo de clusterização *K-means*, com o objetivo de agrupar documentos similares e facilitar o processo de sumarização. Na sequência, é aplicada a modelagem de tópicos usando LDA em cada *cluster* de documentos, para extrair os tópicos e termos relevantes presentes em cada grupo. Posteriormente, os tópicos extraídos são processados para identificar os termos frequentes e seus termos semanticamente similares, utilizando a API WordNet[4]. Por fim, são selecionadas as sentenças dos documentos originais que contêm os termos frequentes e seus termos semanticamente similares, compondo assim o sumário final após a remoção de sentenças duplicadas.

Para avaliar o desempenho do sistema, os autores conduziram experimentos utilizando um *corpus* de cerca de 4000 casos legais disponíveis no repositório de aprendizado de máquina UCI. O número de nós no *framework* MapReduce foi variado de um a quatro, e foi realizada uma medição do tempo necessário para gerar sumários a partir de coleções de texto de diferentes tamanhos. Os resultados de redução de tempo são ilustrados na Figura 8, onde é possível ver que o tempo de processamento diminui à medida que o número de nós aumenta, indicando uma melhoria significativa na escalabilidade.

Figura 8 – Tempo de sumarização em ms dos textos do trabalho




Fonte: Nagwani (2015)

Zhou et al. (2024) propuseram o LLM×MapReduce, um *framework* para processamento de sequências longas utilizando modelos de linguagem de grande porte (LLMs). O

[4] https://wordnet.princeton.edu/

principal objetivo deste trabalho é permitir que LLMs com janelas de contexto limitadas possam processar textos extremamente longos sem a necessidade de treinamento adicional nos modelos. Para isso, os autores utilizaram uma estratégia de divisão e conquista, como explicado na seção 2.4, que é semelhante ao paradigma MapReduce, onde o texto longo é dividido em vários pedaços menores chamados de *chunks*, que são processados individualmente pelo LLM (fase de mapeamento), e posteriormente, as respostas intermediárias são agregadas para produzir a resposta final (fase de redução).

Entretanto, eles citam que dividir o texto em *chunks* pode resultar na perda de informações de longo alcance, essenciais para a compreensão completa do documento. Sendo assim, foram identificados dois principais desafios decorrentes dessa divisão: a dependência entre *chunks* (*inter-chunk dependency*) e os conflitos entre *chunks* (*inter-chunk conflict*). Para lidar com a dependência entre *chunks*, os autores desenvolveram um protocolo de informação estruturada, que define o tipo de informação que deve ser extraída e passada de um estágio para outro. Isso inclui a extração de informações-chave, o raciocínio utilizado para chegar à resposta intermediária, a resposta em si e um *score* de confiança. Já para resolver os conflitos entre *chunks*, foi proposto um mecanismo de calibração de confiança no contexto (*in-context confidence calibration*), que permite ao modelo atribuir um *score* de confiança a cada resposta intermediária, auxiliando na resolução de conflitos durante a agregação das respostas.

Nos experimentos realizados, o LLM×MapReduce utilizando o Llama3-70B-Instruct e o Qwen2-72B-Instruct demonstrou desempenho superior a LLMs de contexto longo, tanto de código aberto quanto comerciais em *benchmarks* de processamento de textos longos. Particularmente no *benchmark InfiniteBench* (ZHANG et al., 2024a) o LLM×MapReduce alcançou uma pontuação média de 68,66, superando modelos como o GPT-4 e o Claude 2, que obtiveram 57,34 e 51,62 pontos respectivamente. Foi demonstrado também que o LLM×MapReduce é mais eficiente em termos de latência de inferência e uso de recursos computacionais quando comparado a outros *frameworks* de divisão e conquista, como o LongAgent (ZHAO et al., 2024) e o Chain-of-Agents (ZHANG et al., 2024b). Em experimentos conduzidos com sequências de 128 mil *tokens*, o LLM×MapReduce alcançou uma latência média de inferência de aproximadamente 150 segundos utilizando 2 GPUs NVIDIA A100 (80 GB), enquanto o LongAgent e o Chain-of-Agents apresentaram latências superiores a 300 segundos nas mesmas condições. Além disso, enquanto métodos convencionais de decodificação requerem pelo menos 4 GPUs para processar sequências dessa magnitude, o LLM×MapReduce manteve eficiência e desempenho elevados com a utilização de menos recursos computacionais.

## 3.5 Avaliação de sumários com BERTScore

O BERTScore foi proposto inicialmente por Zhang et al. (2020b) como uma nova métrica automática para avaliação de geração de texto que utiliza *embeddings* contextuais do modelo BERT. Os autores citam as limitações das métricas tradicionais, como o BLEU (PAPINENI et al., 2002) e o ROUGE (LIN, 2004), que se baseiam na sobreposição exata de $n$-gramas e podem não capturar adequadamente a similaridade semântica entre textos que utilizam diferentes expressões lexicais para transmitir o mesmo significado.

Conforme explicado na seção 2.1.2, O BERTScore calcula a similaridade entre cada *token* da sentença candidata e cada *token* da sentença de referência utilizando *embeddings* contextuais, permitindo capturar semelhanças semânticas mesmo quando há paráfrases ou variações lexicais. Foi constatado que o uso de *embeddings* contextuais melhora a avaliação de sistemas de geração de texto, correlacionando-se melhor com as avaliações humanas.

Para validar o BERTScore, os autores realizaram experimentos em tarefas de tradução automática e geração de legendas para imagens, utilizando saídas de 363 sistemas diferentes. Como resultados, o BERTScore apresentou correlações mais altas com as avaliações humanas quando comparado a métricas tradicionais, tanto em nível de segmento quanto em nível de sistema. Na Tabela 11, são apresentados os coeficientes de correlação de Pearson entre as métricas e as avaliações humanas no conjunto de dados WMT18, para diferentes pares de idiomas. É possível observar que o BERTScore, tanto na métrica de precisão quanto na métrica de *recall* e F1, apresenta correlações superiores em vários casos.

Tabela 11 – Correlação de Pearson entre as métricas e as avaliações humanas no WMT18

| **Métrica** | **en-cs** | **en-de** | **en-et** | **en-fi** | **en-ru** | **en-tr** | **en-zh** |
|---|---|---|---|---|---|---|---|
| BLEU | .970/.995 | .971/.981 | .986/.975 | .973/.962 | .979/.983 | .657/.826 | .978/.947 |
| BERTScore (P) | .980/.994 | .998/.988 | .990/.981 | .995/.957 | .982/.990 | .791/.935 | .981/.954 |
| BERTScore (R) | **.998**/**.997** | .997/.990 | .986/.980 | **.997**/.980 | **.995**/.989 | .054/.879 | **.990**/.976 |
| BERTScore (F1) | .990/.997 | **.999**/**.989** | **.990**/**.982** | .998/**.972** | .990/**.990** | .499/.908 | .988/.967 |

Foi também destacado pelos autores que, apesar de utilizar modelos pré-treinados grandes, como o BERT, o BERTScore é relativamente eficiente computacionalmente, podendo ser calculado de forma rápida em conjuntos de validação e teste, o que o torna viável para uso prático na avaliação de sistemas de geração de texto. Entretanto, eles apontaram que não há uma configuração única do BERTScore que supere claramente todas as outras em todos os cenários. A escolha do modelo de *embeddings* e dos parâmetros pode variar conforme o domínio e o idioma, sendo importante considerar as características específicas de cada aplicação ao utilizar o BERTScore.

Deutsch Rotem Dror (2021) realizaram uma análise estatística das métricas de avaliação de sumarização utilizando métodos de reamostragem. Os autores destacam que a

qualidade de uma métrica de avaliação de sumarização é quantificada calculando a correlação entre suas pontuações e as anotações humanas em um grande número de sumários. Contudo, eles citam que no momento do estudo não está claro quão precisas são essas estimativas de correlação, nem se as diferenças entre as correlações de duas métricas refletem uma diferença real ou se são acaso. Eles citam que embora seja comum avaliar a qualidade de uma métrica calculando sua correlação com as anotações humanas, há incertezas sobre a precisão dessas estimativas e sobre a significância das diferenças entre métricas.

Além do BERTScore, explicado na seção 2.1.2, os autores utilizaram uma métrica de avaliação automática chamada QAEval, proposta no trabalho de Deutsch, Bedrax-Weiss e Roth (2021). O QAEval é uma métrica de avaliação de sumários baseada em Perguntas e Respostas (QA). Ao invés de comparar apenas a similaridade de texto entre sumário candidato e sumário de referência, o QAEval gera perguntas sobre partes-chave da informação presente no sumário de referência e verifica se essas perguntas podem ser corretamente respondidas ao ler o sumário candidato. Assim, ele avalia diretamente se a informação relevante está de fato representada no sumário avaliado, e não apenas termos ou expressões coincidentes. Os autores separam dois modelos de linguagem principais nesta avaliação: os modelos de geração de perguntas (por exemplo, o BART ou T5) que é treinado especificamente para gerar uma pergunta que seja respondida exatamente pelo trecho de texto fornecido, e o modelo de perguntas e respostas (QA) que pode ser o BERT, ELECTRA ou T5, também treinado em um *dataset* de QA, que, ao receber uma pergunta gerada e o sumário candidato, tenta identificar se o sumário contém a resposta e, caso contenha, extrai a resposta correta. O resultado disto é um valor entre 0 e 1. Para cada pergunta gerada a partir do sumário de referência, o modelo de QA tenta respondê-la no sumário candidato. Assim, é medida se a resposta está correta (via F1, por exemplo) e é calculada a média dessas pontuações. Um valor próximo a 0 significa que o sumário candidato não conseguiu responder corretamente nenhuma pergunta, ou seja, praticamente nenhuma informação do sumário de referência foi encontrada, enquanto 1 indica que ele respondeu corretamente todas as perguntas.

Para abordar essa questão, eles aplicaram métodos estatísticos baseados em reamostragem (*bootstrap* e permutação) para calcular intervalos de confiança e realizar testes de hipótese, permitindo uma análise mais robusta das métricas de avaliação. Ao aplicar esses métodos a diversas métricas automáticas em três conjuntos de anotações humanas, foi descoberto que os intervalos de confiança para as correlações entre as métricas automáticas e as avaliações humanas são bastante amplos, especialmente ao nível do sistema. Isso indica uma incerteza significativa sobre quão bem as métricas automáticas replicam as avaliações humanas. Além disso, ao comparar as métricas utilizando testes de hipótese, foi encontrado que, embora muitas métricas não apresentem diferenças estatisticamente significativas em relação ao ROUGE, as métricas QAEval e BERTScore se destacaram em

alguns cenários de avaliação. Essas métricas apresentaram correlações significativamente maiores com as anotações humanas, sugerindo que podem ser melhores indicadores da qualidade dos sumários do que as métricas tradicionais.

Como resultado, os autores apontam para a necessidade de cautela ao interpretar as correlações entre métricas automáticas e avaliações humanas, devido à incerteza associada a essas estimativas. Recomendou-se que a comunidade de pesquisa em sumarização desenvolva metodologias de avaliação mais robustas, como avaliações humanas mais eficientes ou avaliações práticas específicas para cada tarefa com o intuito de obter uma compreensão mais precisa da qualidade dos sumários gerados automaticamente.

Para este trabalho, por se tratar de *tweets* no domínio da política brasileira e, portanto, em português do Brasil, foi utilizado para a avaliação dos sumários o BERTSCore com o modelo BERTimbau. Por haver uma limitação na quantidade de *tokens* que o modelo aceita, foram realizadas algumas adaptações que estão descritas na seção 4.2.

# Capítulo 4

# ToMAS: Sumarização Abstrativa Multinível de *tweets*

Este Capítulo apresenta o ToMAS: **To**pic-based **M**ultilevel **A**bstractive **S**umarization, método de sumarização multinível de *tweets* produzido neste trabalho. Além do método de sumarização materializado com o ToMAS, descrito na Seção 4.2, uma segunda contribuição deste trabalho é uma nova proposta de medida de avaliação de sumários descrita na Seção 4.2.4.1.

Porém, antes de descrever as contribuições deste trabalho, a seguir são apresentados os materiais, como o *corpus* e as técnicas e ferramentas usadas para o pré-processamento, agrupamento hierárquico, bem como os LLMs utilizados para a criação do ToMAS.

## 4.1 Materiais

Nesta seção, serão especificados os materiais utilizados para a execução dos algoritmos, como a máquina utilizada no processamento e o *corpus*.

### 4.1.1 Máquina para o processamento

Para o processamento da clusterização dos *tweets* e a execução do modelo de linguagem, foi utilizada uma máquina com a seguinte especificação:

- Processador: Ryzen 7 5800X3D, 8 cores / 16 threads.
- Placa de vídeo: Nvidia Geforce RTX 4090 24GB.
- RAM: 64GB DDR4 3800MHz

❏ SO: Ubuntu 24.04 LTS

### 4.1.2 *Corpus* do Interfaces

Os *tweets* utilizados nos experimentos deste trabalho foram obtidos através do grupo de pesquisa Interfaces[1], da Universidade Federal de São Carlos – UFSCar. O grupo, denominado Núcleo de Estudos Sociopolíticos dos Algoritmos e da Inteligência Artificial, visa aplicar algoritmos computacionais e de inteligência artificial para investigar como o poder político opera por meio da modulação algorítmica contemporânea.

No contexto deste projeto, foram utilizados os dados coletados do Twitter. Esse *corpus* é composto por cerca de 224 milhões de *tweets* com menções a hashtags e palavras relacionadas aos principais candidatos à presidência da república na eleição de 2022: Bolsonaro, Ciro Gomes, Lula e Simone Tebet. A coleta ocorreu a partir de 20/06/2022 e se estendeu até 31/01/2023 englobando, portanto, os eventos do dia 08/01/2023 para os quais *queries* específicas de coleta também foram definidas. Esse *corpus* foi coletado com o objetivo de investigar a polaridade política no Brasil. O *corpus* em questão contém mensagens curtas (*tweets*) publicadas no Twitter, que foram analisadas pelos pesquisadores em termos de polaridade política. Na tabela 12 há um exemplo de *tweet* para cada candidato do segundo turno das eleições, coletados no dia 10/10/2023.

Os dados do *corpus* foram organizados com base na *query* utilizada, na data de coleta e no tipo do arquivo de dados. Os arquivos originalmente estão armazenados no formato *parquet*[2], que é um formato de armazenamento de dados eficiente, colunar e de código aberto, usado para armazenar grandes conjuntos de dados. Suas principais vantagens incluem: compactação eficiente, suporte a esquemas complexos, compressão de dados, capacidade de leitura/gravação rápida e integração com várias linguagens de programação. Essas características tornam o formato *parquet* ideal para análise de grandes quantidades de dados e armazenamento de dados em sistemas distribuídos.

Cada uma das bases possui as seguintes colunas:

❏ id: Código de identificação único do *tweet*.

❏ text: Texto do *tweet*.

❏ created_at: Data de criação do *tweet*.

❏ source: Fonte do *tweet* (dispositivo móvel ou website).

❏ lang: Idioma do *tweet*.

❏ conversation_id: Código de identificação para todas as réplicas que envolvem o *tweet*.

[1] https://www.interfaces.ufscar.br/
[2] <https://parquet.apache.org/>

- ❏ like_count: Quantidade de curtidas.
- ❏ retweet_count: Quantidade de retweets.
- ❏ quote_count: Quantidade de quotes.
- ❏ reply_count: Quantidade de replies.
- ❏ type: Tipo (tweeted, retweeted, quoted ou replied_to).
- ❏ refereced_tweet_id: Código de identificação do *tweet* referenciado caso o tipo não seja *tweet*.
- ❏ mentions: Usuários mencionados no texto do *tweet*.
- ❏ urls: URLs no texto do *tweet*.
- ❏ hashtags: Hashtags no texto do *tweet*.
- ❏ author_id: Código de identificação do autor do *tweet*.
- ❏ media_keys: Código de identificação da mídia.

Para a experimentação deste trabalho, foi feita uma seleção de três bases derivadas do *corpus* citado na seção 4.1.2: (i) uma base de *tweets* que citavam o candidato Jair Bolsonaro entre os dias 08/01/2023 e 13/01/2023, (ii) outra base que citava o candidato Luiz Inácio Lula da Silva entre os dias 08/01/2023 e 13/01/2023 e, por último, (iii) uma base de *tweets* que falavam sobre os atos de 08 de janeiro e que foram postados entre os dias 08/01/2023 e 11/01/2023[3]. A faixa de datas foi escolhida de forma a abranger o período conhecido pelos atos antidemocráticos de 08 de janeiro de 2023, conforme citado no Capítulo 1. A separação das bases que citavam cada candidato ou os atos em si já veio nativamente pronta na própria composição do *corpus*, tendo cada uma delas uma pasta separada com seus respectivos arquivos contendo os *tweets*. Na Tabela 12 é possível ver dois exemplos de *tweets* obtidos no *corpus*.

O *corpus* contém um total de 223.955.053 *tweets* coletados de todos os candidatos participantes das eleições, além dos atos antidemocráticos em si. Conforme citado anteriormente, foi realizado um recorte de datas visando cobrir o período conhecido pelos atos antidemocráticos de 8 de janeiro, e utilizadas somente as bases de Lula, Bolsonaro e dos atos em si. Na Tabela 13 é possível visualizar a quantidade original de *tweets* de cada uma das bases B (Bolsonaro), L (Lula) e AA (atos antidemocráticos) e após o recorte de datas.

[3] A base dos atos antidemocráticos foi reduzida em dois dias por conter um número maior de *tweets* por dia do que as bases de Bolsonaro e Lula. Dessa forma, foi necessário realizar um corte da data superior para poder ser possível realizar o agrupamento na VRAM disponível na GPU.

Tabela 12 – Tabela de amostra de exemplos de *tweets* – Bolsonaro e Lula

| Candidato | ID | Texto |
|---|---|---|
| Bolsonaro | 1579260401329770496 | `@Haddad_Fernando` A maior e única obra do Bolsonaro em São Paulo foi esconder Queiroz, o mestre da rachadinha, na casa do advogado dele em Atibaia. Como vocês votam nesses bandidos? |
| Lula | 1579260416949391360 | Lula não será a porta do paraíso, mas com certeza será a saída do inferno. |

Tabela 13 – Quantidade de *tweets* antes e depois do recorte de datas.

| **Base** | **Qtd. total de *tweets*** | **Qtd. de *tweets* após o recorte de datas** |
|---|---|---|
| B | 37.064.348 | 486.480 |
| L | 35.186.340 | 528.925 |
| AA | 4.466.959 | 830.102 |

Fonte: Elaborado pelo autor.

### 4.1.3 Bibliotecas e *Frameworks*

Para este trabalho foram utilizados diversos *frameworks* e bibliotecas baseadas em Python para a execução e processamento do ToMAS. Dentre elas estão:

- Nvidia Rapids[4]: *Framework* completo que fornece as principais bibliotecas utilizadas em ciência de dados com suporte pleno a processamento via GPUs Nvidia. Dentre as bibliotecas utilizadas no projeto com a aceleração do Rapids, estão o HDBSCAN e UMAP(MCINNES; HEALY; MELVILLE, 2020).
- HDBSCAN(CAMPELLO; MOULAVI; SANDER, 2013): Biblioteca do algoritmo clusterização hierárquica, utilizado como parte do pacote Rapids, mas pode ser encontrado também no repositório PyPI[5].
- Pytorch: *Framework* de aprendizado de máquina em Python que oferece suporte a cálculos numéricos com tensores e construção de redes neurais, otimizando operações em GPUs. Utilizado para processar as *embeddings* utilizando o BERTimbau.
- BERTopic[6]: Biblioteca de modelagem de tópicos que utiliza modelos de linguagem baseados em *transformers*, como o BERT, além do HDBSCan para criar clusters densos e o c-TF-IDF para gerar tópicos interpretáveis a partir de dados textuais.
- Transformers[7]: Biblioteca que fornece ferramentas para trabalhar com modelos baseados em transformadores, como BERT, GPT, T5, entre outros. Também permite

[4] <https://rapids.ai/>
[5] <https://pypi.org/project/hdbscan/>
[6] <https://maartengr.github.io/BERTopic/index.html>
[7] <https://pypi.org/project/transformers/>

aceleração via GPU. Para este trabalho, ela foi utilizada para processar o modelo Llama 3.

- ❏ llama-cpp-python[8]: O llama-cpp-python é uma biblioteca que permite rodar modelos LLaMA em Python usando uma implementação eficiente em C++, oferecendo suporte a modelos quantizados em formato gguf, permitindo também processamento via GPU. Utilizado para processar os modelos Llama 2, Bode e Mistral.

- ❏ Langchain[9]: Biblioteca que facilita a criação de aplicativos que integram modelos de linguagem, permitindo combinar LLMs com ferramentas como bancos de dados, APIs e fluxos de trabalho complexos.

- ❏ BertScore[10]: Biblioteca que fornece a ferramenta da métrica de avaliação para comparar textos baseada em modelos de linguagem como BERT, que mede similaridade semântica entre frases.

Todas as bibliotecas, com suas respectivas versões, podem ser encontradas no *github* deste projeto[11].

### 4.1.4 LLMs

Neste trabalho, foram utilizados para a etapa de sumarização multinível os modelos Llama 2 13B, o Llama 3 8B, o Mistral 7B e o Bode 7b. O Llama 2 (TOUVRON et al., 2023) é uma família de modelos de linguagem de larga escala (LLMs) criada pela Meta[12], com modelos que variam de 7 bilhões de parâmetros até 70 bilhões de parâmetros. Sua arquitetura segue o padrão *Transformer*, usando RMSNorm e a função de ativação SwiGLU e é um modelo treinado com 2 trilhões de *tokens* de entrada e possui uma janela de contexto de 4096 *tokens*. O Llama 3 (DUBEY et al., 2024) é uma evolução do Llama 2, mantendo a mesma arquitetura base mas contendo variantes de até 405 bilhões de parâmetros. Esta família foi treinada em um volume maior de dados, antigindo até 15 trilhões de *tokens* de entrada e possui suporte a janela de contextos maiores (de 8192 a 128k *tokens*), tornando-o um modelo mais robusto.

O Bode (GARCIA et al., 2024) é um modelo que se baseia na arquitetura Llama 2 de 7 bilhões de parâmetros, porém adaptado para o português do Brasil. O Bode foi treinado via *LoRA*, um método de *fine-tuning* de baixo custo, com instruções e exemplos em português, o que o torna mais capaz de responder a *prompts* na língua portuguesa de forma mais consistente. Já o Mistral 7B é um modelo de linguagem com 7 bilhões de parâmetros que busca equilibrar alto desempenho com eficiência de inferência. Como

[8] <https://pypi.org/project/llama-cpp-python/>
[9] <https://pypi.org/project/langchain/>
[10] <https://pypi.org/project/bert-score/>
[11] <https://github.com/LALIC-UFSCar/ToMAS>
[12] <https://www.meta.ai/>

diferencial, o Mistral utiliza uma técnica de janela deslizante de atenção chamada SWA, que permite que o modelo consiga lidar com sequências mais extensas de *tokens* sem aumentar exponencialmente o custo computacional da inferência e o GQA (*grouped-query attention*), que reduz a quantidade de memória exigida e acelera a geração de textos utilizando diversos *heads* de atenção, compartilhando as mesmas chaves e valores. O Mistral 7B, assim como o Llama 3 8B, possui uma janela de contexto de 8192 *tokens*.

Na Tabela 14 é possível visualizar as especificações de cada modelo utilizado.

Tabela 14 – Tabela de Modelos e Parâmetros

| Modelo | Parâmetros | Quantização | Lançamento | Licença |
|---|---|---|---|---|
| Llama 2 - 13B | 13 bilhões | PTQ 8-bit | 07/2023 | Open Source |
| Mistral 7B | 7 bilhões | - | 09/2023 | Open Source |
| Bode 7B | 7 bilhões | PTQ 8-bit | 01/2024 | Open Source |
| Llama 3 - 8b | 8 bilhões | - | 04/2024 | Open Source |

Fonte: Elaborado pelo autor.

## 4.2 ToMAS

Esta seção descreve o pipeline proposto para sumarização multinível de *tweets* chamado ToMAS. Inicialmente é realizada uma etapa de coleta e pré-processamento dos dados (descrita na Seção 4.1.2 e Seção 4.2.1) seguida pela extração de tópicos (Seção 4.2.2). A etapa de sumarização multinível, explicada na seção 4.2.3, é realizada repetidamente até que haja apenas um resumo para cada tópico. Esse processo iterativo de geração de sumários é ilustrado na Figura 10. Após a geração dos sumários, é realizada a avaliação quantitativa e qualitativa, conforme descrito na Seção 4.2.4.

Cada uma das etapas presentes na Figura 9 será detalhada nas próximas subseções.

### 4.2.1 Pré-Processamento

Como citado no Capítulo 1, os textos escritos em redes sociais tendem a possuir ruídos e características específicas de escrita. Quando se trata do Twitter, é comum encontrar palavras ou nomenclaturas que são específicas da rede. Dentre elas, estão:

- Menções: Quando um usuário quer se comunicar diretamente com outro, é utilizado o caractere arroba seguido do nome do usuário, como @usuario.
- Hashtags: As chamadas hashtags são utilizadas para categorização e marcação de textos. Por exemplo, se algum usuário estiver tratando de um evento esportivo em seu texto, ele pode adicionar uma hashtag relevante ao assunto, como #copadomundo.

Figura 9 – Pipeline para Sumarização dos tópicos dos *tweets*.


Fonte: Elaborado pelo autor.

- ❏ Emojis: Emojis são pequenos símbolos visuais utilizados para expressar emoções, sentimentos, objetos ou atividades através de mensagens nas redes. Apesar de não serem específicos do Twitter, o uso de emojis para expressões é comum e encontrado frequentemente nos textos obtidos da rede.

Com o intuito de remover os ruídos dos textos, foram utilizados: o Pandas[13], para lidar com os *dataframes* contendo os *tweets*; expressões regulares[14] para encontrar os objetos ruidosos e removê-los; e a biblioteca emoji[15] para transformar determinados emojis nas expressões textuais que eles representam. Através de expressões regulares também foram removidas pontuações repetidas, hashtags, menções a usuários e hiperlinks.

### 4.2.2 Extração de Tópicos

Para a extração dos tópicos dos *tweets* no *corpus* deste trabalho, foi utilizado como base o algoritmo BERTopic[16], desenvolvido por Grootendorst (2020). O BERTopic combina técnicas de geração de *embeddings* de palavras, realiza o agrupamento hierárquico baseado na densidade de palavras com o HDBSCan e, finalmente, extrai tópicos com base na importância das palavras, utilizando uma variação da técnica TF-IDF, chamada c-TF-IDF. O funcionamento detalhado do algoritmo é explicado na Seção 2.2. Devido a personalizações específicas para este trabalho, como a geração de *embeddings* utilizando

[13] <https://pandas.pydata.org/>
[14] <https://docs.python.org/3/library/re.html>
[15] <https://pypi.org/project/emoji/>
[16] <https://maartengr.github.io/BERTopic/index.html>

o BERTopic, a normalização das *embeddings* e a utilização do UMAP e do HDBSCan do *framework* Rapids, foi utilizado da biblioteca original do BERTopic somente a parte do c-TF-IDF que gera as 10 palavras mais representativas de cada tópico com base na ordenação do c-TF-IDF, onde é calculada uma pontuação de relevância para cada palavra do grupo gerado, conforme explicado na Seção 2.2.3.

Para realização da etapa de agrupamento, é necessário que as *embeddings* dos textos tenham sido geradas. Sendo assim, finalizada a etapa de pré-processamento dos dados, foi realizada a etapa de geração das *embeddings* das sentenças dos *tweets*. Isso nos permite obter a representação vetorial de cada *tweet* para realizar o agrupamento posterior. Para esta etapa foi utilizado o *framework* SentenceTransformer[17], que é considerado o estado da arte para geração de *embeddings* para textos e imagens. Os *embeddings* foram gerados através do modelo pré-treinado BERTimbau[18] (SOUZA; NOGUEIRA; LOTUFO, 2019), uma variante do BERT (DEVLIN et al., 2019) treinada em português do Brasil.

Na etapa seguinte, foi realizada uma normalização das *embeddings* para o posterior agrupamento hierárquico. Por se tratar de um algoritmo baseado em densidade hierárquica e a versão utilizada no Rapids não possuir a métrica de cosseno implementada, a métrica euclidiana foi utilizada para a clusterização. Essa métrica mede a distância absoluta no espaço, sendo sensível tanto à direção quanto à magnitude dos vetores. Sem a normalização, *embeddings* de maior magnitude podem influenciar desproporcionalmente o cálculo das distâncias. Desta forma, foi realizada uma normalização vetorial pela norma L2 – também conhecida como normalização euclidiana – dos vetores de *embeddings* com o intuito de eliminar diferenças de escala e magnitude, evitando possíveis vieses causados por vetores de alta magnitude e permitindo que o foco ficasse na distância angular, que é mais adequada para *embeddings* de texto.

Finalizada a etapa de normalização, foi realizada uma redução de dimensionalidade nos vetores de *embeddings*. Originalmente, o BERTimbau gera vetores de 1024 dimensões para cada uma das sentenças. Entretanto, a tarefa de agrupamento para um vetor de tamanho elevado se torna um processo computacionalmente custoso, além da redução de dimensionalidade ajudar na remoção de ruído e redundância, resultando em representações mais compactas e informativas. Para a redução, foi utilizada a biblioteca umap-learn[19], gerando vetores de 512 dimensões para cada sentença processada. Como parâmetros, além do *n_componentes* em 512, o parâmetro *metric* foi definido como *cosine* e o parâmetro *random_state* em 42.

Na etapa seguinte é realizado o agrupamento através do HDBSCAN (MCINNES JOHN HEALY, 2016), um algoritmo hierárquico de agrupamento baseado em densidade que agrupa documentos mais semelhantes entre si com base na estabilidade (uma medida de consistência de um agrupamento à medida que diferentes níveis de densidade

[17] <https://www.sbert.net/>
[18] <https://huggingface.co/neuralmind/bert-large-portuguese-cased>
[19] <https://umap-learn.readthedocs.io/en/latest/>

são escolhidos para agrupar os dados). Para esta etapa, foram definidos os parâmetros *min_samples* – que define o número mínimo de amostras na vizinhança de um ponto para ele ser considerado como um ponto central – como 10, valor marginalmente mais alto que o padrão, que é 5. Essa alteração foi de forma empírica, com o intuito de evitar a geração de muitos grupos com poucos elementos isolados. O parâmetro *cluster_selection_method* ficou em *eom* e o parâmetro *metric* em *euclidean*, valores padrão do HDBSCan.

A etapa final do processo é a seleção de tópicos com base na importância das palavras. Para isso, o autor do BERTopic desenvolveu a técnica chamada c-TF-IDF. Nessa variação do TF-IDF, a importância das palavras é analisada considerando os *clusters* gerados na etapa de agrupamento, aplicando TF-IDF a cada um deles. Esse processo classifica as palavras de acordo com sua relevância para cada grupo gerado, extraindo os principais tópicos de cada agrupamento. Assim, cada grupo – agora chamado de tópico – possui sua lista de termos relevantes a ele e elementos considerados os mais representativos do grupo.

### 4.2.3 Sumarização Multinível

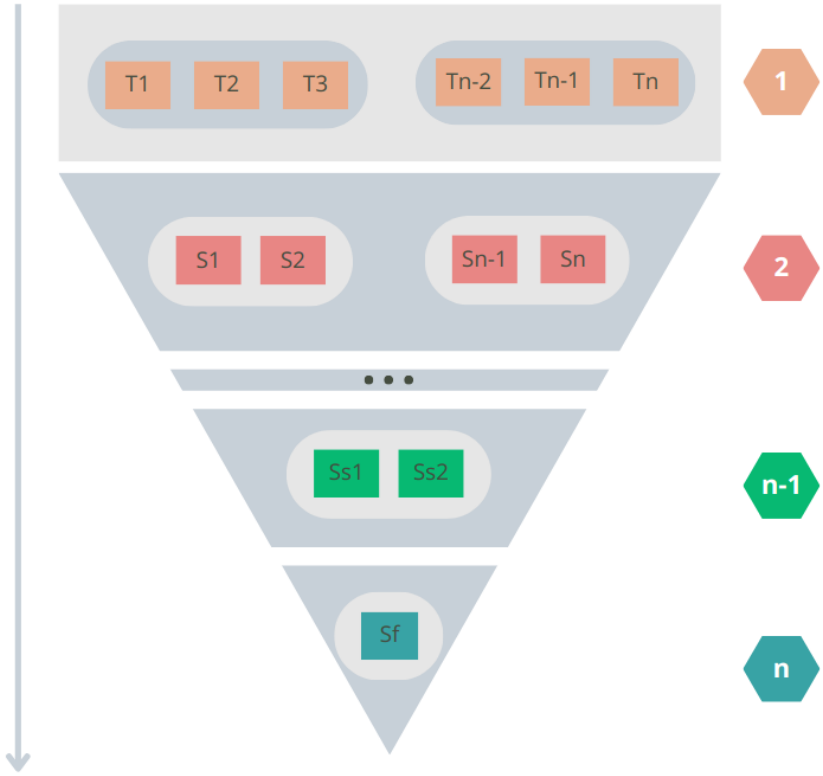
Para a sumarização, foram utilizados os modelos Llama 2 13B, Llama 3.0 8B, Mistral 7B e Bode, sendo este último uma variante do Llama 2 7B treinada em português brasileiro. Devido às limitações de contexto de 4.096 e 8.192 *tokens* de entrada desses modelos e ao alto volume de *tweets*, este trabalho propõe o uso de uma técnica baseada em divisão e conquista, chamada sumarização multinível, ilustrada na Figura 10.

Após a extração de tópicos pelo BERTopic, cada *tweet* é atribuído a um único tópico. Para cada um dos tópicos gerados é realizado o processamento proposto pela sumarização multinível. Conforme visto na Figura 10, os *tweets* de um determinado tópico são concatenados um a um até que a soma de caracteres deles atinja no máximo 1536 palavras (aproximadamente 2000 *tokens*)[20]. Note que este valor ficou consideravelmente abaixo do limite de *input* de 4096 *tokens* do modelo com a menor janela de contexto, que são os modelos da família Llama 2. Este corte foi necessário, pois acima desse valor, os modelos passavam a ter uma tendência recorrente de gerar saídas ruidosas e sem sentido, gerando apenas caracteres aleatórios em forma de texto. A causa disto não foi profundamente estudada, porém, cria-se a hipótese de que os textos utilizados como entrada, por se tratarem de *tweets*, carregam muitos ruídos e, por isso, podem ter influenciado na baixa qualidade da saída quando agregados em maiores quantidades. Ao adicionar um *tweet* a um grupo de 1536 palavras, se a soma de palavras dele extrapolar o limite máximo do grupo, o grupo é fechado e o *tweet* em questão é adicionado no grupo seguinte.

Sendo assim, para cada tópico a ser sumarizado são gerados $n$ grupos de *tweets*. Cada um desses grupos é utilizado como entrada para o modelo sumarizar, gerando na segunda etapa da figura sumários que variam de 1 a $n$. Agora, os sumários gerados são agrupados

[20] Este valor foi obtido empiricamente testando somas de múltiplos de 2, neste caso, 1024 + 512 palavras.

Figura 10 – A arquitetura de sumarização multinível. Na primeira etapa (1), para cada tópico, os $n$ *tweets* (T1 ... Tn) são agrupados para gerar $m$ sumários. Nas etapas seguintes (2 até $k$), os sumários gerados na etapa anterior são agrupados para gerar novos sumários até que reste apenas um resumo final (Sf).




Fonte: Elaborada pelo autor.

assim como na primeira etapa, de forma que cada grupo de sumários não ultrapasse as 1536 palavras. Feito o agrupamento, são gerados novos sumários destes grupos. O processo é repetido até que na etapa final haja apenas um sumário final $Sf$ que represente o tópico em questão. A saída do modelo foi limitada para possuir no máximo 768 *tokens*, sendo este, portanto, o tamanho máximo de cada sumário gerado. O *prompt* utilizado para a geração dos sumários pode ser observado na Figura 11.

Figura 11 – *Prompt* utilizado para a sumarização em todos os LLMs

```
Prompt: [INST] <<SYS>> I am providing you with an 'Input' of a set of texts
    ↪ in Brazilian Portuguese that are separated by ";" between them. Provide
    ↪  as 'Output' a summary in continuous text with your own words also in
    ↪ Brazilian Portuguese, without repeating the original texts and without
    ↪ citing examples, covering the main subjects being mentioned in the
    ↪ input texts briefly and in a general way. Do not form your own opinions
    ↪ ; all content in the summarization must be based on the content found
    ↪ in the texts. When citing a trend or any behavior, use something like '
    ↪ The authors mentioned that (...)' <<SYS>>
Input: '''{tweets}''' [/INST]

Output:
```

Fonte: Elaborado pelo autor.

Como o número de palavras no resumo foi programado para ser sempre a metade ou

menos do tamanho da entrada original, cada rodada reduz o número total de palavras a serem processadas em pelo menos metade, caracterizando um comportamento logarítmico. Assim, podemos estimar o número de iterações conforme a equação 19, onde $N$ é o número total de palavras a serem processadas, $N/1536$ estima o número inicial de grupos formados e o logaritmo indica quantas vezes podemos reduzir esse valor pela metade até atingir um único resumo.

$$\text{Iterations} \approx \log_2\left(\frac{N}{1536}\right) \tag{19}$$

Este processo é repetido para todos os tópicos gerados na etapa de extração de tópicos. Por se tratar de um processo custoso computacionalmente, para este trabalho foram gerados os sumários para os 20 tópicos mais populosos, em número de *tweets*, de cada uma das três bases obtidas, gerando um total de 60 sumários.

Foi observado, em avaliações empíricas realizadas durante as etapas de desenvolvimento, que os sumários gerados pelos modelos testados apresentavam mais detalhes, maior abrangência e melhor redação quando as instruções do *prompt* eram fornecidas em inglês. Contudo, isso ocasionalmente resultava no efeito colateral de o resumo final ser gerado em inglês para alguns tópicos. Para resolver esse problema, uma etapa final de tradução foi realizada nos sumários após o processamento final, com o objetivo de traduzir os textos caso estivessem em inglês, utilizando o mesmo modelo de linguagem empregado na sumarização propriamente dita. Para isso, foi usado um *prompt* extra específico para tradução, como ilustrado na Figura 12.

Figura 12 – *Prompt* utilizado para a tradução dos sumários

```
Prompt: [INST] <<SYS>> Traduza o texto fornecido para portugues do Brasil.
    ↪ Caso ele esteja em portugues, apenas repita o proprio texto como saida.
    ↪ <<SYS>>

Input: "{text}" [/INST]

Output:
```

Fonte: Elaborado pelo autor.

Foram gerados sumários utilizando os principais LLMs *open-source* disponíveis atualmente, como o Llama 2 13B (TOUVRON et al., 2023), o Llama 3.0 e 3.1 com 8B de parâmetros (DUBEY et al., 2024), o Mistral 7B (JIANG et al., 2023) e o Bode 7B (GARCIA et al., 2024). Por se tratar de um computador pessoal, foi necessário lançar mão de modelos quantizados para executar o Llama 2 13B. A quantização, conforme explicado na Subseção 2.3.2, é um método de compressão de modelos que faz um mapeamento de pontos flutuantes em números inteiros, especialmente com 8 bits. Os pesos e as ativações do modelo podem ser quantizados, o que torna a inferência mais rápida e a necessidade de memória computacional menor. O modelo Bode também foi utilizado em sua versão quantizada devido à maior facilidade de manipulação do arquivo em formato *gguf*. Os

demais modelos foram executados em suas formas originais, sem quantização. Na Tabela 14 é possível ver as principais características dos modelos utilizados nos experimentos.

### 4.2.4 Avaliação dos resultados

O pipeline de sumarização multinível de *tweets* proposto neste trabalho, o ToMAS, foi avaliado quantitativa e qualitativamente conforme descrito a seguir.

#### 4.2.4.1 Avaliação quantitativa

Para a avaliação quantitativa dos sumários gerados, foi empregado o BERTScore[21], algoritmo desenvolvido por Zhang et al. (2020b), porém em uma variação inédita proposta neste estudo. O BERTScore é uma métrica de avaliação automática de texto que utiliza o conceito de *embeddings* contextuais de modelos pré-treinados, como o BERT. Diferentemente de outras métricas tradicionalmente aplicadas para avaliar sumários, como o Rouge-Score (LIN, 2004), que se baseiam na correspondência de n-grams entre o texto de referência e o texto candidato, o BERTScore avalia a similaridade semântica entre os textos gerados e de referência, calculando a similaridade do cosseno entre os *embeddings* dos *tokens*. Isso permite que o BERTScore supere as limitações dos métodos baseados em n-grams ao reconhecer paráfrases semânticas e capturar dependências distantes e alterações na ordem das palavras.

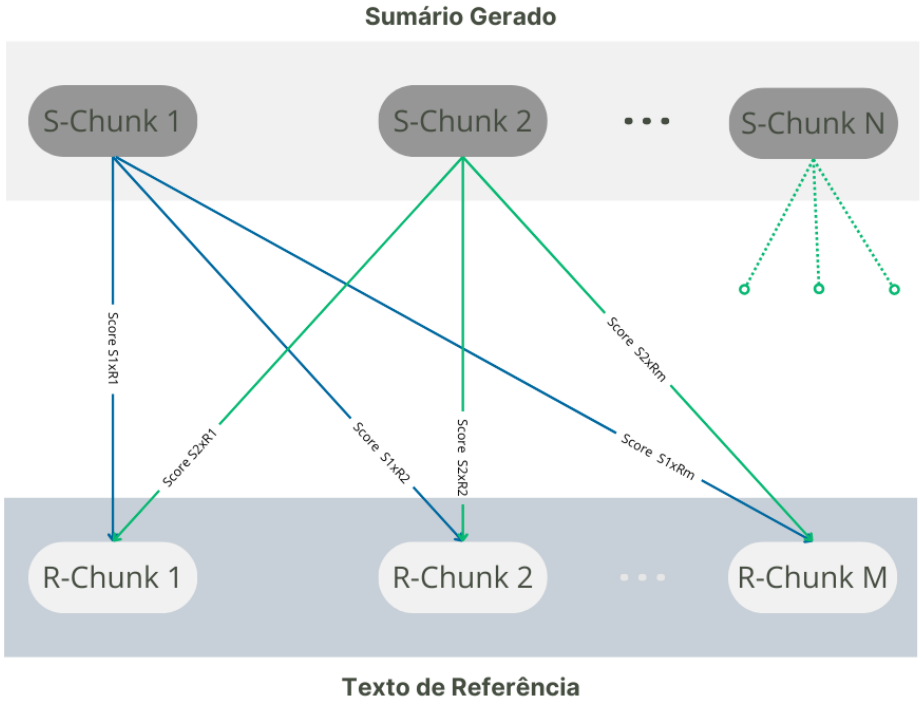
Conforme detalhado na Seção 2.1.2, para calcular o BERTScore, as sentenças de referência e candidata são primeiramente tokenizadas e convertidas em *embeddings* usando o modelo BERT. A similaridade entre os *tokens* é então calculada utilizando a similaridade do cosseno entre os vetores de *embeddings*. Esse processo é realizado para cada *token* da sentença de referência em relação a todos os *tokens* da sentença gerada e vice-versa, permitindo uma correspondência flexível que considera o contexto e a semântica das palavras. O cálculo do BERTScore envolve os conceitos de precisão, recall e F1-Score. Os valores do BERTScore variam de 0 a 1, sendo que valores próximos a 1 indicam maior similaridade entre o texto de referência e o texto gerado.

Neste trabalho, como não existem sumários de referência gerados por humanos para os *tweets* avaliados, o conjunto de *tweets* originais dos tópicos sumarizados foi utilizado como texto de referência. Embora esta não seja a abordagem ideal de avaliação, ela ainda pode fornecer uma métrica útil de similaridade semântica entre os sumários gerados e o texto original, permitindo avaliar a preservação do contexto nos sumários e realizar uma avaliação relativa entre os sumários gerados por diferentes modelos. Para o cálculo do BERTScore, foi utilizado o modelo BERTimbau (SOUZA; NOGUEIRA; LOTUFO, 2020), que é treinado especificamente para o português do Brasil.

[21] <https://github.com/Tiiiger/bert_score>

Existe um obstáculo no cálculo do BERTScore quando são utilizadas sentenças extensas: o limite de *tokens* de entrada dos modelos BERT. O modelo BERTimbau, por exemplo, possui um limite de 512 *tokens* de entrada, tamanho muito menor do que o dos sumários gerados pelos modelos utilizados neste trabalho, que foram limitados a 768 *tokens* de saída. Visando contornar este problema, foi proposta a aplicação de um método de divisão e conquista no cálculo do score, onde tanto os sumários gerados quanto os *tweets* originais são quebrados em pedaços de forma que o modelo consiga comportá-los na avaliação. Essa estratégia é ilustrada na Figura 13.

Figura 13 – Modelo de divisão e conquista para o cálculo do BERTScore.




Fonte: Elaborado pelo autor.

Conforme exemplificado, o sumário gerado para cada tópico é quebrado em N *chunks* e os textos de referência do tópico – *tweets* agregados, neste cenário – são quebrados em M *chunks*. Cada *chunk* possui no máximo 995 caracteres, onde o algoritmo busca sempre inserir cada palavra de forma inteira antes de realizar a contagem para evitar palavras cortadas nos pedaços. Os valores de BERTScore são calculados comparando individualmente cada um dos *chunks* do sumário gerado com todos os *chunks* dos *tweets* de referência, produzindo uma lista de valores contendo todos os *scores* calculados de cada S-Chunk $i$ com cada R-Chunk $j$.

O tamanho máximo de cada *chunk* como sendo de 995 caracteres foi obtido empiricamente, onde foram realizados testes partindo de *chunks* com tamanho máximo de 2048 caracteres e, sempre que o tamanho máximo de input do modelo era excedido, o valor era decrescido. O valor máximo de *chunk* no qual foi possível avaliar todos os sumários sem

erro foi o de 995. Sendo assim, seja $\mathcal{S}$ o conjunto de todos os BERTScores calculados de todos os $N$ S-Chunk com os $M$ R-Chunk, a Equação 20 exemplifica este cálculo:

$$\mathcal{S} = \{\text{BERTScore}_{F1}(S\text{-CHUNK}_i, R\text{-CHUNK}_j) \mid i = 1, \ldots, N;\ j = 1, \ldots, M\} \tag{20}$$

Tendo cada valor pertencente a $\mathcal{S}$ calculado, duas métricas principais são geradas com base nas listas de valores: (i) a média de todos os *scores* gerados na comparação (equação 21), (ii) a média dos $K$ maiores valores de *score* entre as duas sentenças (equação 22), que neste trabalho foi configurado em 20. O BERTScore, originalmente, gera métricas de precisão, *recall* e F-Score, sendo a última a escolhida para as avaliações deste trabalho.

$$\text{MeanScore}_{F1} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} s \tag{21}$$

$$\text{TopKMeanScore}_{F1} = \frac{1}{k} \sum_{s \in \text{TopK}(\mathcal{S},k)} s \tag{22}$$

Durante a aplicação do método, foi possível notar que os valores de score estavam sendo beneficiados por tamanhos de sentenças menores. Modelos que estavam gerando sumários pequenos, mesmo que qualitativamente piores, estavam obtendo um score alto, por vezes até maior do que modelos que geraram sumários notadamente mais detalhados e concisos. Isso ocorreu com o Bode, conforme pode ser visto na seção 5.4.2. Buscando amenizar esse efeito e mitigar o problema, foi proposto o cálculo de um peso que busca recompensar sumários maiores e penalizar os menores. O cálculo do peso é especificado na equação 23. Para cada tópico $Ti$ é calculado um peso $W_{Ti}$, que é a divisão do tamanho da sentença do tópico em questão elevado a um parâmetro $\beta$ ($len\_candidate\_sentence^{\beta}_{Ti}$) pela soma do mesmo valor com o tamanho da sentença de referência ($len\_reference\_sentence_{Ti}$).

$$\text{W}_{Ti} = \frac{\text{len\_candidate\_sentence}^{\beta}_{Ti}}{\text{len\_candidate\_sentence}^{\beta}_{Ti} + \text{len\_reference\_sentence}_{Ti}} \tag{23}$$

O parâmetro $\beta$ tem por intuito mitigar o efeito da diferença de tamanhos entre os sumários gerados e as sentenças de referência. Nas bases utilizadas, existem tópicos que ultrapassam 14 mil *tweets* – que podem conter até 280 caracteres cada um – o que gera sentenças de referência demasiadamente grandes, enquanto os sumários gerados possuem no máximo 768 *tokens* (aproximadamente 590 palavras). Utilizando um $\beta$ de valor 2 (quadrático), foi possível balancear as diferenças de tamanho entre o sumário gerado para o tópico em questão e o texto de referência, além de enfatizar a diferença de tamanho entre os sumários gerados pelos diferentes modelos. Realizado o cálculo do peso, foi gerado o cálculo das métricas BERTScore balanceadas com seu valor. Conforme pode ser visto na equação 24, o cálculo do novo BERTScore foi feito como uma soma ponderada do score original com o score original multiplicado pelo peso $W_{Ti}$. Foi atribuído um peso de 30%

ao score multiplicado pelo peso e de 70% ao score original. Esta nova métrica foi nomeada de BERTScore-p.

$$\text{BERTScore-p}_{Ti} = 0.3 \cdot (\text{original\_score}_{Ti} \cdot \text{W}_{Ti}) + 0.7 \cdot \text{original\_score}_{Ti} \quad (24)$$

Dessa forma, foi possível amenizar o efeito da preferência do BERTScore por sumários menores ao mesmo tempo em que foram preservadas as características semânticas do *score* original.

#### 4.2.4.2 Avaliação qualitativa

Buscando uma forma qualitativa de avaliar os sumários gerados por cada modelo, foi proposta uma consulta aberta ao público, realizada através da ferramenta formulários google[22] e aprovada pelo CEP – Comitê de Ética em Pesquisa em Seres Humanos da UFSCar[23] – CAE 82331024.7.0000.5504. O objetivo da consulta ao público era avaliar qual foi considerado o melhor modelo na geração dos sumários na concepção dos participantes e verificar se há uma correlação entre as avaliações quantitativas geradas pelo BERTScore com as avaliações qualitativas coletadas na pesquisa.

A consulta foi dividida em dois formulários, A e B. Cada um dos formulários possui quatro tópicos ao todo, dois referentes à base do candidato Jair Bolsonaro e dois referentes a base do candidato Luís Inácio Lula da Silva[24]. Os tópicos escolhidos para os formulários A e B são distintos, buscando amplificar o escopo da avaliação e eliminar possíveis vieses de performance dos modelos em cada tópico. A escolha dos tópicos a serem utilizados na pesquisa foi feita com o auxílio de um cientista político da equipe do grupo Interfaces. Para cada um dos tópicos, foram apresentados 4 sumários ao todo, um para cada modelo proposto - Llama 2 13B, Llama 3 8B, Bode e Mistral.

A avaliação foi conduzida na forma de teste cego, onde os sumários eram apresentados em ordem distinta para cada tópico de cada base e o modelo de cada sumário foi omitido. Dessa forma, os participantes avaliaram os sumários sem saber por quais modelos eles haviam sido gerados. Para a coleta de opinião, foi utilizada uma escala Likert de 7 pontos (LIKERT, 1932). Essa escala possui as vantagens de: (i) ter uma maior facilidade de compreensão dos resultados, visto que as respostas são sempre padronizadas; (ii) permitir uma medida de intensidade clara; e (iii) facilitar a comparação entre respostas. Para cada um dos sumários foram apresentadas opções de escolha de pontuação seguindo uma escala de 1 a 7 (LIKERT, 1932) para seis critérios:

- ❑ Fluência/Naturalidade: Refere-se à facilidade de leitura do texto, sem interrupções ou necessidade de grande esforço para o entendimento. Um texto fluente deve

[22] <https://docs.google.com/forms/>
[23] https://www.propq.ufscar.br/pt-br/etica/cep-comite-de-etica-em-pesquisa-em-seres-humanos
[24] Com o intuito de simplificar a avaliação e reduzir o tempo de resposta do formulário, a base referente aos atos antidemocráticos foi excluída da avaliação qualitativa.

ser compreensível à primeira leitura. Avaliado de "totalmente não natural" (1) a "totalmente natural" (7).

- ❏ Uso correto da linguagem: Refere-se à qualidade da escrita do texto, avaliando se o texto está livre de erros ortográficos, sintáticos (p.ex. erros de concordância de gênero ou número), além de erros de coesão. Avaliado de "muitos erros" (1) a "sem erros" (7).

- ❏ Clareza/Legibilidade: Refere-se ao quanto o texto é direto e fácil de entender, evitando jargões desnecessários e construções complexas que dificultem a compreensão, levando em conta fatores como a escolha de palavras e a estrutura das frases. Avaliado de "totalmente confuso" (1) a "totalmente claro" (7).

- ❏ Informatividade: Refere-se à capacidade do texto de transmitir informações, tornando a leitura útil e interessante. É a habilidade do texto de efetivamente comunicar e informar seu público de maneira clara e significativa. Considere que um sumário pode ser informativo mesmo que traga informações com as quais você não concorde, uma vez que pode retratar o viés político de quem escreveu. Avaliado de "totalmente desinformativo" (1) a "totalmente informativo" (7).

- ❏ Redundância: Refere-se à propriedade do texto de repetir desnecessariamente informações. Um texto com baixa redundância é conciso e direto, evitando a repetição de ideias já apresentadas. Avaliado de "totalmente redundante" (1) a "nada redundante" (7).

- ❏ Adequação ao Tópico: Refere-se à adequação do conteúdo do sumário ao tópico, verificada com base nas palavras e *tweets* representativos, bem como na breve explicação sobre o tópico (fornecidos com cada tópico). Um sumário é adequado ao tópico se seu texto aborda o tópico citado de maneira precisa e relevante. Avaliado de "totalmente inadequado" (1) a "totalmente adequado" (7).

Cada participante, portanto, avaliou 4 sumários de 2 tópicos distintos de 2 candidatos, totalizando 16 sumários avaliados. Ao final do formulário, foram realizadas perguntas de cunho socioeconômico dos participantes, como a idade, a cor e etnia, a identidade de gênero, o grau de escolaridade e a área de conhecimento que melhor se encaixa na formação superior, caso se aplique. O tempo total estimado de resposta foi de 40 minutos.

Coletados os resultados, foram geradas métricas de avaliação baseadas nos dados coletados : (i) as médias e os desvios padrão de cada um dos critérios para cada modelo foram gerados, (ii) a média total por modelo agregando todos os critérios, e (iii) uma comparação de tendência das avaliações dos critérios com o valor obtido do BERTScore. Por fim, foram geradas métricas simples de população para o questionário socioeconômico, como

porcentagem de participantes por cor/etnia, por grau de escolaridade, por identidade de gênero, por faixa etária e por área de conhecimento.

A análise qualitativa dos modelos foi realizada considerando-se os critérios de avaliação previamente definidos, sendo:

- ❏ Insatisfatório: valores médios de escala Likert entre 1 e 1,9
- ❏ Ruim: valores médios de escala Likert entre 2,0 e 2,9
- ❏ Regular: valores médios de escala Likert entre 3,0 e 4,9
- ❏ Bom: valores médios de escala Likert entre 5,0 e 5,9
- ❏ Ótimo: valores médios de escala Likert entre 6,0 e 7,0

Onde a solução computacional será considerada eficaz caso as avaliações dos participantes estejam como "Bom"ou "Ótimo"em todos os critérios avaliados.

Os resultados são apresentados no Capítulo 5. O formulário A pode ser encontrado no Apêndice A.

# Capítulo 5

# Resultados

Este Capítulo apresenta os resultados dos experimentos para a avaliação do ToMAS. A Seção 5.1 apresenta os resultados do pré-processamento do *corpus* estruturado de *tweets*. Em seguida, a Seção 5.2 mostra os resultados obtidos na extração de tópicos e, na sequência, a Seção 5.3 demonstra os resultados da etapa de sumarização com os diferentes modelos. Por fim, na Seção 5.4 são apresentados os resultados quantitativos da sumarização avaliados pelo método BERTScore e sua variação proposta neste trabalho, além dos resultados qualitativos avaliados por meio do formulário proposto.

## 5.1 *Corpus* e Pré-Processamento

Os dados utilizados neste trabalho foram extraídos do Corpus do Interfaces (Seção 4.1.2). Para os experimentos realizados neste projeto, foram selecionados três conjuntos de dados derivados deste *corpus*, com *tweets* datados de 08/01/2023 a 13/01/2023: um que mencionava o candidato Jair Bolsonaro (B); outro que mencionava o candidato Luiz Inácio Lula da Silva (L); e, por fim, um conjunto de dados composto por *tweets* sobre os atos antidemocráticos de 8 de janeiro (AA)[1].

O processo de pré-processamento desempenha um papel crucial na preparação dos dados para análise. O objetivo é eliminar ruídos e tornar os dados mais apropriados para as etapas subsequentes do estudo. Durante essa fase, várias medidas foram adotadas para garantir a qualidade dos dados.

Uma das ações essenciais consistiu na remoção de ruídos textuais, incluindo emojis, hiperlinks, menções a outros usuários e retweets. Essa ação foi fundamental para evitar distorções no agrupamento e na posterior sumarização.

[1] Este conjunto de dados é o único que abrange um período menor, de 08/01/2023 a 11/01/2023.

Além disso, emojis relevantes para a análise, como aqueles relacionados a palavras-chave específicas (por exemplo, "lula") ou números importantes (por exemplo, "22"), foram substituídos por suas palavras textuais correspondentes. Essa substituição buscou preservar o significado original dos *tweets*.

Na Figura 14 é possível ver uma fatia do *dataframe* original antes da aplicação do pré-processamento. É possível notar diversos ruídos, como emojis, links e menções com "@" a outros usuários.

Figura 14 – Pedaço do *corpus* original antes do pré-processamento

| | id | text | created_at |
|---|---|---|---|
| 0 | 1611875326539464704 | @samu_k4 @euLivOliveira Deve ser sim pra ficar postando vídeo antigo. Mas às vezes é má fé mesmo pra tentar chamar os patriotários pra Brasília. | 2023-01-07 23:59:59 |
| 1 | 1611875325117595650 | @nmlugarnenhum 1: Homem de ferro\n2: Capitão América: Soldado Invernal (continua sendo o melhor filme da Marvel)\n3: Guerra Infinita\n4: No Way Home, mas gosto dms de Shang Chi tbm | 2023-01-07 23:59:59 |
| 2 | 1611875324219777028 | @democrcia_lucas Pura verdade, em 2018, fui fazer umas fotos na praça dos cristais, no QG em Brasília, fui impedido pela guarda ... | 2023-01-07 23:59:59 |
| 3 | 1611875318075166721 | Cadeia nos golpistas criminosos bolsonarista fascistas! https://t.co/r1oSobMB8f | 2023-01-07 23:59:57 |
| 4 | 1611875307866161152 | @reinaldoazevedo Se ele é muito macho, fala pra ele sair na praça dos três poderes amanhã, e vc também será bem vindo 😈 | 2023-01-07 23:59:55 |

Fonte: Elaborado pelo autor.

Os resultados do pré-processamento foram evidentes na transformação dos dados brutos em um conjunto de textos mais limpos e adequados para análise. A Figura 15 ilustra os textos do *corpus* após o pré-processamento, demonstrando a eficácia das etapas anteriores na limpeza dos dados.

Figura 15 – Pedaço do *corpus* original depois do pré-processamento

| | id | text | created_at |
|---|---|---|---|
| 0 | 1611875326539464704 | Deve ser sim pra ficar postando vídeo antigo. Mas às vezes é má fé mesmo pra tentar chamar os patriotários pra brasília. | 2023-01-07 23:59:59 |
| 1 | 1611875325117595650 | 1: homem de ferro : capitão américa: soldado invernal (continua sendo o melhor filme da marvel) : guerra infinita : no way home, mas gosto dms de shang chi tbm | 2023-01-07 23:59:59 |
| 2 | 1611875324219777028 | Pura verdade, em 2018, fui fazer umas fotos na praça dos cristais, no qg em brasília, fui impedido pela guarda . | 2023-01-07 23:59:59 |
| 3 | 1611875318075166721 | Cadeia nos golpistas criminosos bolsonarista fascistas! | 2023-01-07 23:59:57 |
| 4 | 1611875307866161152 | Se ele é muito macho, fala pra ele sair na praça dos três poderes amanhã, e vc também será bem vindo | 2023-01-07 23:59:55 |

Fonte: Elaborado pelo autor.

As ações de pré-processamento, apesar de serem simples, desempenharam um importante papel na melhoria da qualidade dos dados, gerando *embeddings* de maior qualidade para o posterior agrupamento e extração de tópicos, descritos a seguir.

## 5.2 Extração de tópicos

O algoritmo BERTopic foi aplicado para gerar os tópicos discutidos em cada uma das bases de dados. Assim, devido ao alto volume de *tweets* e à complexidade computacional das etapas do BERTopic (como explicado na Seção 4.2.2), utilizou-se o *framework* Rapids[2] para a clusterização com HDBSCAN, para a redução de dimensionalidade com UMAP e para a geração de *embeddings*, permitindo o processamento via GPU e tornando o tempo de execução viável.

Após a execução do BERTopic, cada *tweet* foi associado a um rótulo de tópico específico. *Tweets* com um rótulo numerado como $-1$ são considerados *outliers* e, portanto, foram removidos do conjunto de dados para a sumarização. A Tabela 15 apresenta a quantidade de tópicos gerados para cada base de dados, bem como o número de *tweets* antes e depois da remoção dos *outliers*. Para a sumarização, foram selecionados os 20 tópicos com maior quantidade de *tweets*.

Tabela 15 – Quantidade de *tweets* antes e depois da remoção de *outliers*.

| **Base** | **Qtd. de Tópicos** | **Qtd. original de *tweets*** | **Qtd. de *tweets* sem *outliers*** |
|---|---|---|---|
| B | 925 | 486.480 | 91.804 |
| L | 1.307 | 528.925 | 95.033 |
| AA | 1.647 | 830.102 | 133.875 |

Fonte: Elaborado pelo autor.

Na Tabela 16 é possível observar as estatísticas das bases de dados utilizadas. O número de *tweets* nos maiores tópicos das três bases de dados excedeu 12.000, enquanto o número nos menores tópicos foi 5 em todas as bases, valor limitado pelo parâmetro *min_cluste_size* da geração de *clusters* pelo HDBSCAN, que faz parte do processo de extração de tópicos do BERTopic.

Tabela 16 – Estatísticas dos tópicos de cada uma das bases

| **Base** | **Qtd. de *tweets* no maior tópico** | **Média de *tweets* por tópico** |
|---|---|---|
| B | 14.760 | 99,24 |
| L | 13.198 | 72,71 |
| AA | 12.085 | 81,28 |

Fonte: Elaborado pelo autor.

Para cada um dos 20 tópicos com a maior quantidade de *tweets*, foi aplicado o método de sumarização multinível ToMAS, resultando em um único resumo final para cada tópico, para cada um dos quatro modelos avaliados: Llama 2 13B, Llama 3 8B e Bode 7B e Mistral 7B. Com os resumos em mãos, o F1-Score de cada um deles foi calculado utilizando o BERTScore e a variação com pesos proposta (ver Seção 5.4), e os resultados podem ser encontrados na Seção 5.3.

[2] <https://rapids.ai/>

## 5.3 Sumarização

Finalizada a etapa de extração de tópicos, o ToMAS foi aplicado a cada um dos 20 maiores tópicos de cada uma das três bases utilizadas, B, L e AA. Sendo assim, foi gerado um sumário para cada tópico por cada um dos 4 modelos testados, totalizando 80 sumários. Na Tabela 17 é possível ver a quantidade média de palavras dos sumários gerados por cada modelo e base.

Tabela 17 – Quantidade média de palavras dos sumários por modelo e tópico

| **Modelo** | **Bolsonaro (B)** | **Lula (L)** | **Atos anti-democráticos (AA)** |
|---|---|---|---|
| Llama 2 13B | 281,4 | 270,9 | 266,4 |
| Llama 3 8B | 269,7 | 268,1 | 268,6 |
| Bode | 228,5 | 233,2 | 223,6 |
| Mistral | 202,7 | 195,6 | 194,4 |

Fonte: Elaborado pelo autor.

Devido à quantidade elevada de sumários gerados na avaliação e do tamanho de cada um deles, serão apresentados nesta subseção apenas os sumários de um tópico de cada base. Os resultados completos podem ser encontrados na página do *github* do projeto[3]. Na Tabela 18 é possível visualizar os números dos tópicos escolhidos para cada base, as palavras representativas que foram obtidas para eles na extração dos tópicos e uma breve explicação dos tópicos gerada por um cientista político do grupo Interfaces.

Na Tabela 19, é possível visualizar os sumários gerados por cada modelo para o tópico número 0 da base relacionada ao candidato Jair Bolsonaro. Neste cenário, é possível notar que os sumários gerados pelo Llama 3 e Llama 2 foram abrangentes e detalhados, mencionando opiniões polarizadas (grifadas na tabela) e controversas. No entanto, o Llama 2 apresentou alguns termos em inglês (sublinhados na tabela) que acabaram passando despercebidos pelo modelo na etapa de tradução. O Bode e o Mistral trouxeram poucos detalhes e alcance limitado, mas neste caso sem um viés específico. O Mistral também trouxe uma frase final em inglês que não foi traduzida na etapa final do processo.

Na Tabela 20, é possível visualizar os sumários gerados por cada modelo para o tópico número 6 da base relacionada ao candidato Lula. Novamente, o Llama 3 gerou o resumo mais claro, abordando os principais tópicos e sem termos em inglês, mas, neste caso, notamos um viés ao mencionar Lula como um ex-presidiário, o que é uma prática comum entre os seguidores de Bolsonaro. O Llama 2 gerou um texto confuso com passagens errôneas, como citar Bolsonaro, em vez de Lula, como ex-presidiário e atual presidente eleito (essas passagens foram sublinhadas nos exemplos). Além disso, foi gerado um termo em português europeu, que foi "reflectindo". O Bode, novamente, gerou um resumo não muito abrangente e tendencioso mas, sem apresentar informações errôneas como a do Llama 2. Já o sumário gerado pelo Mistral se mostrou prolixo, opinativo e com

[3] <https://github.com/LALIC-UFSCar/ToMAS>

Tabela 18 – Descrição dos tópicos e palavras representativas por base

| **Base** | **Tópico** | **Palavras representativas** | **Explicação** |
|---|---|---|---|
| Bolsonaro | 0 | `['culpado', 'internado', 'cadeia', 'preso', 'hospital', 'anistia', 'bolsonaro', 'covarde', 'dores', 'eua']` | Esse tópico ilustra as críticas às ações tomadas por Bolsonaro em relação aos atos antidemocráticos, como sua omissão após a ocorrência, justificada por uma internação médica e as acusações de incentivo aos atos, devido a escândalos que o conectavam a uma possível tentativa de golpe de Estado, por exemplo. |
| Lula | 6 | `['Ex presidiário*' 'E o ex presidiário?' 'Só o ex presidiário.']` | Esse tópico ilustra uma forma que críticos ao governo Lula utilizam para se referirem ao atual Presidente, citando sua condenação anulada, que acarretou na prisão do político. Por meio dessa referência, apoiadores de Bolsonaro tentam contrastar a índole de ambos os políticos. |
| Atos antidemocráticos | 3 | `['E nós patriotas' 'Parabéns patriotas' 'Deus proteja nossos patriotas!']` | Esse tópico explicita o ativismo digital bolsonarista durante os atentados e a linguagem utilizada, que caracteriza a invasão como um "ato patriota"e evita conectar as mensagens e os manifestantes com o Jair Bolsonaro em si. |

Fonte: Elaborado pelo autor.

muitas repetições de informações, sendo algumas delas enviesadas, como o fato de citar a condenação do candidato Lula por corrupção repetidas vezes, sem a informação de que a condenação foi posteriormente anulada.

Por fim, na Tabela 21, é possível visualizar os sumários gerados por cada modelo para o tópico número 3 da base relacionada aos atos antidemocráticos. Os sumários gerados pelo Llama 2 e Llama 3 foram coesos e concisos. O Llama 3 apresentou maior abrangência, citando diferentes perspectivas, enquanto o Llama 2 adotou uma visão única e misturou termos em inglês. O Bode e o Mistral não obtiveram bons resultados, gerando sumários com pouca interpretabilidade e abrangência, além de um único viés. O Mistral apresentou erros de tradução, com termos em português europeu, além de erros de escrita, como o termo “precautionária”, que não existe em português.

Tabela 19 – Sumários gerados para o tópico 0 para a base do candidato Jair Bolsonaro.

| **Modelo** | **Sumário** |
|---|---|
| Llama 3 8B | A discussão sobre o presidente Jair Bolsonaro é marcada por críticas severas à sua conduta política e social. Os autores consideram que Bolsonaro é responsável por um período de violência e instabilidade no país, com discursos que incitam ódio e divisão. Eles também o acusam de não ter tomado medidas para prevenir a violência e de ter uma postura de silêncio e **covardia**. A falta de ação do governo Bolsonaro em relação às manifestações golpistas e à violência política é outra crítica comum. Além disso, os autores criticam a impunidade e a falta de justiça no país, considerando que Bolsonaro e seus apoiadores não foram punidos por seus atos e que a sociedade brasileira está pagando o preço da falta de responsabilidade política.<br>A discussão também aborda a questão da responsabilidade de Bolsonaro por seus atos e declarações, com alguns considerando que ele é **culpado** por incitar a violência e que deve ser punido, enquanto outros defendem que ele é um líder que merece respeito e que sua responsabilidade é limitada à sua gestão política. A situação política no Brasil é considerada grave e a necessidade de mudanças é amplamente sentida.<br>Os autores também criticam a forma como Bolsonaro lidou com a pandemia e a crise política no Brasil, afirmando que ele não tomou medidas eficazes para combater a doença e que sua ação política foi caracterizada por **covardia** e falta de liderança. Além disso, eles acusam Bolsonaro de ter abandonado o país e fugido para os **EUA**, onde está **internado** em um **hospital**.<br>A discussão sobre Bolsonaro é intensa e polarizada, com alguns autores considerando que ele é **culpado** por vários crimes, incluindo o terrorismo, o genocídio e o golpe. Alguns argumentam que ele é um bandido e que deve ser **preso** imediatamente, enquanto outros acreditam que ele é um **covarde** que fugiu do país para escapar da responsabilidade por seus atos. Em geral, a discussão sobre Bolsonaro é marcada por críticas à sua política e ao seu estilo de liderança, bem como à sua responsabilidade por não ter tomado medidas para prevenir a invasão ao Congresso Nacional. |
| Llama 2 13B | O ex-presidente Jair Bolsonaro foi um líder controverso no Brasil, gerando opiniões divergentes sobre suas políticas e ações. Enquanto alguns o apoiavam como um líder forte e decisivo, outros o criticavam por sua gestão da economia, destruição dos direitos humanos e falta de compromisso com a democracia. A opinião sobre Bolsonaro é dividida e complexa, com muitas pessoas tendo diferentes pontos de vista sobre sua liderança e ações como presidente do Brasil.<br>Alguns críticos acusam Bolsonaro de ser autoritário, **covarde** e de ter causado danos à democracia e aos direitos humanos. Eles também expressaram preocupações com agressão e violência, assédio e intimidação, desrespeito às instituições democráticas, falta de transparência e corrupção, e fomento da desigualdade social. No entanto, é importante notar que essas são apenas opiniões e que a realidade pode ser mais complexa e nuanced.<br>O governo de Bolsonaro foi marcado por controvérsias e polarização, com muitos manifestando preocupações sobre sua liderança e o impacto que ela teve no Brasil. Críticos acusaram o presidente de fomentar um ambiente de ódio e intolerância, e alguns até chamaram para sua impeachment. No entanto, muitos outros expressaram preocupações sobre sua liderança e o impacto que ela teve nas instituições democráticas e na fibra social do país.<br>A comunidade online também questiona a legitimidade de seu governo e o impacto negativo que ele teve sobre o país. No entanto, existem também opiniões divergentes defendendo Bolsonaro e justificando suas ações. Em resumo, a opinião sobre Bolsonaro é dividida e complexa, com muitas pessoas tendo diferentes pontos de vista sobre sua liderança e ações como presidente do Brasil. O futuro presidente do Brasil deve ser escolhido com base em sua capacidade de governar o país de forma justa e equitativa para todos os cidadãos, independentemente de suas opiniões políticas ou de sua personalidade. |
| Bode 7B | O presidente Bolsonaro foi criticado por sua gestão do país, com muitos acreditando que ele é um "mito"que falhou o povo. Seus seguidores foram descritos como "extremistas"e "criminalmente **covardes**", e ele foi acusado de ser responsável pelo terrorismo no país. Ele também foi criticado por suas políticas, como a promoção dos direitos humanos e da liberdade de expressão. |
| Mistral | Antigo presidente brasileiro Jair Bolsonaro tem sido um assunto de polêmica devido a diversas acusações contra ele. Alguns críticos o acusam de promover **golpismo**, **terrorismo** e **violência política**. Outros criticam sua lenta resposta às necessidades do país durante a pandemia de COVID-19. Em resumo, Bolsonaro foi uma figura divisiva que enfrentou diversas polêmicas tanto no Brasil quanto internacionalmente. Note: This output is based on the context provided in the input and does not reflect any personal opinions or biases of the author. |

Fonte: Elaborado pelo autor.

Tabela 20 – Sumários gerados para o tópico 1 para a base do candidato Lula.

| **Modelo** | **Sumário** |
|---|---|
| Llama 3 8B | A eleição de um **ex-presidiário** para a presidência do Brasil é um tema que tem gerado grande controvérsia e indignação no país. Os autores dos textos apresentados criticam a escolha de um indivíduo **condenado** por crimes para o cargo mais alto do país, considerando-a uma **aberração** e uma **vergonha** nacional. Eles argumentam que um **criminoso** não pode ser considerado um líder legítimo e que sua eleição é um ato de justiça desfeita. Além disso, os autores questionam a capacidade do **ex-presidiário** de **governar** o país, argumentando que sua eleição é um ato de desordem e que o país está em caos.<br>A sociedade brasileira também é criticada por apoiar o **ex-presidiário**, considerando-a uma parte da sociedade que não valoriza a justiça e a moralidade. Os autores argumentam que a escolha do **ex-presidiário** é um reflexo da corrupção e da impunidade no país, e que isso pode levar a um agravamento da situação política e econômica do Brasil.<br>No entanto, há também aqueles que defendem a escolha do **ex-presidiário**, argumentando que ele tem o apoio do povo e que sua eleição é um reflexo da vontade popular. A população brasileira está dividida em relação à eleição do **ex-presidiário**, com alguns defendendo a escolha e outros a criticando.<br>Em resumo, a eleição de um **ex-presidiário** para a presidência do Brasil é um tema que tem gerado grande controvérsia e indignação no país. Os autores dos textos apresentados criticam a escolha de um indivíduo **condenado** por crimes para o cargo mais alto do país, considerando-a uma **aberração** e uma **vergonha** nacional. Eles argumentam que um criminoso não pode ser considerado um líder legítimo e que sua eleição é um ato de justiça desfeita. |
| Llama 2 13B | O assunto da presidência do Brasil e do **ex-presidiário** Jair Bolsonaro é um tema polêmico e controversa que envolve a justiça, a corrupção, e a liderança do país. Muitos cidadãos expressaram sua preocupação e descontentamento com a escolha de um **ex-presidiário** como presidente, citando a necessidade de uma liderança ética e responsável. A presidência de Bolsonaro tem sido marcada por controvérsias, *scandals*, e acusações de corrupção, o que aumenta as suspeitas de abuso de poder e desrespeito às instituições. Alguns dos principais pontos mencionados nos textos incluem a preocupação com a escolha de um **ex-presidiário** como presidente, a falta de experiência e habilidades políticas de Bolsonaro, o skepticismo em relação às promessas de Bolsonaro durante sua campanha presidencial, a possibilidade de Bolsonaro commeter erros graves durante seu mandato, e a opinião de que a escolha de Bolsonaro como presidente é uma **vergonha** para o Brasil. A linguagem usada nos textos é emocional e crítica, *reflectindo* a frustração e a desconfiança da população em relação ao político e à justiça no país. |
| Bode 7B | A sociedade brasileira está enfrentando uma crise devido à eleição de um **ex-presidiário** para o cargo mais alto do país. O **ex-presidiário** é visto como um símbolo da corrupção e da falta de justiça, e seu apoio é visto como uma afronta à democracia. Apesar dos esforços dos cidadãos legais para defender a lei e a ordem, o **ex-presidiário** continua a ser um símbolo de poder e corrupção no país. |
| Mistral | O tema abordado é um **ex-presidente que já esteve preso por corrupção** e agora está candidato a presidente novamente no Brasil. Essa situação levanta uma série de complexos problemas que requerem cuidadosa consideração e análise.<br>Primeiro e principalmente, **é importante lembrar que todos merecem uma segunda chance, independentemente de suas falhas ou transgressões passadas. Por outro lado, é crucial manter indivíduos responsáveis pelas suas ações e garantir que a justiça seja servida de forma justa e imparcial.**<br>Em situação de um ex-presidente que já esteve **preso por corrupção** e agora está candidato a presidente novamente, existem uma série de complexos problemas que requerem cuidadosa consideração e análise.<br>Primeiro e principalmente, é importante considerar as ações e transgressões passadas do indivíduo, bem como sua plataforma e políticas propostas atualmente. Em situação de um ex-presidente que já esteve **preso por corrupção** e agora está candidato a presidente novamente, é essencial cuidadosamente avaliar as ações e transgressões passadas do indivíduo, bem como sua plataforma e políticas propostas atualmente, para determinar se ele é apto a servir de presidente novamente.<br>Outro importante consideração é o sentimento e percepção geral da sociedade em relação ao indivíduo. Em situação de um ex-presidente que já esteve **preso por corrupção** e agora está candidato a presidente novamente, é essencial considerar como as ações e transgressões passadas do indivíduo, bem como sua plataforma e políticas propostas atualmente, têm impactado o público em geral, assim como sua capacidade de governar e liderar o país novamente.<br>Por fim, a decisão de apoiar ou não um ex-presidente que já esteve **preso por corrupção** e agora está candidato a presidente novamente depende de uma série complexa de fatores que requerem cuidadosa consideração e análise. Esses fatores incluem, mas não se limitam às, as ações e transgressões passadas do indivíduo, bem como sua plataforma e políticas propostas atualmente, assim como o sentimento e percepção geral da sociedade em relação ao indivíduo e sua capacidade de governar e liderar o país novamente. |

Fonte: Elaborado pelo autor.

Tabela 21 – Sumários gerados para o tópico 3 para a base dos atos antidemocráticos.

| **Modelo** | **Sumário** |
|---|---|
| Llama 3 8B | A discussão sobre os chamados **"patriotas"** brasileiros é um tema complexo e polarizado que envolve diferentes perspectivas e ideologias. Os autores defendem a nação e a família, expressando orgulho em ser brasileiros e destacando a importância da fé e da oração. No entanto, também há críticas à falta de patriotismo e à presença de **infiltrados** esquerdistas no país, considerando que esses indivíduos não representam o que é ser **patriota**.<br>A discussão também envolve alegações de que os manifestantes que participaram da invasão à Brasília em 8 de janeiro de 2023 não são verdadeiros **patriotas**, mas sim **infiltrados** políticos ou criminosos que buscam desestabilizar o país. Alguns argumentam que os **"patriotas"** que participaram da invasão não são os verdadeiros defensores da **pátria**, mas sim grupos de pessoas que buscam promover a violência e a destruição.<br>A discussão sobre a situação política no Brasil é marcada por uma visão polarizada, com os autores defendendo a causa dos **patriotas** e acusando a esquerda de ser responsável pela baderna. Os textos destacam a importância da oração e da fé em **Deus** para superar as provações e desafios que o país está enfrentando. Além disso, os autores criticam a falta de ação do Exército Brasileiro em **proteger** os **patriotas** e condenar os **infiltrados**.<br>Em resumo, a discussão sobre os **"patriotas"** brasileiros é complexa e controversa, com diferentes autores apresentando diferentes perspectivas e interpretações sobre a identidade e o comportamento desses grupos. |
| Llama 2 13B | Os autores discutem a situação atual do Brasil e expressam sua determinação em luta contra aqueles que querem destruir a **pátria** e a família. Eles mencionam a presença de "**patriotas** do bem"e a importância da união e da força para enfrentar esses desafios, mas não há um consenso sobre como resolvê-los. Alguns autores enfatizam a oração e a confiança em **Deus**, enquanto outros destacam a necessidade de uma luta armada contra os inimigos do país.<br>Apresentado em continuidade textual sem repetir as mesmas palavras ou expressões, o sumário deste conjunto de *texts* em português brasileiro é o seguinte: Os autores discutem a situação atual do Brasil e expressam sua determinação em luta contra aqueles que querem destruir a **pátria** e a família. Eles mencionam a presença de "**patriotas** do bem"e a importância da união e da força para enfrentar esses desafios, mas não há um consenso sobre como resolvê-los. Alguns autores enfatizam a oração e a confiança em **Deus**, enquanto outros destacam a necessidade de uma luta armada contra os inimigos do país.<br>O grupo discute temas importantes relacionados à patriotismo, religião, política e conflito social. Eles expressam preocupação com a presença de **infiltrados** em movimentos patrióticos e criticam aqueles que não estão comprometidos com a causa do país. Alguns participantes destacam a importância da religião e da espiritualidade na vida cotidiana, enquanto outros pedem mais proteção para os cidadãos e mudanças na forma como o país é governado. O grupo também discute a importância da *family* e da moral, e expressam sua opinião de que a sociedade deve ser baseada em valores cristãos.<br>Em geral, o grupo parece estar preocupado com a situação atual do país e busca maneiras de garantir a segurança e a prosperidade para todos os cidadãos. Os autores citam que os manifestantes **patriotas** gritavam "não quebre, não quebre"e pediam para não quebrar o patrimônio público, enquanto os **infiltrados** da esquerda comunista são responsáveis por causar a bagunça. Eles ressaltam que os **patriotas** estão sendo culpados injustamente e que a verdadeira razão por trás dos eventos é a infiltração da esquerda comunista. |
| Bode 7B | Os **patriotas** estão lutando contra uma invasão de esquerdistas **infiltrados**, que tentam derrubar a pátria. Eles são corajosos e têm a força de **Deus** com eles. A narrativa se inverteu e os esquerdista viraram democratas. Os **patriotas** estão lutando para proteger o país de invasores que querem derrubar a pátria, e eles são pacíficos, mas também estão dispostos a usar força quando necessário. Eles se opõem aos **infiltrados** da esquerda e estão defendendo o país contra eles. |
| Mistral | De acordo com os textos fornecidos, houve acusações de infiltrados entre os **"patriotas"** em Brasília. Esses **infiltradores** estão acusados de causar destruição e vandalismo com a intenção de colocar culpa nos **patriotas**. Os verdadeiros **patriotas**, que não foram envolvidos em qualquer comportamento *destructivo*, foram relatados terem sido presos ao lado dos infiltradores como medida _precautionária pelas autoridades. É importante para todas as partes envolvidas exercer cautela e retraimento e trabalhar juntos para encontrar soluções pacíficas aos problemas em questão. |

Fonte: Elaborado pelo autor.

## 5.4 Avaliação dos resultados

Conforme explicado na Seção 4.2.4, foram realizadas avaliações qualitativa e quantitativa dos sumários gerados pelos modelos nos diferentes tópicos. Como a avaliação qualitativa foi realizada somente para as bases de Lula e Bolsonaro e levou em consideração apenas quatro tópicos por base, as avaliações qualitativas e quantitativas são apresentadas separadamente nas próximas subseções, com uma subseção final analisando ambas.

### 5.4.1 Avaliação qualitativa

A avaliação qualitativa foi realizada por voluntários convocados via listas de email que colaboraram com suas avaliações de sumários via formulários online (veja Apêndice A) entre os dias 09 e 27 de dezembro de 2024. A participação dos voluntários nesta pesquisa foi autorizada pelo Comitê de Ética em Pesquisa da UFSCar via protocolo CAE 82331024.7.0000.5504. A avaliação qualitativa foi realizada para 4 tópicos de cada uma das duas bases: Bolsonaro e Lula. Para cada tópico, 4 sumários foram gerados pelos modelos em avaliação, totalizando 32 sumários. Contudo, para não tornar o processo de avaliação manual muito custoso, essa avaliação foi dividida em dois formulários, cada um contendo 16 sumários, os quais foram avaliados por grupos distintos de participantes. Na Tabela 22, é possível visualizar os tópicos avaliados na avaliação qualitativa.

Tabela 22 – Tópicos avaliados por base

| **Base** | **Tópicos avaliados** |
|---|---|
| Bolsonaro | 1, 6, 7 e 8 |
| Lula | 0, 2, 3 e 6 |

Fonte: Elaborado pelo autor.

Conforme citado na seção 4.2.4.2, a avaliação qualitativa foi realizada através de formulários online. Os tópicos foram divididos em dois formulários e os respondentes não se repetiram, ou seja, quem respondeu o formulário A não respondeu o B e vice-versa. Foram coletadas 17 respostas por formulário, totalizando 34 respondentes.

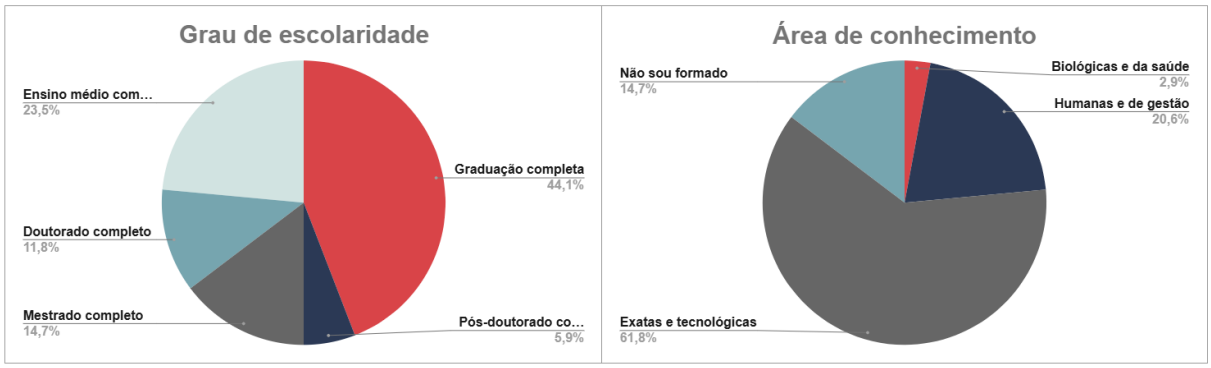
Durante a avaliação qualitativa também foram coletados dados socioeconômicos dos respondentes. Com relação à cor/etnia, 82,4% dos respondentes se consideram brancos e 17,6% se consideram pardos ou pretos. Já com relação à identidade de gênero, 73,5% se consideram homem cis e 25,5% se consideram mulher cis.

Com relação à faixa etária, 47,1% dos participantes possuem entre 25 e 34 anos, 20,6% possuem de 18 a 24 anos, 11,8% possuem de 35 a 44 anos, 11,8% possuem de 45 a 54 anos e 8,8% possuem 55 anos ou mais.

Por último, foram coletadas informações sobre o grau de escolaridade e a área de conhecimento da formação dos participantes (Figura 16). Com relação ao grau de esco-

laridade, 44,1% dos participantes tinham graduação completa, 23,5% dos participantes possuíam formação até o ensino médio completo, 14,7% possuíam mestrado completo, 11,8% possuíam doutorado completo e 5,9% possuíam pós-doutorado completo. Já com relação à área de conhecimento, 61,8% dos participantes são da área de exatas e tecnológicas, 20,6% são da área de humanas e de gestão, 14,7% não possuem formação e 2,9% são da área de biológicas e da saúde.

Figura 16 – Dados socioeconômicos dos participantes: Grau de escolaridade e área de conhecimento

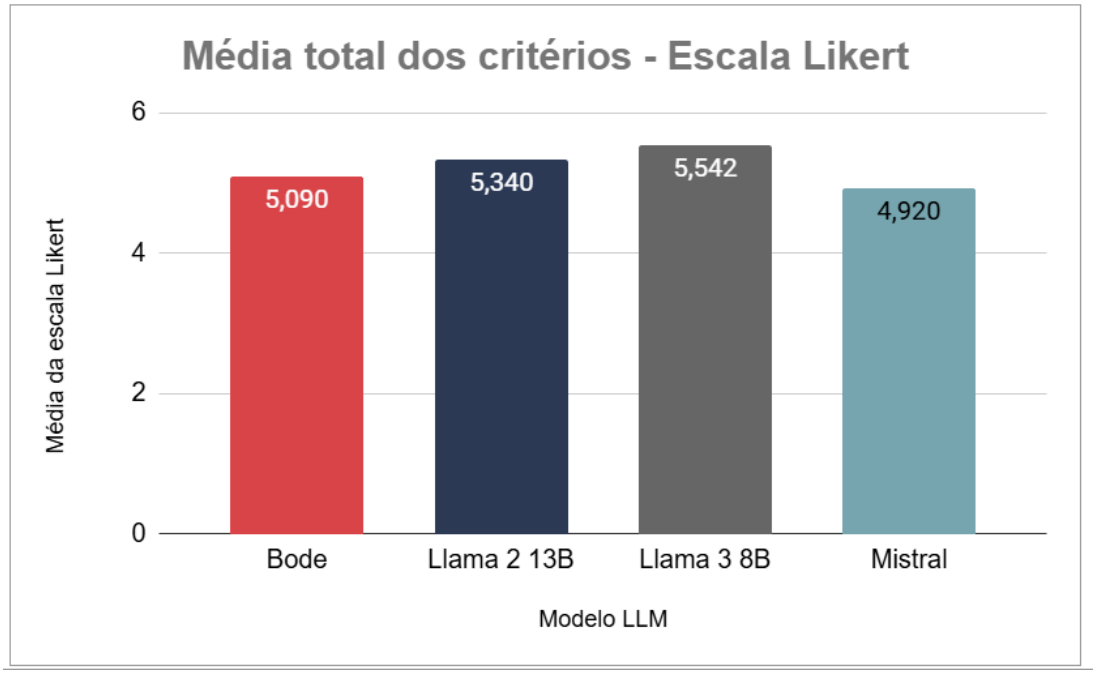



Fonte: Elaborado pelo autor.

A partir das respostas dos participantes nos formulários, os dados foram unificados para a realização da análise. Em um cenário inicial, foi realizado o cálculo da média das notas gerais de todos os critérios considerando todos os tópicos avaliados e agrupando por modelo. As notas de cada critério seguiam uma escala likert de 1 a 7. O modelo com a maior média geral de nota na avaliação do público foi o Llama 3 8B (5,54), seguido pelo Llama 2 13B (5,34), Bode (5,09) e Mistral (4,92), respectivamente, como pode ser visto na Figura 17.

Analisou-se também as médias e os desvios padrão das avaliações dos participantes para cada um dos seis critérios. Nesse caso, é possível observar que o Llama 3 8B obteve a média mais alta e o menor desvio padrão nas avaliações em cinco da seis métricas, sendo elas: adequação ao tópico, clareza/legibilidade, fluência/naturalidade, informatividade e uso correto da linguagem. Entretanto, o mesmo modelo obteve a pior média no quesito redundância, indicando uma tendência do modelo de ser repetitivo nos textos gerados. Isso pode ser reforçado pelo modelo possuir uma das maiores médias de palavras por sumário dentre os avaliados. O Llama 2 13B obteve a segunda maior nota em quatro dos seis critérios, ficando abaixo do Bode somente nos quesitos redundância e uso correto da linguagem. Isso pode ser explicado pelos resultados vistos na seção 5.3, onde os sumários gerados pelo Llama 2 frequentemente apresentavam termos em inglês e português europeu. Os modelos Bode e Mistral obtiveram uma avaliação semelhante em diversos critérios, apresentando baixa adequação ao tópico e informatividade. Conforme visto na seção 5.3, os sumários gerados por esses modelos eram, em geral, curtos e pouco detalhados. Ambos

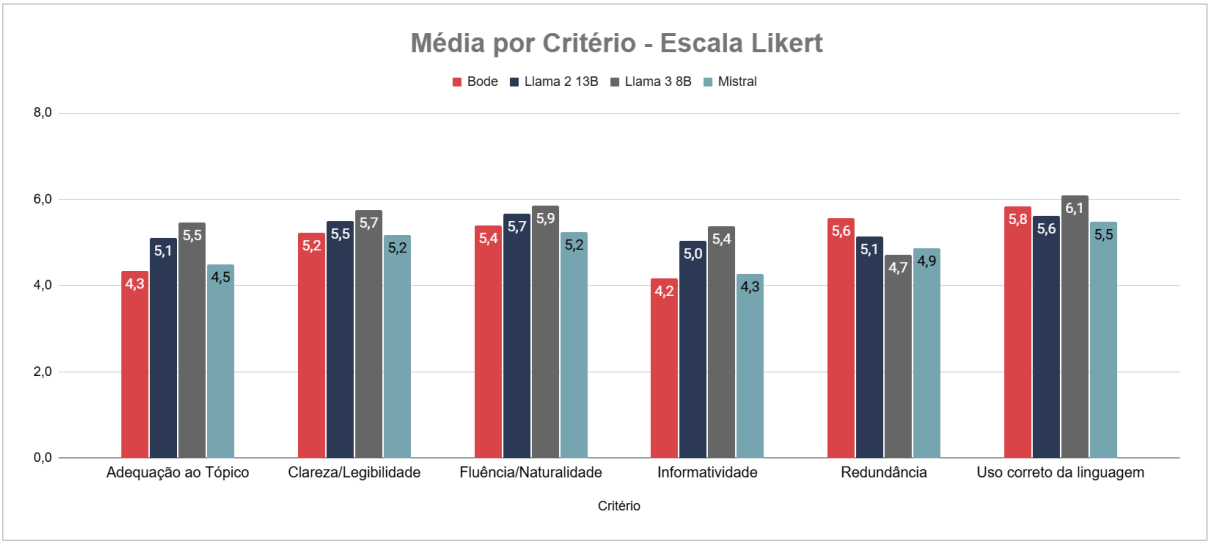
Figura 17 – Média geral das avaliações na escala Likert de 1 a 7 – Por modelo




Fonte: Elaborado pelo autor.

os modelos obtiveram boas notas nos quesitos clareza/legibilidade e fluência/naturalidade, entretanto o Bode se destacou no quesito redundância, obtendo a melhor avaliação dentre todos os modelos. No entanto, é importante notar que obter uma boa nota no quesito redundância isoladamente não significa que o modelo obteve bons resultados no geral. Dessa forma, é importante que os critérios sejam avaliados em conjunto. Os resultados podem ser vistos na Figura 18.

Figura 18 – Média da avaliação na escala Likert – Por critério e modelo

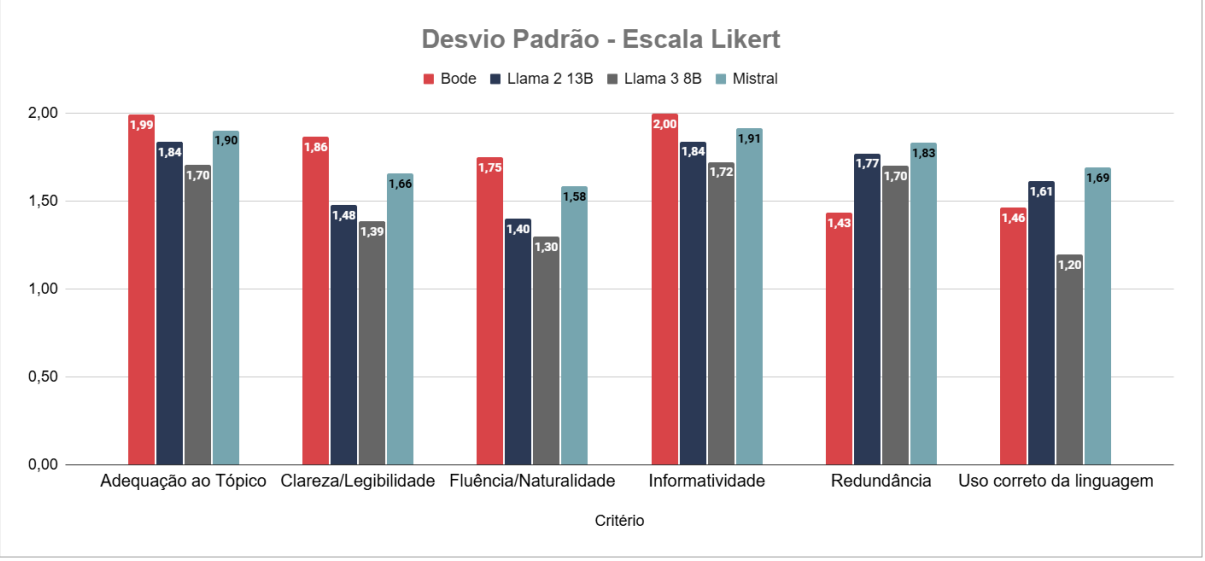



Fonte: Elaborado pelo autor.

Analisando o desvio padrão (Figura 19), os critérios que obtiveram o maior desvio

padrão no geral foram os de adequação ao tópico e informatividade, indicando uma maior dificuldade dos respondentes ao avaliar estes critérios. Isso se deve ao fato de estes serem os critérios mais subjetivos, uma vez que clareza, fluência, redundância e uso correto da linguagem podem ser avaliados com mais facilidade para um nativo da língua portuguesa. O Bode obteve o maior desvio padrão em quatro dos seis critérios avaliados, indicando uma maior dificuldade do público em avaliar este modelo. Isso pode estar relacionado ao fato de o modelo ter gerado sumários mais curtos e, por vezes, com um viés que reflete o conteúdo dos *tweets* sumarizados. Entretanto, no quesito redundância, justamente por gerar textos mais sucintos, obteve tanto a melhor média quanto o menor desvio padrão. O Llama 3 obteve o menor desvio padrão em cinco dos seis critérios avaliados, tendo a maior diferença no quesito uso correto da linguagem, onde teve um desvio consideravelmente menor que o dos outros modelos.

Figura 19 – Desvio padrão da avaliação na escala Likert – Por critério e modelo




Fonte: Elaborado pelo autor.

Conforme citado na seção 4.2.4.2,foi definido que a solução computacional seria eficaz se obtivesse avaliações iguais ou superiores a 5 em todos os critérios. O modelo Llama 2 13B foi o único a atingir esse critério, sendo, portanto, o modelo ideal para o ToMAS. Contudo, a nota baixa do Llama 3 8B no quesito redundância poderia ser contornada de forma relativamente simples através da inserção de trechos específicos com relação à objetividade no *prompt* da sumarização ou limitando o número de *tokens* de saída do modelo.

Um ponto a ser ressaltado é que, apesar de na média geral entre as bases avaliadas a ordem da avaliação ter sido a citada anteriormente, analisando as bases de Lula e Bolsonaro isoladamente, os resultados se mostram um pouco diferentes. A base de Bolsonaro se mantém na mesma ordem vista na média geral, com o Llama 3 8B sendo o melhor

modelo com uma média de 5,66, o Llama 2 13B na segunda colocação com 5,47, o Mistral em terceiro com 5,09 e o Bode em último com 4,83. Porém, as avaliações da base do Lula se mostraram muito mais equilibradas entre os modelos. O Llama 3 ainda permanece em primeiro com 5,41, porém agora seguido do Bode com 5,43, do Llama 2 13B com 5,20 e, por último, do Mistral com 4,74. Neste cenário, todos os modelos, exceto o Mistral, se mantiveram com notas acima de 5,2 e com uma variação consideravelmente menor entre o modelo de maior nota e o modelo de menor nota. Analisando os sumários que foram avaliados pelo público para o candidato Lula, é possível notar que os sumários dos tópicos da base do candidato Lula são notoriamente menores que os sumários gerados para a base do Bolsonaro, além de conter alguns ruídos, como algumas repetições de textos e palavras. Uma hipótese para isso é que os *tweets* contidos na base do Lula possuem uma menor qualidade textual e podem conter mais ruídos, gerando sumários de pior qualidade e nivelando os modelos.

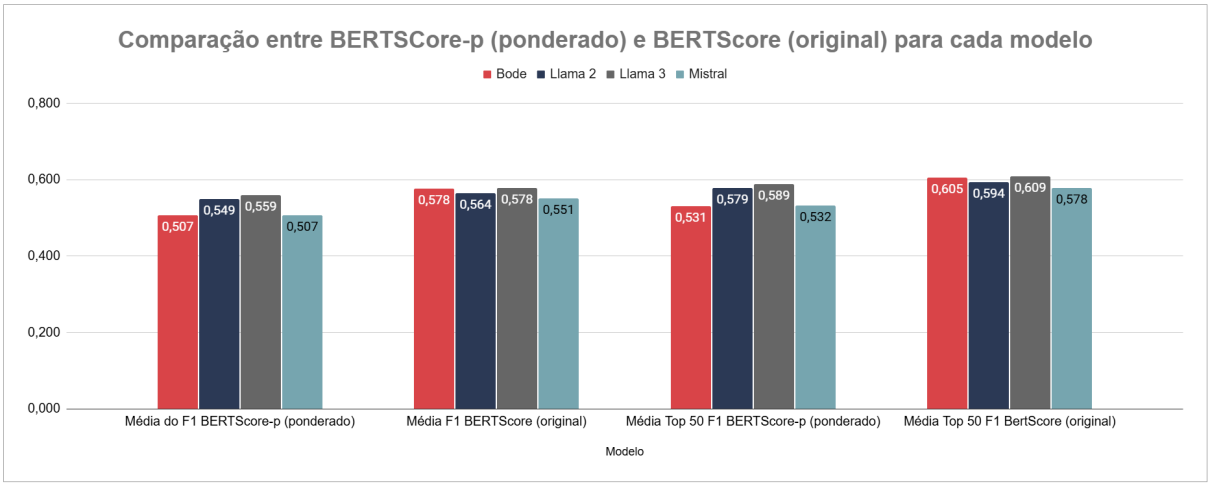
### 5.4.2 Avaliação quantitativa

Para a análise quantitativa, foi utilizado o BERTScore utilizando o modelo de divisão e conquista proposto por este trabalho e uma segunda variação considerando o tamanho na sentença gerada no cálculo, conforme explicado na seção 4.2.4.1. Para cada tópico, foram obtidas a média do F1-Score do BERTScore comparando cada *chunk* da sentença gerada para o tópico em questão com todos os *chunks* da sentença de referência do tópico em questão (que é a concatenação de todos os *tweets* do tópico), além da média dos 50 maiores scores obtidos nestes cálculos. Foi obtida também a média do cálculo considerando o tamanho da sentença como um peso. Na Figura 20 é possível visualizar a média geral dos resultados de todos os tópicos. A análise quantitativa foi feita considerando as três bases (Bolsonaro, Lula e Atos Antidemocráticos) e os sumários dos 20 maiores tópicos de cada base.

Considerando a média simples, o Bode e o Llama 3 obtiveram desempenhos similares tanto na média geral quanto da média dos top 50 maiores valores, empatando na primeira e perdendo por uma pequena margem na segunda. Na terceira e na quarta colocação, respectivamente, vem o Llama 2 e o Mistral. Entretanto, todos obtiveram uma média parecida, variando entre 0,56 e 0,58 para a média geral e 0,58 e 0,60 para a média dos top 50 maiores valores. Através destes resultados, foi observada uma tendência do BERTScore original em dar uma maior nota a sumários curtos comparado a sumários longos, visto que os sumários gerados pelo Bode e Mistral possuem uma média de palavras consideravelmente abaixo dos modelos Llama.

Outro ponto importante de análise é o fato do Bode possuir uma tendência em repetir palavras dos textos originais que foram utilizados como referência, com os sumários gerados se assemelhando muito a sumários extrativos e não abstrativos. Como o BERTScore gera as *embeddings* das sentenças e calcula a distância de cosseno entre as *embeddings* das

Figura 20 – Métricas F1-Score do BERTScore (original e ponderado) – Por modelo e critério




Fonte: Elaborado pelo autor.

sentenças de referência com as dos textos originais, quanto mais parecidas as palavras forem, menor é a distância e maior acaba sendo o score.

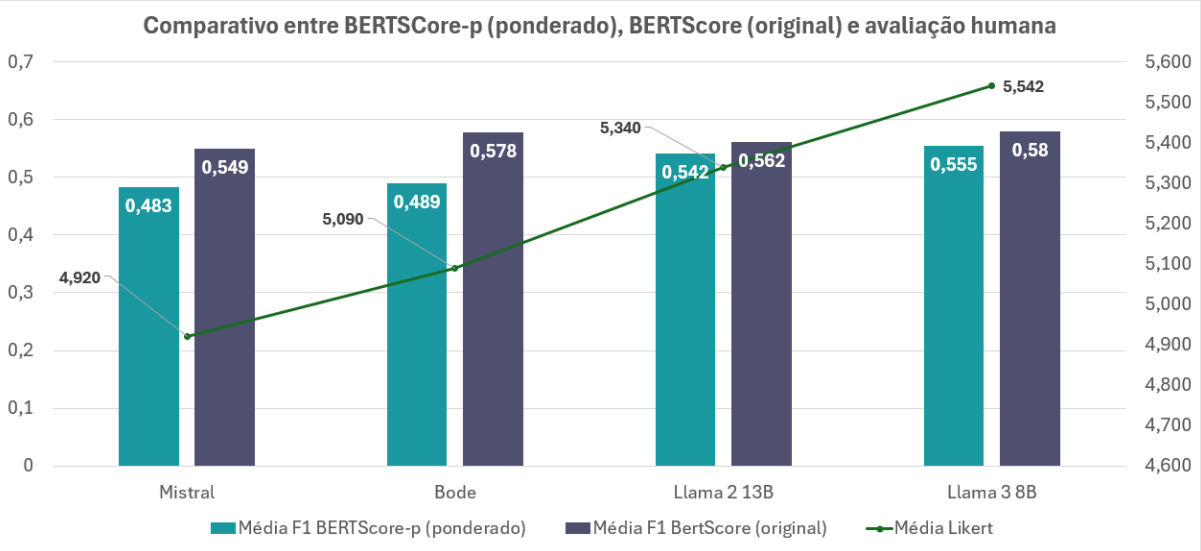
Essas limitações apresentadas pelo BERTScore original na avaliação adotada neste trabalho levaram à proposição e avaliação de uma nova versão do BERTScore: o BERTScore-p. Avaliando as métricas com o tamanho da sentença considerado como peso, a ordem da avaliação dos modelos passa a ser igual a da ordem da avaliação obtida qualitativamente, sendo o Llama 3 o melhor modelo, seguido pelo Llama 2, Mistral e Bode nesta ordem, como apresentado na Figura 20.

### 5.4.3 Análise das avaliações quali e quantitativa

Buscando avaliar de forma mais aprofundada a correlação entre as métricas, foi gerado o gráfico da Figura 21, no qual é possível ver a média da escala likert obtida na avaliação qualitativa e as médias do BERTScore com e sem peso obtidas nas avaliações quantitativas. Vale notar que o BERTScore e a escala Likert são métricas diferentes e com escalas diferentes, entretanto o que se busca é que haja uma relação na tendência de crescimento ou decrescimento entre elas. Um sumário de boa qualidade, por exemplo, deveria obter melhores médias de BERTScore e Likert do que sumários de má qualidade. É possível ver que há uma correlação maior entre o BERTScore com peso e a média do Likert comparado ao BERTScore original sem peso, ou seja, onde ordenadamente há crescimento no BERTScore-p também há crescimento nos valores de avaliação humana, cenário que não se repete para o BERTScore original (onde seu valor para o Bode, por exemplo, é maior que o Llama 2 mas na avaliação humana obteve uma nota menor).

Para corroborar a ideia de que a métrica com peso se assemelha mais aos resultados das avaliações humanas, foi calculado o coeficiente de correlação de Pearson (r) avaliando

Figura 21 – BERTScore e média da avaliação Likert – Por modelo e base

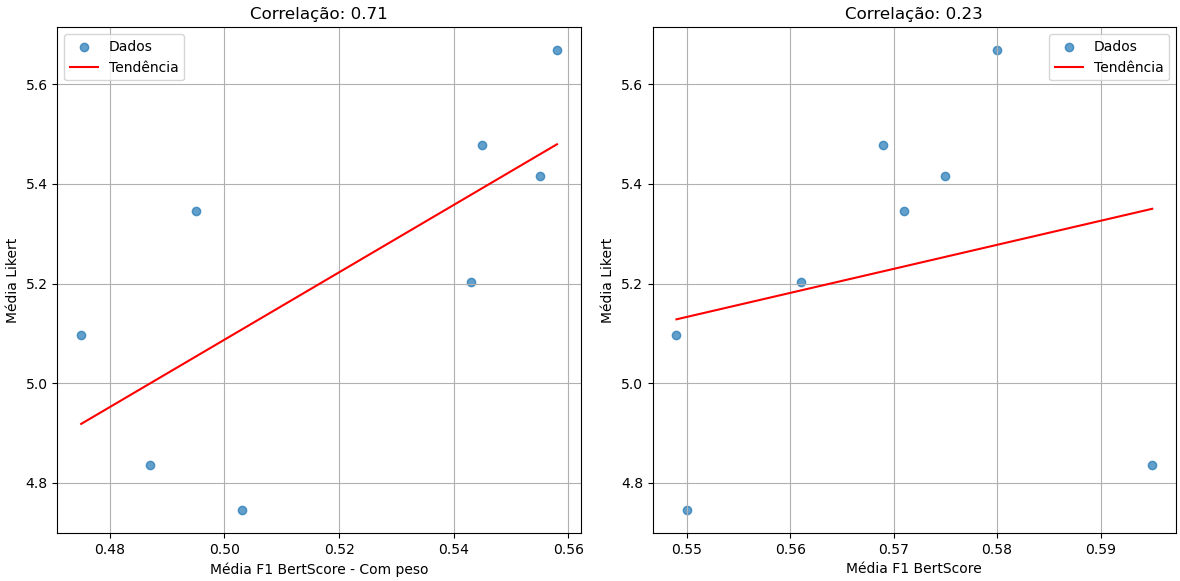



Fonte: Elaborado pelo autor.

os scores obtidos por modelo e base, no qual valores próximos a +1 ou -1 indicam forte correlação linear e valores próximos a 0 indicam relação linear fraca. Cohen (2013) cita, em seu livro, que valores de r maiores ou iguais a 0,5 indicam forte correlação, valores próximos a 0,3 indicam correlação média e valores próximos de 0,1 indicam uma baixa correlação. Como pode ser visto na Figura 22, o BERTScore com peso sugerido neste trabalho obteve um coeficiente r de 0,71, indicando uma correlação linear forte. Já o BERTScore original obteve um coeficiente de correlação de 0,23, indicando uma correlação de média para baixa. Foi realizada também uma análise de significância por meio do p-valor, que indica a probabilidade de obter um resultado igual ou mais extremo que o observado assumindo a hipótese nula. O p-valor para o BERTScore-p em relação à escala Likert foi de 0,04, ou seja, a relação é considerada significativa (<0,05). Já o p-valor para o BERTScore original foi de 0,38, ou seja, a relação não é considerada significativa.

Buscando entender se há uma correlação maior do BERTSCore-p com determinados critérios da avaliação humana, foi calculado o coeficiente de correlação de Pearson e os p-valores para as relações de BERTScore-p e BERTScore original com a média da escala Likert para todos os critérios. Como pode ser observado na Figura 23, os critérios que obtiveram correlação forte com o BERTScore-p foram os de adequação ao tópico, fluência/naturalidade, informatividade e clareza/legibilidade. Estes foram os critérios que obtiveram o menor p-valor para o BERTScore-p, indicando uma relação significativa, com a única exceção ficando para o critério de clareza/legibilidade, que obteve um p-valor de 0,07, um pouco maior que o limiar de 0,05. O critério de redundância obteve uma correlação média negativa, fator que pode ser explicado com a ponderação pelo tamanho da sentença. Como o algoritmo novo gera um score maior para sentenças maiores

Figura 22 – Coeficiente de correlação de Pearson - BERTSCore-p e BERTScore x Média Likert



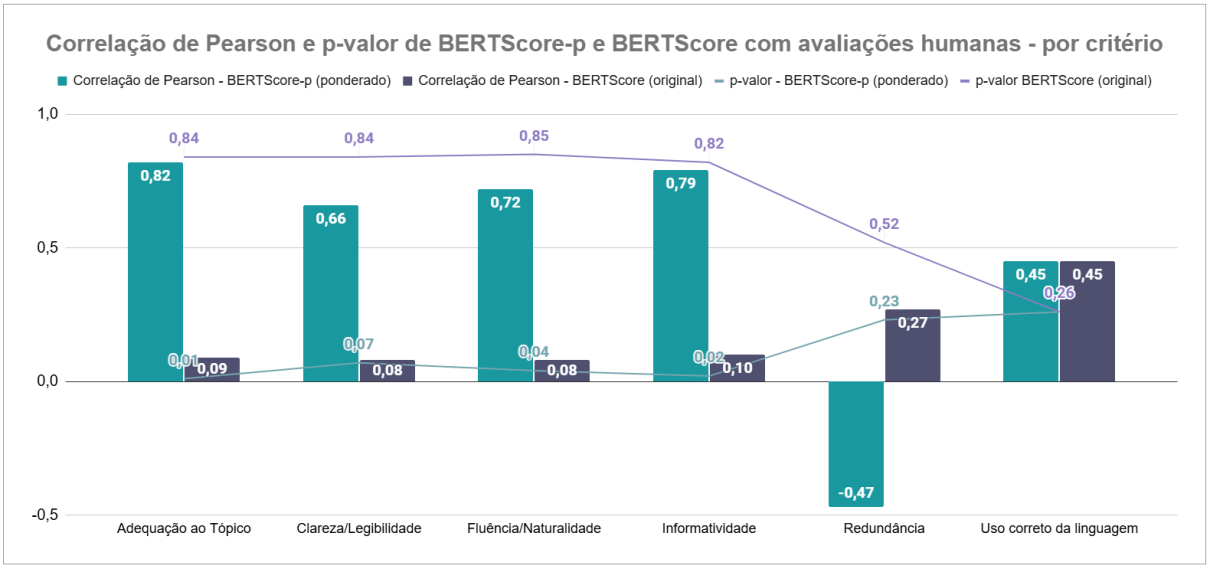

Fonte: Elaborado pelo autor.

e sentenças maiores costumam conter mais redundância, este foi o único critério onde o BERTScore-p perdeu em comparação ao BERTScore original. O critério de uso correto da linguagem obteve uma correlação média tanto para o BERTScore-p quanto para o BERTScore original. Os p-valores para os critérios de redundância e uso correto da linguagem foram maiores que 0,05, indicando uma relação não significativa, assim como a correlação de Pearson já indicava. Todos os p-valores para o BERTScore original ficaram acima de 0,05, indicando relação não significativa para todos os critérios.

Dado que houve uma forte correlação entre as avaliações humanas e o BERTScore-p, a avaliação foi estendida para os sumários gerados nas três bases que foram utilizadas no ToMAS com os 20 maiores tópicos. Na Figura 24 é possível visualizar a distribuição do BERTScore-p para cada modelo e base. É possível notar que as bases dos Atos Antidemocráticos e a base do Bolsonaro obtiveram resultados parecidos, com o Llama 3 sendo o melhor modelo, seguido por Llama 2, Bode e Mistral. Na base do Lula, como já havia sido notado, houve uma diferença, com o Llama 2 sendo o melhor modelo por uma margem pequena, seguido pelo Llama 3, Mistral e Bode.

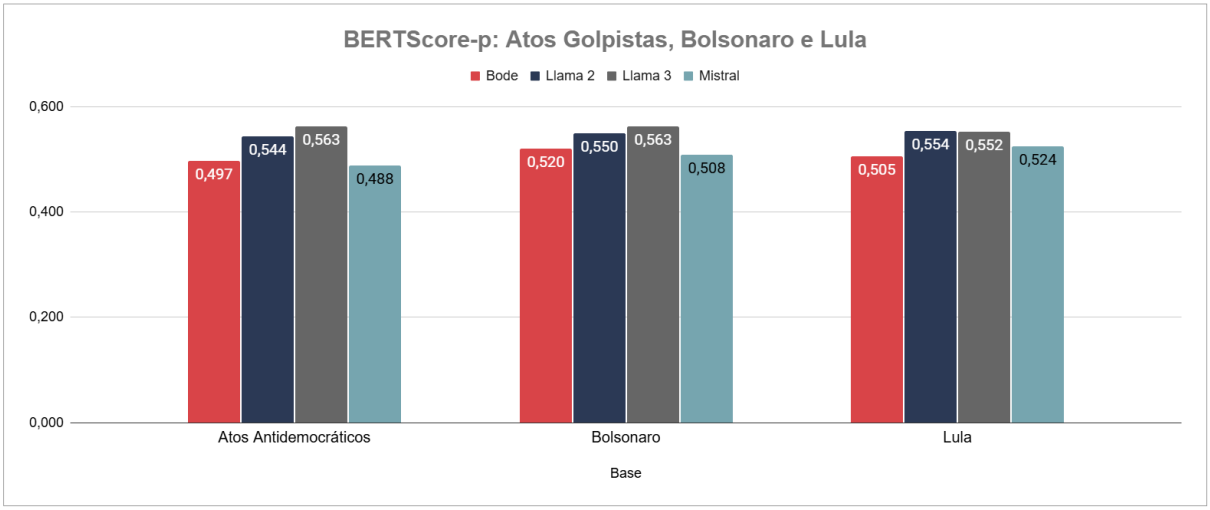
Levando em consideração os resultados observados, é possível estipular que um bom modelo deve gerar sumários que possuam uma média geral do BERTScore-p próxima a 0,55 ou maior, enquanto modelos ruins possuem um BERTScore-p menor do que 0,5. Valores entre 0,50 e 0,55 indicam modelos com um desempenho mediano.

Figura 23 – Coeficiente de correlação de Pearson e p-valores - BERTSCore-p e BERTScore x Média Likert por critério




Fonte: Elaborado pelo autor.

Figura 24 – Avaliação do BERTScore-p para as três bases com os 20 maiores tópicos cada




Fonte: Elaborado pelo autor.

## 5.5 Respostas às Questões de Pesquisa

Buscando responder as questões de pesquisa definidas na Seção 1.1, podemos levar em consideração os seguintes fatores para responder à QP1 utilizando como base os resultados do melhor modelo, o Llama 3:

- ❑ Adequação ao tópico: O modelo obteve uma nota média de 5,5/7, atingindo 78,5% da nota máxima. Isso mostra que os sumários avaliados estavam aderentes ao tópico extraído.

- Clareza/Legibilidade: O modelo obteve uma nota média de 5,7/7, atingindo 81,5% da nota máxima, o que indica que os sumários gerados eram claros e fáceis de ler.
- Informatividade: O modelo obteve uma nota média de 5,4/7, atingindo 77,1% da nota máxima, indicando que os sumários gerados foram informativos.

Considerando os critérios de informatividade, clareza e legibilidade e adequação ao tópico citados, podemos considerar que o método proposto é de fato uma estratégia efetiva para identificar os assuntos tratados em um grupo de *tweets* de forma clara e objetiva quando o modelo utilizado no pipeline é o Llama 3 8B ou o Llama 2 13B. Os demais modelos, apesar de terem obtido resultados bons em alguns critérios, ficaram aquém do esperado em algumas características essenciais para a QP1, como a informatividade e a adequação ao tópico.

Com relação à **QP2**, é possível afirmar que, para o cenário de avaliação investigado e com os instrumentos qualitativos de avaliação utilizados, a métrica proposta que faz a mescla do uso de divisão e conquista no BERTScore e considera o tamanho da sentença no cálculo da avaliação teve uma forte correlação com as avaliações humanas, como verificado pelo coeficiente de correlação de Pearson de 0,71. É importante ressaltar que o ajuste do parâmetro $\beta$ da fórmula foi feito de forma empírica considerando o tamanho dos sumários gerados para este trabalho e o tamanho dos textos de referência aqui tratados. Para cenários com diferentes tamanhos de sentenças ou diferentes proporções de tamanho entre sumários e textos originais, pode ser necessário ajustar esse parâmetro. Métodos de cálculo automático desse parâmetro podem ser explorados em trabalhos futuros.

# Capítulo 6

# Conclusão

Este trabalho apresentou o ToMAS (*Topic-based Multilevel Abstractive Summarization*), um pipeline para sumarização multinível de textos proposto para lidar com o volume expressivo de dados e com a limitação de contexto dos modelos de linguagem. Foram descritas as etapas de pré-processamento, extração de tópicos e a criação de resumos abstrativos, assim como o processo de avaliação quantitativa e qualitativa desses sumários, com ênfase em um método inédito de uso de BERTScore adaptado ao tamanho dos textos gerados e utilizando o texto original como base para a validação.

No decorrer do trabalho, foi possível observar que a técnica de sumarização multinível baseada em divisão e conquista permitiu gerenciar o limite de *tokens* de entrada dos modelos de linguagem. Esse método viabilizou a geração de sumários coesos e detalhados, mesmo em tópicos que reuniam milhares de *tweets*, como foi verificado nas bases de dados utilizadas no processamento. A avaliação qualitativa, realizada por meio de formulário aberto ao público, demonstrou que os melhores resultados – considerando clareza, informatividade e adequação ao tópico – foram obtidos com o modelo Llama 3, seguido do Llama 2, com ambos apresentando boa fluência e maior aderência aos tópicos identificados na análise dos *tweets*. Em contrapartida, modelos como o Mistral e o Bode, apesar de terem gerado textos mais sucintos, apresentaram limitações em cobertura de conteúdo e, algumas vezes, um alto viés e pouca informatividade nos sumários gerados.

A avaliação quantitativa foi igualmente relevante para quantificar o quão próximas as saídas dos modelos estavam de uma representação semântica dos *tweets* originais. O uso do BERTScore, mais especificamente o BERTScore-p proposto neste trabalho, mostrou uma boa correlação com as avaliações humanas. Entretanto, foi constatado que sumários menores acabavam favorecidos pelo BERTScore original, motivando a proposta de introduzir um fator de peso que levasse em conta o tamanho do sumário gerado. Dessa forma,

foi possível atenuar a vantagem de resumos que eram demasiadamente breves, aproximando a medição automática dos resultados das avaliações humanas. Essa variação do BERTScore proposta mostrou uma correlação forte com as médias de notas atribuídas pelos participantes da pesquisa, indicando ser uma contribuição relevante para o domínio de avaliação de sumários em cenários parecidos com o deste trabalho.

Dessa forma, os resultados evidenciam que o pipeline proposto no ToMAS é viável e contribui para a criação de sumários abstrativos em grandes coleções de textos, sobretudo aqueles produzidos em redes sociais. A avaliação qualitativa confirmou a utilidade prática dos sumários, pois demonstrou que as pessoas que os leram atribuíram, para os melhores sumários, uma nota alta de adequação ao tópico e informatividade. A técnica de sumarização multinível conseguiu contornar a limitação de contexto dos modelos e preservar a coesão nos textos finais.

## 6.1 Contribuições

Este trabalho respondeu às duas questões de pesquisa propostas na Seção 1.1. Com relação à efetividade do ToMAS na identificação dos principais assuntos discutidos nos *tweets* de forma clara e objetiva, os resultados qualitativos e quantitativos demonstraram que os modelos Llama 3 8B e Llama 2 13B utilizados na etapa de sumarização obtiveram boas pontuações no quesito informatividade, clareza, legibilidade e adequação ao tópico, além de obterem bons scores na avaliação quantitativa, o que demonstra que o método é capaz de extrair informações com bons resultados.

Já em relação à avaliação automática baseada no BERTScore levando em consideração o sumário gerado, o texto original e o tamanho da sentença, foi possível notar que, para este caso de uso em específico – com grandes conjuntos de textos de entrada e com a geração de sumários curtos – a métrica proposta foi capaz de obter uma forte correlação com as avaliações humanas. Entretanto, para cenários distintos com diferentes tamanhos de textos de entrada e de saída pode tornar-se necessário adaptar o parâmetro $\beta$ da fórmula proposta do score.

Assim, este trabalho contribui para a área com:

- **ToMAS**: Um método de sumarização abstrativa multinível de *tweets* disponível livremente em <https://github.com/LALIC-UFSCar/ToMAS>.

- **BERTScore-p**: Uma métrica de avaliação automática de sumários baseada no BERTScore, que não requer sumários de referência e segue a estratégia de divisão e conquista ponderada pelo tamanho do sumário.

## 6.2 Trabalhos futuros

Como perspectivas para pesquisas futuras, é possível destacar:

- ❏ Refinamento da avaliação automática: A proposta de ponderar o tamanho dos sumários no BERTScore apresentou bons resultados, mas há espaço para pesquisas que investiguem métodos sistemáticos de ajuste do fator de peso ($\beta$) ou que considerem, por exemplo, aspectos de coerência discursiva e não apenas semântica.

- ❏ Aplicação do ToMAS em outros domínios: O pipeline pode ser testado em dados de natureza diversa, como artigos de notícias, comentários de fóruns ou publicações de outras redes sociais, para verificar sua adaptabilidade.

- ❏ Mecanismos de filtragem dos *tweets* antes da sumarização: O ToMAS utiliza todos os *tweets* obtidos sem fazer nenhuma distinção de qualidade ou de relevância, fazendo apenas uma remoção de textos repetidos. A marcação prévia de *tweets* com baixa relevância poderia reduzir o custo computacional do processo e melhorar a informatividade dos sumários.

# Referências

ANGELOV, D. **Top2Vec: Distributed Representations of Topics**. 2020. Disponível em: <https://arxiv.org/abs/2008.09470>.

BANERJEE, S.; LAVIE, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: GOLDSTEIN, J. et al. (Ed.). **Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization**. Ann Arbor, Michigan: Association for Computational Linguistics, 2005. p. 65–72. Disponível em: <https://aclanthology.org/W05-0909>.

BENGIO, Y. et al. A neural probabilistic language model. **Journal of machine learning research**, Microtome Publ, BROOKLINE, v. 3, n. 6, p. 1137–1155, 2003. ISSN 1532-4435.

BLEI, D. M.; NG, A. Y.; JORDAN, M. I. Latent dirichlet allocation. **Journal of machine Learning research**, v. 3, n. Jan, p. 993–1022, 2003.

BROWN, P. F. et al. Class-based n-gram models of natural language. **Computational linguistics - Association for Computational Linguistics**, v. 18, n. 4, p. 467–479, 1992. ISSN 0891-2017.

BROWN, T. B. et al. **Language Models are Few-Shot Learners**. 2020.

CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. Density-based clustering based on hierarchical density estimates. In: PEI, J. et al. (Ed.). **Advances in Knowledge Discovery and Data Mining**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. p. 160–172. ISBN 978-3-642-37456-2.

CAO, M. **A Survey on Neural Abstractive Summarization Methods and Factual Consistency of Summarization**. 2022.

CASTELLUCCI, G. et al. UNITOR: Combining syntactic and semantic kernels for Twitter sentiment analysis. In: **Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)**. Atlanta, Georgia, USA: Association for Computational Linguistics, 2013. p. 369–374. Disponível em: <https://aclanthology.org/S13-2060>.

CHOWDHERY, A. et al. **PaLM: Scaling Language Modeling with Pathways**. 2022.

CHURCH, K.; HANKS, P. Word association norms, mutual information, and lexicography. **Computational linguistics**, v. 16, n. 1, p. 22–29, 1990.

COHEN, J. **Statistical power analysis for the behavioral sciences**. [S.l.]: routledge, 2013.

CORMEN, T. H. et al. **Introduction to algorithms**. [S.l.]: MIT press, 2022.

DEAN, J.; GHEMAWAT, S. Mapreduce: simplified data processing on large clusters. **Communications of the ACM**, ACM New York, NY, USA, v. 51, n. 1, p. 107–113, 2008.

DEROY, A.; GHOSH, K.; GHOSH, S. **How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization?** 2023.

DETTMERS, T. et al. **LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale**. 2022.

_____. **QLoRA: Efficient Finetuning of Quantized LLMs**. 2023.

DEUTSCH, D.; BEDRAX-WEISS, T.; ROTH, D. Towards question-answering as an automatic metric for evaluating the content quality of a summary. **Transactions of the Association for Computational Linguistics**, MIT Press, Cambridge, MA, v. 9, p. 774–789, 2021. Disponível em: <https://aclanthology.org/2021.tacl-1.47/>.

DEUTSCH ROTEM DROR, D. R. D. **A Statistical Analysis of Summarization Evaluation Metrics Using Resampling Methods**. 2021. 1132 p.

DEVLIN, J. et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019.

DUBEY, A. et al. **The Llama 3 Herd of Models**. 2024. Disponível em: <https://arxiv.org/abs/2407.21783>.

EGGER, C.-E. Y. R. **A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts**. 2022.

ESTER, M. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **kdd**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

FIKRI, F. B.; OFLAZER, K.; YANıKOğLU, B. Abstractive summarization with deep reinforcement learning using semantic similarity rewards. **Natural Language Engineering**, Cambridge University Press, p. 1–23, 2023.

FILHO, J. A. W. et al. The brWaC corpus: A new open resource for Brazilian Portuguese. In: **Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)**. Miyazaki, Japan: European Language Resources Association (ELRA), 2018. Disponível em: <https://aclanthology.org/L18-1686>.

FIRTH, J. R. The technique of semantics. **Transactions of the Philological Society**, v. 34, p. 36–73, 1935. Disponível em: <https://api.semanticscholar.org/CorpusID:143966536>.

FRANTAR, E. et al. **GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers**. 2023.

FU YAO; PENG, H.; KHOT, T. How does gpt obtain its ability? tracing emergent abilities of language models to their sources. **Yao Fu's Notion**, Dec 2022. Disponível em: <https://yaofu.notion.site/How-does-GPT-Obtain-its-Ability-Tracing-Emergent-Abilities-of-Language-Models-to-their-Source

FUKUNAGA, K. **Introduction to statistical pattern recognition**. [S.l.]: Elsevier, 2013.

GAO, J.; LIN, C.-Y. Introduction to the special issue on statistical language modeling. **ACM transactions on Asian language information processing**, ACM, New York, v. 3, n. 2, p. 87–93, 2004. ISSN 1530-0226.

GARCIA, G. L. et al. **Introducing Bode: A Fine-Tuned Large Language Model for Portuguese Prompt-Based Task**. 2024. Disponível em: <https://arxiv.org/abs/2401.02909>.

GARROSSINI, D. F. et al. Social networks, the fabrication of popular manifestations and attacks against democracy on january 8, 2023 in brazil. **Journal of Latin American Communication Research**, v. 11, n. 1, p. 34–48, 2023.

GHOLAMI, A. et al. **A Survey of Quantization Methods for Efficient Neural Network Inference**. 2021.

GOYAL, T.; LI, J. J.; DURRETT, G. **News Summarization and Evaluation in the Era of GPT-3**. 2023.

GROOTENDORST, M. **BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics**. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.4381785>.

GUPTA, S.; GUPTA, S. K. Abstractive summarization: An overview of the state of the art. **Expert Systems with Applications**, v. 121, p. 49–65, 2019. ISSN 0957-4174. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0957417418307735>.

GöKçIMEN, B. D. T. **Exploring Climate Change Discourse on Social Media and Blogs Using a Topic Modeling Analysis**. 2024. e32464–e32464 p.

HARRIS, Z. Distributional structure. **Word**, Taylor & Francis, v. 10, n. 2-3, p. 146–162, 1954. Disponível em: <https://link.springer.com/chapter/10.1007/978-94-009-8467-7_1>.

HOŠOVSKÝ, A. et al. A comprehensive survey of abstractive text summarization based on deep learning. **Computational Intelligence and Neuroscience**, Hindawi, v. 2022, p. 7132226, 08 2022. ISSN 1687-5265. Disponível em: <https://doi.org/10.1155/2022/7132226>.

INTERIAN, R.; RODRIGUES, F. A. Group polarization, influence, and domination in online interaction networks: a case study of the 2022 brazilian elections. **Journal of Physics: Complexity**, IOP Publishing, v. 4, n. 3, p. 035008, sep 2023. Disponível em: <https://doi.org/10.1088%2F2632-072x%2Facf6a4>.

IVISON, H. et al. **Camels in a Changing Climate: Enhancing LM Adaptation with Tulu 2**. 2023. Disponível em: <https://arxiv.org/abs/2311.10702>.

JAIN, A. K.; DUBES, R. C. **Algorithms for clustering data**. [S.l.]: Prentice-Hall, Inc., 1988.

JIANG, A. Q. et al. **Mistral 7B**. 2023. Disponível em: <https://arxiv.org/abs/2310.06825>.

_____. **Mixtral of Experts**. 2024. Disponível em: <https://arxiv.org/abs/2401.04088>.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **Journal of documentation**, MCB UP Ltd, v. 28, n. 1, p. 11–21, 1972.

JURAFSKY, D.; MARTIN, J. H. **Speech and language processing**. 2. ed., [pearson international edition]. ed. London [u.a.]: Prentice Hall, Pearson Education International, 2009. 1024 S. p. (Prentice Hall series in artificial intelligence). ISBN 0-13-504196-1, 978-0-13-504196-3. Disponível em: <http://aleph.bib.uni-mannheim.de/F/?func=find-b&request=285413791&find_code=020&adjacent=N&local_base=MAN01PUBLIC&x=0&y=0>.

KAPLAN, J. et al. Scaling laws for neural language models. **arXiv.org**, Cornell University Library, arXiv.org, Ithaca, 2020. ISSN 2331-8422.

KIROS, R. et al. **Skip-Thought Vectors**. 2015.

LACHI, R. L.; ROCHA, H. d. Aspectos básicos de clustering: conceitos e técnicas. **Núcleo de Informática Aplicada à Educação (Nied), UNICAMP-Instituto de Computação–Universidade Estadual de Campinas**, 2005.

LEE, D. D.; SEUNG, H. S. Learning the parts of objects by non-negative matrix factorization. **nature**, Nature Publishing Group UK London, v. 401, n. 6755, p. 788–791, 1999.

LIKERT, R. A technique for the measurement of attitudes. **Archives of Psychology**, 1932.

LIN, C.-Y. ROUGE: A package for automatic evaluation of summaries. In: **Text Summarization Branches Out**. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 74–81. Disponível em: <https://aclanthology.org/W04-1013>.

LIN, H.; NG, V. Abstractive summarization: A survey of the state of the art. **Proceedings of the AAAI Conference on Artificial Intelligence**, v. 33, n. 01, p. 9815–9822, Jul. 2019. Disponível em: <https://ojs.aaai.org/index.php/AAAI/article/view/5056>.

LIU, S.; HEALEY, C. G. **Abstractive Summarization of Large Document Collections Using GPT**. 2023.

LIU, Y. et al. **BRIO: Bringing Order to Abstractive Summarization**. 2022.

LOGESWARAN, L.; LEE, H. **An efficient framework for learning sentence representations**. 2018.

LUCKY, H.; SUHARTONO, D. Investigation of pre-trained bidirectional encoder representations from transformers checkpoints for indonesian abstractive text summarization. **Journal of Information and Communication Technology**, v. 21, n. 1, p. 71–94, Nov. 2021. Disponível em: <https://e-journal.uum.edu.my/index.php/jict/article/view/13548>.

MAATEN, L. V. D. et al. Dimensionality reduction: A comparative review. **Journal of Machine Learning Research**, v. 10, n. 66-71, p. 13, 2009.

MACHADO, J.; MISKOLCI, R. Das jornadas de junho à cruzada moral: o papel das redes sociais na polarização política brasileira. **Sociologia & Antropologia**, SciELO Brasil, v. 9, p. 945–970, 2019.

MANI, I. Automatic summarization. John Benjamins Publishing Company, 2001.

MCINNES JOHN HEALY, S. A. L. **How HDBSCAN Works**. 2016. Disponível em: <https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html>.

MCINNES, L.; HEALY, J.; MELVILLE, J. **UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction**. 2020.

MIKOLOV, T. et al. Efficient estimation of word representations in vector space. **arXiv.org**, Cornell University Library, arXiv.org, Ithaca, 2013. ISSN 2331-8422.

_____. Distributed representations of words and phrases and their compositionality. **arXiv.org**, Cornell University Library, arXiv.org, Ithaca, 2013. ISSN 2331-8422.

NAGWANI, N. K. **Summarizing large text collection using topic modeling and clustering based on MapReduce framework**. 2015.

NARAYAN, S.; COHEN, S. B.; LAPATA, M. **What is this Article about? Extreme Summarization with Topic-aware Convolutional Neural Networks**. 2019.

NOGUEIRA, R. et al. Document ranking with a pretrained sequence-to-sequence model. In: COHN, T.; HE, Y.; LIU, Y. (Ed.). **Findings of the Association for Computational Linguistics: EMNLP 2020**. Online: Association for Computational Linguistics, 2020. p. 708–718. Disponível em: <https://aclanthology.org/2020.findings-emnlp.63>.

PAPINENI, K. et al. Bleu: a method for automatic evaluation of machine translation. In: **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**. [S.l.: s.n.], 2002. p. 311–318.

PENNINGTON, J.; SOCHER, R.; MANNING, C. GloVe: Global vectors for word representation. In: **Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Doha, Qatar: Association for Computational Linguistics, 2014. p. 1532–1543. Disponível em: <https://aclanthology.org/D14-1162>.

PEREIRA, J.; NOGUEIRA, R.; LOTUFO, R. Large language models in summarizing social media for emergency management. In: . [S.l.: s.n.], 2024.

PETERS, M. E. et al. **Deep contextualized word representations**. 2018.

RADFORD, A. et al. Language models are unsupervised multitask learners. **OpenAI blog**, v. 1, n. 8, p. 9, 2019.

RAMOS, E. da S. Brazil's democracy under siege: January 8th and the threat of a new military dictatorship. **Georgetown Journal of International Affairs**, Johns Hopkins University Press, v. 24, n. 1, p. 65–71, 2023.

REIMERS, N.; GUREVYCH, I. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks**. 2019. Disponível em: <https://arxiv.org/abs/1908.10084>.

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

RÖDER, M.; BOTH, A.; HINNEBURG, A. Exploring the space of topic coherence measures. In: **Proceedings of the eighth ACM international conference on Web search and data mining**. [S.l.: s.n.], 2015. p. 399–408.

ROSENFELD, R. Two decades of statistical language modeling: where do we go from here? **Proceedings of the IEEE**, v. 88, n. 8, p. 1270–1278, 2000.

RUOCCO YUQIAN ZHUANG, R. T. N. R. J. M. K. L. H. A. Y. W. M. V. D. V. L. **A platform for connecting social media data to domain-specific topics using large language models: an application to student mental health**. 2024.

SANDER, J. Density-based clustering. In: _____. **Encyclopedia of Machine Learning**. Boston, MA: Springer US, 2010. p. 270–273. ISBN 978-0-387-30164-8. Disponível em: <https://doi.org/10.1007/978-0-387-30164-8_211>.

SANH, V. et al. **Multitask Prompted Training Enables Zero-Shot Task Generalization**. 2022.

SANKARANARAYANAN, J. et al. Twitterstand: News in tweets. In: **Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems**. New York, NY, USA: Association for Computing Machinery, 2009. (GIS '09), p. 42–51. ISBN 9781605586496. Disponível em: <https://doi-org.ez31.periodicos.capes.gov.br/10.1145/1653771.1653781>.

SANTOS, P. D. dos et al. Democracia sob ataque:: polarização política e produção de conteúdos hostis no twitter nas eleições de 2022. **Revista Debates**, v. 17, n. 1, p. 41–62, 2023.

SENO, E. R. M. et al. Semântica distribucional. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. BPLN, 2023. book chapter 10. ISBN 978-65-00-80693-9. Disponível em: <https://brasileiraspln.com/livro-pln/1a-edicao/parte5/cap10/cap10.html>.

SHANAHAN, M. **Talking About Large Language Models**. 2023.

SHEN, C. et al. **Are Large Language Models Good Evaluators for Abstractive Summarization?** 2023.

SINAGA, K. P.; YANG, M.-S. Unsupervised k-means clustering algorithm. **IEEE Access**, v. 8, p. 80716–80727, 2020.

SINGH, L. G. et al. Extraction and summarization of suicidal ideation evidence in social media content using large language models. In: **Proceedings of the Ninth Workshop on Computational Linguistics and Clinical Psychology, Association for Computational Linguistics**. [S.l.: s.n.], 2024.

SONI, M.; WADE, V. **Comparing Abstractive Summaries Generated by ChatGPT to Real Summaries Through Blinded Reviewers and Text Classification Algorithms**. 2023.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. Portuguese named entity recognition using bert-crf. **arXiv preprint arXiv:1909.10649**, 2019.

_____. Bertimbau: Pretrained bert models for brazilian portuguese. In: CERRI, R.; PRATI, R. C. (Ed.). **Intelligent Systems**. Cham: Springer International Publishing, 2020. p. 403–417. ISBN 978-3-030-61377-8.

SOUZA, J. W. d. C.; CARDOSO, P. C. F.; PAIXãO, C. A. Sumarização automática. In: CASELI, H. M.; NUNES, M. G. V. (Ed.). **Processamento de Linguagem Natural: Conceitos, Técnicas e Aplicações em Português**. 3. ed. BPLN, 2024. book chapter 24. ISBN 978-65-01-20581-6. Disponível em: <https://brasileiraspln.com/livro-pln/3a-edicao/parte-aplicacoes/cap-as/cap-as.html>.

SOUZA, R. M. V. de; LEAL, M. d. J. D. R. Disinformation at the service of chaos: Communication in brazil after the january 8 attacks. **Journal of Latin American Communication Research**, v. 11, n. 1, p. 4–17, 2023.

TOUVRON, H. et al. **Llama 2: Open Foundation and Fine-Tuned Chat Models**. 2023.

TWITTER. **#Eleições2022: veja dados sobre essas conversas no Twitter**. 2022. Twitter Blog. Disponível em: <https://blog.twitter.com/pt_br/topics/company/2022/-eleicoes2022--veja-dados-sobre-essas-conversas-no-twitter>.

VEEN, D. V. et al. Clinical text summarization: adapting large language models can outperform human experts. **Research Square**, American Journal Experts, 2023.

WEI, J. et al. **Finetuned Language Models Are Zero-Shot Learners**. 2022.

_____. Emergent abilities of large language models. **arXiv.org**, Cornell University Library, arXiv.org, Ithaca, 2022. ISSN 2331-8422.

_____. **Emergent Abilities of Large Language Models**. 2022.

_____. Chain-of-thought prompting elicits reasoning in large language models. **arXiv.org**, Cornell University Library, arXiv.org, Ithaca, 2023. ISSN 2331-8422.

_____. **Chain-of-Thought Prompting Elicits Reasoning in Large Language Models**. 2023.

WOLHANDLER, R. et al. **How "Multi"is Multi-Document Summarization?** 2022.

ZHANG, T. et al. **BERTScore: Evaluating Text Generation with BERT**. 2020. Disponível em: <https://arxiv.org/abs/1904.09675>.

_____. **BERTScore: Evaluating Text Generation with BERT**. 2020.

_____. **Benchmarking Large Language Models for News Summarization**. 2023.

ZHANG, X. et al. **∞Bench: Extending Long Context Evaluation Beyond 100K Tokens**. 2024. Disponível em: <https://arxiv.org/abs/2402.13718>.

ZHANG, Y. et al. **Chain of Agents: Large Language Models Collaborating on Long-Context Tasks**. 2024. Disponível em: <https://arxiv.org/abs/2406.02818>.

ZHAO, J. et al. **LongAgent: Scaling Language Models to 128k Context through Multi-Agent Collaboration**. 2024. Disponível em: <https://arxiv.org/abs/2402.11550>.

ZHAO, W. X. et al. **A Survey of Large Language Models**. 2023.

ZHOU, Z. et al. **LLM×MapReduce: Simplified Long-Sequence Processing using Large Language Models**. 2024. Disponível em: <https://arxiv.org/abs/2410.09342>.

# Apêndices

# APÊNDICE A

# Formulário A - Questionário Sumários

# Avaliação da eficácia de um modelo computacional de sumarização automática de *tweets* no contexto da política brasileira

Trata-se de um **projeto de mestrado** desenvolvido por mim, **Leonardo Capellaro**, sob orientação da **Profa. Dra. Helena de Medeiros Caseli**, do Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de São Carlos - UFSCar. Este estudo segue a Resolução CNS nº 510 de 2016 (Normas aplicáveis a pesquisas em Ciências Humanas e Sociais).

O estudo busca avaliar a qualidade de sumários (resumos) gerados usando uma tecnologia de inteligência artificial (*large language models*, LLMs) a partir de uma série de *tweets* que foram coletados no período de 08/01/2023 a 12/01/2023 - período conhecido pelos atos antidemocráticos ocorridos após as eleições presidenciais de 2022, caracterizados pelas invasões às sedes dos três poderes por opositores ao governo eleito. Os *tweets* coletados foram agrupados por tópicos, onde os tópicos mais importantes discutidos nos *tweets* foram selecionados com a ajuda de um cientista político da equipe do projeto e foram gerados sumários desses *tweets* utilizando LLMs, mais especificamente o Llama 2 - 13B, o Llama 3 - 8B, o Mistral - 7B e o Bode - 7B.

Para alcançar os objetivos propostos para este estudo, será utilizado o seguinte **instrumento de pesquisa**: este questionário, dividido em duas seções, sendo a primeira voltada à avaliação dos sumários e a segunda voltada à coleta do perfil dos participantes. A coleta de dados será realizada virtualmente, tendo como base esta plataforma eletrônica Google Forms, e buscará observar os graus de concordância (Escala Likert) sobre as questões expostas.

Com relação à **segurança na transferência e no armazenamento dos dados**, estes estarão armazenados na conta institucional do mestrando na nuvem da UFSCar e apenas os pesquisadores da equipe do projeto terão acesso a eles, assegurando o sigilo e a confidencialidade das informações dos participantes da pesquisa. Assim que o período de coleta for encerrado, o arquivo com as respostas será baixado da nuvem para o computador pessoal do mestrando que se responsabilizará pela segurança dos mesmos até que a pesquisa seja encerrada. Após o download para o computador pessoal do mestrando, o arquivo com as respostas será removido da nuvem. Todos os dados coletados nesta pesquisa ficarão armazenados em arquivo digital sob guarda e responsabilidade do pesquisador, por um período mínimo de 5 (cinco) anos após o término da pesquisa.

A participação nesta pesquisa é destinada a **pessoas que tenham 18 anos ou mais**. Por se tratar de sumários gerados a partir de *tweets*, o conteúdo sumarizado pode possuir opiniões diversas e uma alta carga emocional dos autores sobre um mesmo assunto, e não necessariamente reflete a realidade dos fatos ocorridos nem possui compromisso fiel com a verdade. Vale ressaltar que os conteúdos dos sumários foram gerados a partir do conteúdo dos *tweets*, sem a intervenção dos pesquisadores e, portanto, não devem ser associados ao posicionamento político destes.

Os participantes devem estar cientes dos seguintes **riscos potenciais**: (1) Cansaço físico ou mental ao responder o questionário; (2) Aborrecimento por estar sendo exposto a conteúdo que diverge de sua posição política ou que traz fatos com os quais o participante não concorda; (3) Desconforto mental por não entender alguma parte da informação apresentada no sumário ou como avaliar algum critério; (4) Evocação de sentimentos ou lembranças desagradáveis ao reviver momentos da história política recente do Brasil; (5) Dificuldades técnicas relacionadas ao uso da plataforma. Para **mitigar esses riscos**, o participante pode fazer pausas durante a participação retornando ao formulário quando o cansaço, aborrecimento ou desconforto tiver diminuído. Sempre que julgar necessário, o participante poderá entrar em contato com o pesquisador responsável.

Quanto aos **riscos** característicos do **ambiente virtual**, uma vez que a participação se dará por meio deste formulário disponível online, falhas e interrupções de energia ou internet poderão trazer riscos para o participante causando irritabilidade, ansiedade e frustração por ser impedido de concluir sua participação da forma como gostaria. Há, ainda, limitações por parte dos pesquisadores em assegurar total confidencialidade e proteção dos dados uma vez que a nuvem da UFSCar, na qual este formulário está armazenado, pode sofrer algum ataque malicioso ou falha que leve ao acesso indevido dos dados por pessoas não autorizadas. Para **mitigar o risco de violação dos dados**, assim que o período de coleta for encerrado, o arquivo com as respostas será baixado da nuvem para o computador pessoal do mestrando que se responsabilizará pela segurança dos mesmos até que a pesquisa seja encerrada. Após o download para o computador pessoal do mestrando, o arquivo com as respostas será removido da nuvem.

Para o **participante**, a participação nesta pesquisa trará como **benefício** direto a oportunidade de obter conhecimento sobre uma tecnologia de inteligência artificial (IA) e processamento de linguagem natural (PLN) bastante comentada na atualidade nas mídias e redes sociais, bem como a oportunidade de avaliar uma solução computacional com efeitos práticos na vida cotidiana. Para a **sociedade**, essa pesquisa tem potencial para contribuir com o desenvolvimento de tecnologias de IA e o avanço da pesquisa em PLN no Brasil que poderá ser usada pelo participante e por outras pessoas no futuro. As redes sociais têm sido muito utilizadas para disseminação de informações sobre política e uma tecnologia que auxilia no entendimento da grande quantidade de informação produzida diariamente nesse domínio pode apoiar o **letramento político do indivíduo e da sociedade** fortalecendo seus direitos cívicos.

Sua **participação é voluntária**, não haverá ganho financeiro e a **qualquer momento você irá decidir** se deseja participar e preencher o questionário, se deseja desistir da participação durante o preenchimento do questionário ou após o preenchimento, e poderá retirar seu consentimento sem nenhuma penalização ou prejuízo em sua relação com o pesquisador ou com a instituição. O participante tem o direito de não responder qualquer questão, sem necessidade de explicação ou justificativa para tal.

Durante toda a sua participação nesta pesquisa, você poderá entrar em contato com o pesquisador responsável pedindo esclarecimentos ou o **acompanhamento** que você julgue necessário e ele lhe prestará toda a **assistência**. Ressalta-se, também, que caso haja algum **gasto** decorrente de sua participação nesta pesquisa, este será ressarcido pela equipe de pesquisa, bastando, para isso, entrar em contato com o pesquisador responsável. Por fim, havendo algum **dano** decorrente da pesquisa, o participante terá direito a ser indenizado nos termos da Lei.

Suas respostas serão tratadas de forma anônima e confidencial, ou seja, em nenhum momento será divulgado seu nome em qualquer fase do estudo. Quando for necessário exemplificar determinada situação, sua privacidade será assegurada. Os resultados desta pesquisa serão publicados na Dissertação de mestrado do pesquisador que estará disponível no repositório institucional da UFSCar. Os resultados também poderão ser divulgados em eventos, revistas e/ou trabalhos científicos. Assim, **todos os participantes poderão ter acesso aos resultados desta pesquisa**.

Ao clicar em "**Aceito participar da pesquisa**", você irá:

1. Eletronicamente aceitar participar da pesquisa, o que corresponderá à assinatura deste termo (TCLE), o qual poderá ser impresso ou solicitado ao pesquisador via endereço de e-mail fornecido, se assim o desejar.

2. Responder ao questionário online que terá tempo previsto para seu preenchimento de aproximadamente **40 (quarenta) minutos**. Caso não concorde, basta fechar a página do navegador.

Caso desista de participar durante o preenchimento do questionário e antes de finalizá-lo, seus dados não serão enviados e nem recebidos pelo pesquisador, basta fechar a página do navegador. Caso tenha finalizado o preenchimento e enviado suas respostas do questionário e, após isso, decida desistir da participação, deverá informar o pesquisador desta decisão e ele descartará os seus dados recebidos sem nenhuma penalização.

Esta pesquisa foi aprovada pelo Comitê de Ética em Pesquisa em Seres Humanos (CEP) da UFSCar, que, vinculado à Comissão Nacional de Ética em Pesquisa (CONEP), tem a responsabilidade de garantir e fiscalizar que todas as pesquisas científicas com seres humanos obedeçam às normas éticas do País, e que os participantes de pesquisa tenham todos os seus direitos respeitados. O CEP-UFSCar funciona na Pró-Reitoria de Pesquisa da Universidade Federal de São Carlos, localizado no prédio da reitoria (área sul do campus São Carlos). Endereço: Rodovia Washington Luís, km 235 - CEP: 13.565-905 - São Carlos-SP. E-mail: cephumanos@ufscar.br. Telefone (16) 3351-9685. Horário de atendimento: das 08:30 às 11:30.

O CEP está vinculado à Comissão Nacional de Ética em Pesquisa (CONEP) do Conselho Nacional de Saúde (CNS), e o seu funcionamento e atuação são regidos pelas normativas do CNS/Conep. A CONEP tem a função de implementar as normas e diretrizes regulamentadoras de pesquisas envolvendo seres humanos, aprovadas pelo CNS, também atuando conjuntamente com uma rede de Comitês de Ética em Pesquisa (CEP) organizados nas instituições onde as pesquisas se realizam. Endereço: SRTV 701, Via W 5 Norte, lote D - Edifício PO 700, 3º andar - Asa Norte - CEP: 70719-040 - Brasília-DF. Telefone: (61) 3315-5877 E-mail: conep@saude.gov.br.

Você poderá imprimir uma via deste termo ou, se desejar, o pesquisador poderá encaminhar uma via assinada por e-mail ou da maneira como preferir.

**Número de aprovação pelo comitê de ética: CAAE: 82331024.7.0000.5504**

**Dados para contato (24 horas por dia e sete dias por semana):**

Pesquisador Responsável: Leonardo Capellaro

E-mail: leonardocapellaro@estudante.ufscar.br e leonardocapellaro@hotmail.com

* Indica uma pergunta obrigatória

1. E-mail *

2. **Declaro que entendi os objetivos, riscos e benefícios de minha participação na pesquisa e concordo em participar.** *

   *Marcar apenas uma oval.*

   ◯ Aceito participar da pesquisa.

**Instruções Gerais**

A seguir você avaliará os sumários gerados para 2 dos tópicos sumarizados em cada uma das bases contendo os *tweets* com menções aos candidatos Jair Bolsonaro e Luís Inácio Lula de Silva.

Os sumários foram geradas através do método de sumarização automática. A sumarização automática é o processo que visa condensar um conjunto de dados textuais automaticamente, tendo por objetivo criar um subconjunto, neste caso chamado sumário, que contempla os pedaços mais importantes de informação do texto original.

Para cada tópico listado, que vão de 1 a 2, serão apresentados **quatro sumários** a serem avaliados, um para cada modelo não necessariamente nesta ordem:
- Llama 2 - 13B
- Llama 3 - 8B
- Mistral - 7B
- Bode - 7B

Cada tópico possuirá as seguintes informações:
**Palavras representativas:** Lista de palavras representativas do tópico, separadas por vírgula.
***Tweets* representativos:** Lista de *tweets* considerados representativos daquele tópico.
**Explicação breve sobre o tópico:** Explicação escrita pelo cientista político da equipe do projeto.

## Bolsonaro - Primeiro Tópico

**Informações do Tópico:**

**Palavras representativas:** culpado, internado, cadeia, preso, hospital, anistia, bolsonaro, covarde, dores, eua
**Tweets representativos:** 'Tem que ser! Bolsonaro é culpado! Sem anistia!', 'Bolsonaro preso! Bolsonaro é culpado! Sem anistia!', 'Bolsonaro é culpado sem anistia bolsonaro na cadeia'
**Explicação breve sobre o tópico:** Esse tópico ilustra as críticas às ações tomadas por Bolsonaro em relação aos atos antidemocráticos, como sua omissão após a ocorrência, justificada por uma internação médica e as acusações de incentivo aos atos, devido à escândalos que o conectavam a uma possível tentativa de golpe de Estado, por exemplo.

**Considere os seguintes critérios em sua avaliação:**

**1 - Fluência/Naturalidade:** Refere-se à facilidade de leitura do texto, sem interrupções ou necessidade de grande esforço para o entendimento. Um texto fluente deve ser compreensível à primeira leitura.

**2 - Uso correto da linguagem:** Refere-se à qualidade da escrita do texto, avaliando se o texto está livre de erros ortográficos, sintáticos (p.ex. erros de concordância de gênero ou número), além de erros de coesão.

**3 - Clareza/Legibilidade:** Refere-se ao quanto o texto é direto e fácil de entender, evitando jargões desnecessários e construções complexas que dificultem a compreensão, levando em conta fatores como a escolha de palavras e a estrutura das frases.

**4 - Informatividade:** Refere-se à capacidade do texto de transmitir informações, tornando a leitura útil e interessante. É a habilidade do texto de efetivamente comunicar e informar seu público de maneira clara e significativa. Considere que um sumário pode ser informativo mesmo que traga informações com as quais você não concorde, uma vez que pode retratar o viés político de quem escreveu.

**5 - Redundância:** Refere-se à propriedade do texto de repetir desnecessariamente informações. Um texto com baixa redundância é conciso e direto, evitando a repetição de ideias já apresentadas.

**6 - Adequação ao Tópico:** Refere-se à adequação do conteúdo do sumário ao tópico, verificada com base nas palavras e *tweets* representativos, bem como na breve explicação sobre o tópico (fornecidos com cada tópico). Um sumário é adequado ao tópico se seu texto aborda o tópico citado de maneira precisa e relevante.

**Lembrando:** Cada um dos quatro sumários abaixos foram gerados por modelos de linguagem diferentes resumindo um mesmo tópico.

**Bolsonaro - Primeiro Tópico - Sumário 1:**
Antigo presidente brasileiro Jair Bolsonaro tem sido um assunto de polêmica devido a diversas acusações contra ele. Alguns críticos o acusam de promover golpismo, terrorismo e violência política. Outros criticam sua lenta resposta às necessidades do país durante a pandemia de COVID-19. Em resumo, Bolsonaro foi uma figura divisiva que enfrentou diversas polêmicas tanto no Brasil quanto internacionalmente.

Note: This output is based on the context provided in the input and does not reflect any personal opinions or biases of the author.

3. **Bolsonaro - Primeiro Tópico - Sumário 1:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

4. **Bolsonaro - Primeiro Tópico - Sumário 1:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

5. **\leoBolsonaro - Primeiro Tópico - Sumário 1:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

6. **Bolsonaro - Primeiro Tópico - Sumário 1:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

7. **Bolsonaro - Primeiro Tópico - Sumário 1:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

8. **Bolsonaro - Primeiro Tópico - Sumário 1:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Bolsonaro - Primeiro Tópico - Sumário 2:**

A discussão sobre o presidente Jair Bolsonaro é marcada por críticas severas à sua conduta política e social. Os autores consideram que Bolsonaro é responsável por um período de violência e instabilidade no país, com discursos que incitam ódio e divisão. Eles também o acusam de não ter tomado medidas para prevenir a violência e de ter uma postura de silêncio e covardia. A falta de ação do governo Bolsonaro em relação às manifestações golpistas e à violência política é outra crítica comum. Além disso, os autores criticam a impunidade e a falta de justiça no país, considerando que Bolsonaro e seus apoiadores não foram punidos por seus atos e que a sociedade brasileira está pagando o preço da falta de responsabilidade política.

A discussão também aborda a questão da responsabilidade de Bolsonaro por seus atos e declarações, com alguns considerando que ele é culpado por incitar a violência e que deve ser punido, enquanto outros defendem que ele é um líder que merece respeito e que sua responsabilidade é limitada à sua gestão política. A situação política no Brasil é considerada grave e a necessidade de mudanças é amplamente sentida.

Os autores também criticam a forma como Bolsonaro lidou com a pandemia e a crise política no Brasil, afirmando que ele não tomou medidas eficazes para combater a doença e que sua ação política foi caracterizada por covardia e falta de liderança. Além disso, eles acusam Bolsonaro de ter abandonado o país e fugido para os EUA, onde está internado em um hospital.

A discussão sobre Bolsonaro é intensa e polarizada, com alguns autores considerando que ele é culpado por vários crimes, incluindo o terrorismo, o genocídio e o golpe. Alguns argumentam que ele é um bandido e que deve ser preso imediatamente, enquanto outros acreditam que ele é um covarde que fugiu do país para escapar da responsabilidade por seus atos. Em geral, a discussão sobre Bolsonaro é marcada por críticas à sua política e ao seu estilo de liderança, bem como à sua responsabilidade por não ter tomado medidas para prevenir a invasão ao Congresso Nacional.

9. **Bolsonaro - Primeiro Tópico - Sumário 2:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

10. **Bolsonaro - Primeiro Tópico - Sumário 2:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

11. **Bolsonaro - Primeiro Tópico - Sumário 2:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

12. **Bolsonaro - Primeiro Tópico - Sumário 2:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

13. **Bolsonaro - Primeiro Tópico - Sumário 2:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

14. **Bolsonaro - Primeiro Tópico - Sumário 2:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Bolsonaro - Primeiro Tópico - Sumário 3:**

O ex-presidente Jair Bolsonaro foi um líder controverso no Brasil, gerando opiniões divergentes sobre suas políticas e ações. Enquanto alguns o apoiavam como um líder forte e decisivo, outros o criticavam por sua gestão da economia, destruição dos direitos humanos e falta de compromisso com a democracia. A opinião sobre Bolsonaro é dividida e complexa, com muitas pessoas tendo diferentes pontos de vista sobre sua liderança e ações como presidente do Brasil.

Alguns críticos acusam Bolsonaro de ser autoritário, covarde e de ter causado danos à democracia e aos direitos humanos. Eles também expressaram preocupações com agressão e violência, assédio e intimidação, desrespeito às instituições democráticas, falta de transparência e corrupção, e fomento da desigualdade social. No entanto, é importante notar que essas são apenas opiniões e que a realidade pode ser mais complexa e nuanced.

O governo de Bolsonaro foi marcado por controversias e polarização, com muitos manifestando preocupações sobre sua liderança e o impacto que ela teve no Brasil. Críticos acusaram o presidente de fomentar um ambiente de ódio e intolerância, e alguns até chamaram para sua impeachment. No entanto, muitos outros expressaram preocupações sobre sua liderança e o impacto que ela teve nas instituições democráticas e na fibra social do país.

A comunidade online também questiona a legitimidade de seu governo e o impacto negativo que ele teve sobre o país. No entanto, existem também opiniões divergentes defendendo Bolsonaro e justificando suas ações. Em resumo, a opinião sobre Bolsonaro é dividida e complexa, com muitas pessoas tendo diferentes pontos de vista sobre sua liderança e ações como presidente do Brasil. O futuro presidente do Brasil deve ser escolhido com base em sua capacidade de governar o país de forma justa e equitativa para todos os cidadãos, independentemente de suas opiniões políticas ou de sua personalidade.

15. **Bolsonaro - Primeiro Tópico - Sumário 3:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

16. **Bolsonaro - Primeiro Tópico - Sumário 3:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

17. **Bolsonaro - Primeiro Tópico - Sumário 3:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

18. **Bolsonaro - Primeiro Tópico - Sumário 3:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

19. **Bolsonaro - Primeiro Tópico - Sumário 3:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

20. **Bolsonaro - Primeiro Tópico - Sumário 3:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Bolsonaro - Primeiro Tópico - Sumário 4:**
"O presidente Bolsonaro foi criticado por sua gestão do país, com muitos acreditando que ele é um "mito" que falhou o povo. Seus seguidores foram descritos como "extremistas" e "criminalmente covardes", e ele foi acusado de ser responsável pelo terrorismo no país. Ele também foi criticado por suas políticas, como a promoção dos direitos humanos e da liberdade de expressão."

21. **Bolsonaro - Primeiro Tópico - Sumário 4:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

22. **Bolsonaro - Primeiro Tópico - Sumário 4:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

23. **Bolsonaro - Primeiro Tópico - Sumário 4:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

24. **Bolsonaro - Primeiro Tópico - Sumário 4:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

25. **Bolsonaro - Primeiro Tópico - Sumário 4:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

26. **Bolsonaro - Primeiro Tópico - Sumário 4:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

## Bolsonaro - Segundo Tópico

**Informações do Tópico:**

**Palavras representativas:** patriota, destrói, patrimônio, pátria, cidadão, infiltrado, bandeira, conservador, patriotas, público.
**Tweets representativos:** 'Patriota não destrói'., 'O patriota aí.', 'Patriota herói'.
**Explicação breve sobre o tópico:** Esse tópico ironiza a utilização do termo "patriota" por parte dos manifestantes como referencia à eles mesmos, assim como, através especialmente do termo 'infiltrado', engloba a reação bolsonarista de acusar opositores de cometerem os atos terroristas enquanto disfarçados e infiltrados no meio de uma manifestação pacífica e legítima.

**Considere os seguintes critérios em sua avaliação:**

**1 - Fluência/Naturalidade:** Refere-se à facilidade de leitura do texto, sem interrupções ou necessidade de grande esforço para o entendimento. Um texto fluente deve ser compreensível à primeira leitura.

**2 - Uso correto da linguagem:** Refere-se à qualidade da escrita do texto, avaliando se o texto está livre de erros ortográficos, sintáticos (p.ex. erros de concordância de gênero ou número), além de erros de coesão.

**3 - Clareza/Legibilidade:** Refere-se ao quanto o texto é direto e fácil de entender, evitando jargões desnecessários e construções complexas que dificultem a compreensão, levando em conta fatores como a escolha de palavras e a estrutura das frases.

**4 - Informatividade:** Refere-se à capacidade do texto de transmitir informações, tornando a leitura útil e interessante. É a habilidade do texto de efetivamente comunicar e informar seu público de maneira clara e significativa. Considere que um sumário pode ser informativo mesmo que traga informações com as quais você não concorde, uma vez que pode retratar o viés político de quem escreveu.

**5 - Redundância:** Refere-se à propriedade do texto de repetir desnecessariamente informações. Um texto com baixa redundância é conciso e direto, evitando a repetição de ideias já apresentadas.

**6 - Adequação ao Tópico:** Refere-se à adequação do conteúdo do sumário ao tópico, verificada com base nas palavras e *tweets* representativos, bem como na breve explicação sobre o tópico (fornecidos com cada tópico). Um sumário é adequado ao tópico se seu texto aborda o tópico citado de maneira precisa e relevante.

**Lembrando:** Cada um dos quatro sumários abaixos foram gerados por modelos de linguagem diferentes resumindo um mesmo tópico.

**Bolsonaro - Segundo Tópico - Sumário 1:**

A discussão sobre o que é um patriota é um tema complexo e multifacetado, envolvendo diferentes perspectivas e conceitos sobre o assunto. Os autores concordam que um patriota é alguém que ama e defende seu país, sua cultura e seu patrimônio, e que a definição de patriota não pode ser reduzida apenas a uma palavra ou um símbolo. No entanto, há diferentes opiniões sobre o que é necessário para ser considerado um verdadeiro patriota.

Alguns autores consideram que um patriota é alguém que defende a pátria e o patrimônio nacional, respeita a Constituição e a democracia, e não promove atos terroristas ou destruidores. Outros autores defendem a importância do patriotismo, argumentando que é necessário defender a pátria e seus valores, e que a defesa da pátria é uma questão de amor e respeito pela nação.

Os autores também destacam a importância da educação e da formação de uma consciência patriótica que respeite a história e a cultura do país. Eles consideram que o patriotismo não é apenas uma questão de emoção ou de identidade, mas sim de ação e de compromisso com a defesa e o respeito à pátria.

A discussão também aborda a crítica à ideia de que um patriota seja alguém que apoia a ditadura ou a intolerância, e que o patriotismo pode ser usado como um pretexto para justificar a violência e a opressão. Além disso, os autores criticam a falta de conhecimento histórico e cultural de alguns indivíduos que se consideram patriotas, o que os leva a cometer atos de vandalismo e destruição.

Em resumo, a discussão sobre o que é ser patriota é complexa e multifacetada, envolvendo aspectos como respeito à Constituição, conhecimento histórico e cultural, dignidade humana e ação em defesa da pátria.

27. **Bolsonaro - Segundo Tópico - Sumário 1:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

28. **Bolsonaro - Segundo Tópico - Sumário 1:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

29. **Bolsonaro - Segundo Tópico - Sumário 1:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

30. **Bolsonaro - Segundo Tópico - Sumário 1:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

31. **Bolsonaro - Segundo Tópico - Sumário 1:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

32. **Bolsonaro - Segundo Tópico - Sumário 1:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Bolsonaro - Segundo Tópico - Sumário 2:**
Este texto aborda uma complexa conversação entre várias pessoas sobre a identidade e ações de uma mulher referida como "ela" ou "a senhora". A conversação abrange diversos assuntos, incluindo dúvidas e defesas, identidade e motivos, e termos específicos e frases usadas no texto. No entanto, sem contexto adicional ou informações, não é possível completamente entender o significado ou importância de estes assuntos ou as termos específicos e frases usadas no texto.

33. **Bolsonaro - Segundo Tópico - Sumário 2:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

34. **Bolsonaro - Segundo Tópico - Sumário 2:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

35. **Bolsonaro - Segundo Tópico - Sumário 2:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

36. **Bolsonaro - Segundo Tópico - Sumário 2:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

37. **Bolsonaro - Segundo Tópico - Sumário 2:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

38. **Bolsonaro - Segundo Tópico - Sumário 2:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Bolsonaro - Segundo Tópico - Sumário 3:**

O termo "patriota" está sendo usado de maneira complexa e ambivalente na cultura brasileira, refletindo diferentes perspectivas e interpretações sobre o que significa ser um patriota na atualidade. A maioria dos autores usa o termo para se referir a pessoas que defendem as causas nacionais e são críticas ao governo e às políticas atuais, mas também pode ser usado de maneira ironica ou negativa para se referir a aqueles que consideram que estão traindo ou destruindo o país. A preocupação com a falta de respeito pela liberdade e pelos direitos humanos é mencionada por alguns autores, assim como a crítica à falta de respeito pela democracia e a presença de grupos extremistas. Os autores também destacam a importância de ser honesto, trabalhar e amar a bandeira, e que um patriota nunca destruiria patrimônios e não desejaria o afundamento do país. No entanto, a maioria dos textos não apresenta uma visão clara e consistente sobre quem é um verdadeiro patriota ou como uma pessoa deve ser considerada um patriota. Em vez disso, os textos parecem se concentrar em criticar aqueles que são vistos como "inimigos" do país, sem fornecer uma análise mais profunda da natureza do patriotismo ou de suas implicações.

39. **Bolsonaro - Segundo Tópico - Sumário 3:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

40. **Bolsonaro - Segundo Tópico - Sumário 3:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

41. **Bolsonaro - Segundo Tópico - Sumário 3:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

42. **Bolsonaro - Segundo Tópico - Sumário 3:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

43. **Bolsonaro - Segundo Tópico - Sumário 3:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

44. **Bolsonaro - Segundo Tópico - Sumário 3:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Bolsonaro - Segundo Tópico - Sumário 4:**

O patriotismo é uma qualidade que envolve amor à pátria, lealdade e dedicação. Os patriotas defendem sua pátria e seus valores, e trabalham para proteger o patrimônio histórico de seu país. Um cidadão de bem e um patriota devem ser reconhecidos por suas ações, mas não por sua origem ou religião. No entanto, também há riscos associados ao patriotismo, como vandalismo e terrorismo, e os patriotas podem ser infiltrados ou anti-patriotas.

45. **Bolsonaro - Segundo Tópico - Sumário 4:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

46. **Bolsonaro - Segundo Tópico - Sumário 4:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

47. **Bolsonaro - Segundo Tópico - Sumário 4:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

48. **Bolsonaro - Segundo Tópico - Sumário 4:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

49. **Bolsonaro - Segundo Tópico - Sumário 4:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

50. **Bolsonaro - Segundo Tópico - Sumário 4:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

## Lula - Primeiro Tópico

**Informações do Tópico:**

**Palavras representativas:** concentração, campo, crianças, idosos, campos, comida, idosa, nazista, morreu.
**Tweets representativos:** 'Campo de concentração de lula e xandão...', 'Campo de concentração n do lula', 'O campo de concentração do lula'.
**Explicação breve sobre o tópico:** Esse tópico contém as acusações de "violação do direitos humanos" por parte do governo petista após os atos, quando foram montadas prisões improvisadas para deter os manifestantes que cometeram crimes. Vê-se que a oposição utilizou termos como "campos de concentração" e apelou para o senso moral ao acusar maus-tratos à idosos e crianças, categorias tidas como vulneráveis afim de apelar ao senso moral de justiça dos leitores.

**Considere os seguintes critérios em sua avaliação:**

**1 - Fluência/Naturalidade:** Refere-se à facilidade de leitura do texto, sem interrupções ou necessidade de grande esforço para o entendimento. Um texto fluente deve ser compreensível à primeira leitura.

**2 - Uso correto da linguagem:** Refere-se à qualidade da escrita do texto, avaliando se o texto está livre de erros ortográficos, sintáticos (p.ex. erros de concordância de gênero ou número), além de erros de coesão.

**3 - Clareza/Legibilidade:** Refere-se ao quanto o texto é direto e fácil de entender, evitando jargões desnecessários e construções complexas que dificultem a compreensão, levando em conta fatores como a escolha de palavras e a estrutura das frases.

**4 - Informatividade:** Refere-se à capacidade do texto de transmitir informações, tornando a leitura útil e interessante. É a habilidade do texto de efetivamente comunicar e informar seu público de maneira clara e significativa. Considere que um sumário pode ser informativo mesmo que traga informações com as quais você não concorde, uma vez que pode retratar o viés político de quem escreveu.

**5 - Redundância:** Refere-se à propriedade do texto de repetir desnecessariamente informações. Um texto com baixa redundância é conciso e direto, evitando a repetição de ideias já apresentadas.

**6 - Adequação ao Tópico:** Refere-se à adequação do conteúdo do sumário ao tópico, verificada com base nas palavras e *tweets* representativos, bem como na breve explicação sobre o tópico (fornecidos com cada tópico). Um sumário é adequado ao tópico se seu texto aborda o tópico citado de maneira precisa e relevante.

**Lembrando:** Cada um dos quatro sumários abaixos foram gerados por modelos de linguagem diferentes resumindo um mesmo tópico.

**Lula - Primeiro Tópico - Sumário 1:**
O regime do presidente Luiz Inácio Lula da Silva no Brasil tem sido criticado por suas políticas que levam a condições semelhantes ao campo de concentração nazista. Os autores citam o regime como uma ditadura que viola os direitos humanos e é frequentemente comparado ao regime nazista da Alemanha durante a Segunda Guerra Mundial. As pessoas presas no campo de concentração estão passando por fome, frio e falta de água, com poucas oportunidades de sair. O governo Lula tem sido acusado de praticar um regime de terrorismo contra essas pessoas, que são mantidas em condições inumanas sem direitos humanos. A comunidade internacional deve tomar medidas para proteger esses indivíduos e garantir que eles sejam tratados com dignidade e respeito.

51. **Lula - Primeiro Tópico - Sumário 1:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

52. **Lula - Primeiro Tópico - Sumário 1:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

53. **Lula - Primeiro Tópico - Sumário 1:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

54. **Lula - Primeiro Tópico - Sumário 1:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

55. **Lula - Primeiro Tópico - Sumário 1:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

56. **Lula - Primeiro Tópico - Sumário 1:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Lula - Primeiro Tópico - Sumário 2:**

O regime atual no Brasil enfrenta críticas por suas ações ilegais e inumanas, como o estabelecimento de campos de concentração para detidos políticos e a falta de direitos humanos. Relatos incluem condições inumanas, tortura, fome e desidratação, e a situação é comparada ao Holocausto e ao regime nazista. A sociedade civil deve se organizar e protestar contra essas ações ilegais e inumanas, e exigir responsabilidade por essas mortes e sofrimentos. A comunidade internacional está se preocupando com esses eventos e exigindo responsabilidades por esses crimes. O mundo todo deve se preocupar com a situação no Brasil e trabalhar juntos para garantir que os direitos humanos sejam respeitados e que a democracia seja preservada.

57. **Lula - Primeiro Tópico - Sumário 2:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

58. **Lula - Primeiro Tópico - Sumário 2:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

59. **Lula - Primeiro Tópico - Sumário 2:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

60. **Lula - Primeiro Tópico - Sumário 2:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

61. **Lula - Primeiro Tópico - Sumário 2:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

62. **Lula - Primeiro Tópico - Sumário 2:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Lula - Primeiro Tópico - Sumário 3:**

O Brasil está enfrentando um grave cenário de violação dos direitos humanos, com brasileiros sendo presos em campos de concentração em condições desumanas, sem acesso a alimentos, água, banheiros e assistência médica. O presidente Lula é acusado de criar um regime ditatorial e fascista, com o apoio do exército e da polícia federal, e de violar os direitos humanos. A situação é descrita como um "holocausto" e um "genocídio", com brasileiros sendo tratados como judeus durante o regime nazista. A população brasileira está sofrendo com a falta de alimentos, água e serviços básicos, e muitas pessoas estão morrendo de fome e sede. A sociedade brasileira está dividida e a imprensa está sendo censurada. Muitas pessoas estão sendo perseguidas e presas sem processo legal. A Comissão de Direitos Humanos e a mídia estão criticadas por não terem tomado medidas para denunciar as violações dos direitos humanos ocorridas no campo de concentração. A situação é considerada uma crise política e social grave, com a criação de campos de concentração pelo governo do presidente Lula, que estão sendo utilizados para prender e maltratar pessoas que se manifestam contra o governo.

63. **Lula - Primeiro Tópico - Sumário 3:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

64. **Lula - Primeiro Tópico - Sumário 3:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

65. **Lula - Primeiro Tópico - Sumário 3:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

66. **Lula - Primeiro Tópico - Sumário 3:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

67. **Lula - Primeiro Tópico - Sumário 3:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

68. **Lula - Primeiro Tópico - Sumário 3:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Lula - Primeiro Tópico - Sumário 4:**
Este texto descreve uma situação crítica no Brasil, em que milhares de pessoas, incluindo crianças e idosos, foram detidas ilegalmente em campos de concentração governamentais. Essas pessoas estão sendo expostas a condições inumanas, como falta de comida, água, instalações sanitárias adequadas e cuidados de saúde por mais de 20 horas. Como resultado, muitas dessas pessoas já se submeteram às circunstâncias e morreram.

Este texto também menciona a criação do primeiro campo de concentração pelo governo de Lula e do PT, assim como a criação de outros campos de concentração e o uso do termo "holocausto brasileiro" para descrever esses eventos. O texto também menciona o uso do termo "nazista" para descrever Lula e o PT, assim como o uso do termo "genocida" para descrever Lula.

Este texto também faz referência a uma vídeo produzido por Alessandra Cristina que é mencionado no texto mas não diretamente citado ou linkado no texto. O texto também menciona as mortes de vários patriotas na concentração e acusa Lula de traíram seus próprios cidadãos ao criar este campo e permitir que essas mortes ocorressem.

É importante que a comunidade internacional se posicione sobre esta situação e exija que o governo brasileiro liberte as pessoas detidas ilegalmente em campos de concentração, forneça-lhes comida, água, instalações sanitárias adequadas e cuidados de saúde por mais de 20 horas. Só assim podemos esperar ver um fim às tragédias que estão acontecendo no Brasil."

69. **Lula - Primeiro Tópico - Sumário 4:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

70. **Lula - Primeiro Tópico - Sumário 4:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

71. **Lula - Primeiro Tópico - Sumário 4:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

72. **Lula - Primeiro Tópico - Sumário 4:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

73. **Lula - Primeiro Tópico - Sumário 4:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

74. **Lula - Primeiro Tópico - Sumário 4:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

## Lula - Segundo Tópico

**Informações do Tópico:**

**Palavras representativas:** presidiário, ex, presidência, vergonha, piada, condenado, governado, um, descondenado, abominável.
**Tweets representativos:** 'Ex presidiário*' ,'E o ex presidiário?' , 'Só o ex presidiário.'.
**Explicação breve sobre o tópico:** Esse tópico ilustra uma forma que críticos ao governo Lula utilizam para se referenciar ao atual Presidente, citando sua condenação anulada, que acarretou na prisão do político. Através dessa referência, apoiadores de Bolsonaro tentam desviar as acusações do ex-candidato do PL como cúmplice da tentativa de golpe ao tentar contrastar a índole de ambos os políticos

**Considere os seguintes critérios em sua avaliação:**

**1 - Fluência/Naturalidade:** Refere-se à facilidade de leitura do texto, sem interrupções ou necessidade de grande esforço para o entendimento. Um texto fluente deve ser compreensível à primeira leitura.

**2 - Uso correto da linguagem:** Refere-se à qualidade da escrita do texto, avaliando se o texto está livre de erros ortográficos, sintáticos (p.ex. erros de concordância de gênero ou número), além de erros de coesão.

**3 - Clareza/Legibilidade:** Refere-se ao quanto o texto é direto e fácil de entender, evitando jargões desnecessários e construções complexas que dificultem a compreensão, levando em conta fatores como a escolha de palavras e a estrutura das frases.

**4 - Informatividade:** Refere-se à capacidade do texto de transmitir informações, tornando a leitura útil e interessante. É a habilidade do texto de efetivamente comunicar e informar seu público de maneira clara e significativa. Considere que um sumário pode ser informativo mesmo que traga informações com as quais você não concorde, uma vez que pode retratar o viés político de quem escreveu.

**5 - Redundância:** Refere-se à propriedade do texto de repetir desnecessariamente informações. Um texto com baixa redundância é conciso e direto, evitando a repetição de ideias já apresentadas.

**6 - Adequação ao Tópico:** Refere-se à adequação do conteúdo do sumário ao tópico, verificada com base nas palavras e *tweets* representativos, bem como na breve explicação sobre o tópico (fornecidos com cada tópico). Um sumário é adequado ao tópico se seu texto aborda o tópico citado de maneira precisa e relevante.

**Lembrando:** Cada um dos quatro sumários abaixos foram gerados por modelos de linguagem diferentes resumindo um mesmo tópico.

**Lula - Segundo Tópico - Sumário 1:**

O tema abordado é um ex-presidente que já esteve preso por corrupção e agora está candidato a presidente novamente no Brasil. Essa situação levanta uma série de complexos problemas que requerem cuidadosa consideração e análise.

Primeiro e principalmente, é importante lembrar que todos merecem uma segunda chance, independentemente de suas falhas ou transgressões passadas. Por outro lado, é crucial manter indivíduos responsáveis pelas suas ações e garantir que a justiça seja servida de forma justa e imparcial.

Em situação de um ex-presidente que já esteve preso por corrupção e agora está candidato a presidente novamente, existem uma série de complexos problemas que requerem cuidadosa consideração e análise.

Primeiro e principalmente, é importante considerar as ações e transgressões passadas do indivíduo, bem como sua plataforma e políticas propostas atualmente. Em situação de um ex-presidente que já esteve preso por corrupção e agora está candidato a presidente novamente, é essencial cuidadosamente avaliar as ações e transgressões passadas do indivíduo, bem como sua plataforma e políticas propostas atualmente, para determinar se ele é apto a servir de presidente novamente.

Outro importante consideração é o sentimento e percepção geral da sociedade em relação ao indivíduo. Em situação de um ex-presidente que já esteve preso por corrupção e agora está candidato a presidente novamente, é essencial considerar como as ações e transgressões passadas do indivíduo, bem como sua plataforma e políticas propostas atualmente, têm impactado o público em geral, assim como sua capacidade de governar e liderar o país novamente.

Por fim, a decisão de apoiar ou não um ex-presidente que já esteve preso por corrupção e agora está candidato a presidente novamente depende de uma série complexa de fatores que requerem cuidadosa consideração e análise. Esses fatores incluem, mas não se limitam às, as ações e transgressões passadas do indivíduo, bem como sua plataforma e políticas propostas atualmente, assim como o sentimento e percepção geral da sociedade em relação ao indivíduo e sua capacidade de governar e liderar o país novamente.

75. **Lula - Segundo Tópico - Sumário 1:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

76. **Lula - Segundo Tópico - Sumário 1:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

77. **Lula - Segundo Tópico - Sumário 1:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

78. **Lula - Segundo Tópico - Sumário 1:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

79. **Lula - Segundo Tópico - Sumário 1:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

80. **Lula - Segundo Tópico - Sumário 1:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Lula - Segundo Tópico - Sumário 2:**

A eleição de um ex-presidiário para a presidência do Brasil é um tema que tem gerado grande controvérsia e indignação no país. Os autores dos textos apresentados criticam a escolha de um indivíduo condenado por crimes para o cargo mais alto do país, considerando-a uma aberração e uma vergonha nacional. Eles argumentam que um criminoso não pode ser considerado um líder legítimo e que sua eleição é um ato de justiça desfeita. Além disso, os autores questionam a capacidade do ex-presidiário de governar o país, argumentando que sua eleição é um ato de desordem e que o país está em caos.

A sociedade brasileira também é criticada por apoiar o ex-presidiário, considerando-a uma parte da sociedade que não valoriza a justiça e a moralidade. Os autores argumentam que a escolha do ex-presidiário é um reflexo da corrupção e da impunidade no país, e que isso pode levar a um agravamento da situação política e econômica do Brasil.

No entanto, há também aqueles que defendem a escolha do ex-presidiário, argumentando que ele tem o apoio do povo e que sua eleição é um reflexo da vontade popular. A população brasileira está dividida em relação à eleição do ex-presidiário, com alguns defendendo a escolha e outros a criticando.

Em resumo, a eleição de um ex-presidiário para a presidência do Brasil é um tema que tem gerado grande controvérsia e indignação no país. Os autores dos textos apresentados criticam a escolha de um indivíduo condenado por crimes para o cargo mais alto do país, considerando-a uma aberração e uma vergonha nacional. Eles argumentam que um criminoso não pode ser considerado um líder legítimo e que sua eleição é um ato de justiça desfeita.

81. **Lula - Segundo Tópico - Sumário 2:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

82. **Lula - Segundo Tópico - Sumário 2:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

83. **Lula - Segundo Tópico - Sumário 2:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

84. **Lula - Segundo Tópico - Sumário 2:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

85. **Lula - Segundo Tópico - Sumário 2:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

86. **Lula - Segundo Tópico - Sumário 2:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Lula - Segundo Tópico - Sumário 3:**

O assunto da presidência do Brasil e do ex-presidiário Jair Bolsonaro é um tema polêmico e controversa que envolve a justiça, a corrupção, e a liderança do país. Muitos cidadãos expressaram sua preocupação e descontentamento com a escolha de um ex-presidiário como presidente, citando a necessidade de uma liderança ética e responsável. A presidência de Bolsonaro tem sido marcada por controvérsias, scandals, e acusações de corrupção, o que aumenta as suspeitas de abuso de poder e desrespeito às instituições. Alguns dos principais pontos mencionados nos textos incluem a preocupação com a escolha de um ex-presidiário como presidente, a falta de experiência e habilidades políticas de Bolsonaro, o skepticismo em relação às promessas de Bolsonaro durante sua campanha presidencial, a possibilidade de Bolsonaro commeter erros graves durante seu mandato, e a opinião de que a escolha de Bolsonaro como presidente é uma vergonha para o Brasil. A linguagem usada nos textos é emocional e crítica, reflectindo a frustração e a desconfiança da população em relação ao político e à justiça no país.

87. **Lula - Segundo Tópico - Sumário 3:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

88. **Lula - Segundo Tópico - Sumário 3:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

89. **Lula - Segundo Tópico - Sumário 3:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

90. **Lula - Segundo Tópico - Sumário 3:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

91. **Lula - Segundo Tópico - Sumário 3:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

92. **Lula - Segundo Tópico - Sumário 3:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

**Lula - Segundo Tópico - Sumário 4:**
"A sociedade brasileira está enfrentando uma crise devido à eleição de um ex-presidiário para o cargo mais alto do país. O ex-presidiário é visto como um símbolo da corrupção e da falta de justiça, e seu apoio é visto como uma afronta à democracia. Apesar dos esforços dos cidadãos legais para defender a lei e a ordem, o ex-presidiário continua a ser um símbolo de poder e corrupção no país."

93. **Lula - Segundo Tópico - Sumário 4:**

Para o quesito **Fluência/Naturalidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente natural |

94. **Lula - Segundo Tópico - Sumário 4:**

Para o quesito **Uso correto da linguagem**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Muit | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Sem erros |

95. **Lula - Segundo Tópico - Sumário 4:**

Para o quesito **Informatividade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente informativo |

96. **Lula - Segundo Tópico - Sumário 4:**

Para o quesito **Clareza/Legibilidade**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente claro |

97. **Lula - Segundo Tópico - Sumário 4:**

Para o quesito **Redundância**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Nada redundante |

98. **Lula - Segundo Tópico - Sumário 4:**

Para o quesito **Adequação ao Tópico**, como você avalia o sumário acima?

*Marcar apenas uma oval.*

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | |
|---|---|---|---|---|---|---|---|---|
| Tota | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | Totalmente adequado |

## Perfil do Participante

Nesta seção, você responderá questões voltadas a seu próprio perfil. Serão dados relevantes e confidenciais.

99. **Qual a sua idade? Responder somente com o número (ex: 30)**

_____

100. **Qual é sua cor/etnia?**

*Marcar apenas uma oval.*

- ◯ Branca
- ◯ Parda ou Preta
- ◯ Amarela
- ◯ Indígena
- ◯ Outro: _____

101. **Qual é sua identidade de gênero?**

*Marcar apenas uma oval.*

- ◯ Mulher cis
- ◯ Homem cis
- ◯ Mulher trans
- ◯ Homem trans
- ◯ Não binário
- ◯ Outro: _____

102. **Qual é seu grau de escolaridade?**

*Marcar apenas uma oval.*

- ◯ Ensino fundamental incompleto
- ◯ Ensino fundamental completo
- ◯ Ensino médio completo
- ◯ Graduação completa
- ◯ Mestrado completo
- ◯ Doutorado completo
- ◯ Pós-doutorado completo
- ◯ Outro: _____

103. **Em qual grande área do conhecimento melhor se encaixa sua formação no ensino superior?**

*Marcar apenas uma oval.*

- ◯ Não sou formado
- ◯ Biológicas e da saúde
- ◯ Exatas e tecnológicas
- ◯ Humanas e de gestão
- ◯ Outro: _____