

Análise da Performance de Modelos de LLM em Sumarização de Artigos Científicos Escritos em Português Brasileiro Após Processo de Fine-Tuning

Komi Parfait Moussougan¹, Giancarlo D. Salton¹

¹Universidade Federal da Fronteira Sul
Campus Chapecó

komi.moussougan@estudante.uffs.edu.br, gian@uffs.edu.br

Resumo. Este trabalho avalia modelos de linguagem de grande porte na sumarização de textos científicos em português brasileiro, com foco no impacto do fine-tuning. Experimentos com PTT5 e Flan-T5 (encoder-decoder) e mBERT e RoBERTa (encoder-only) foram realizados em 100 artigos completos (cinco áreas, ≥ 2020). Para modelos gerativos, aplicou-se chunking (512 tokens com sobreposição) e agregação map-reduce; para os encoder-only, sumarização extrativa por similaridade de cosseno ao resumo do autor. O ajuste fino (LoRA) foi avaliado com ROUGE, BERTScore e análise qualitativa. Os resultados indicam que o alinhamento do pré-treinamento ao português é decisivo: o PTT5 degradou após o ajuste e o Flan-T5 apresentou instabilidade/possível colapso linguístico; em contraste, o fine-tuning melhorou os modelos extrativos, sobretudo o mBERT, sugerindo seleção mais relevante e menos redundância.

Abstract. This paper evaluates large language models (LLMs) for summarizing Brazilian Portuguese scientific texts, focusing on fine-tuning effects. We compare PTT5 and Flan-T5 (encoder-decoder) and mBERT and RoBERTa (encoder-only) on 100 full papers (five fields, ≥ 2020). Generative models use chunking (512 tokens, overlap) with map-reduce aggregation, while encoder-only models perform extractive selection via cosine similarity to the authors' abstracts. Fine-tuning (LoRA) is assessed with ROUGE, BERTScore, and qualitative analysis. Results show Portuguese-aligned pretraining is crucial: PTT5 degrades after fine-tuning and Flan-T5 becomes unstable/consistent with language collapse; conversely, fine-tuning improves extractive models, especially mBERT, indicating more relevant and less redundant sentence selection.

1. Introdução

A produção de informações textuais cresceu de forma expressiva nas últimas décadas, alterando profundamente a maneira como profissionais, instituições e sistemas digitais lidam com o conhecimento. Estudos recentes, como o de [Bornmann et al. 2021], evidenciam que o volume de publicações científicas continua a aumentar em ritmo acelerado, seguindo um padrão de expansão exponencial. De forma complementar, [Fire and Guestrin 2019] destacam o fenômeno da chamada “inflação científica”, caracterizado tanto pelo crescimento contínuo da quantidade de trabalhos publicados quanto pela crescente dificuldade de analisá-los de maneira eficiente. Nesse contexto, torna-se

cada vez mais desafiador acessar, interpretar e sintetizar grandes volumes de informação, especialmente quando é necessário identificar rapidamente as ideias centrais presentes em textos extensos.

Nesse contexto, técnicas de Processamento de Linguagem Natural (PLN) adquiriram papel fundamental na análise e organização de grandes volumes de informação [Jurafsky and Martin 2023]. Entre essas técnicas, destaca-se a sumarização automática de textos, cujo objetivo é condensar documentos preservando sua essência informacional. Tradicionalmente, essa tarefa foi realizada por métodos estatísticos ou baseados em regras linguísticas, predominantemente pela abordagem extrativa [Nenkova and McKeown 2012]. Embora amplamente utilizados, tais métodos apresentam limitações importantes, como falta de fluidez textual, dependência da estrutura original e incapacidade de parafrasear ideias complexas, conforme discutido por [Gambhir and Gupta 2017].

Com o avanço dos modelos baseados na arquitetura Transformer, o panorama da sumarização mudou significativamente. Modelos modernos de aprendizado profundo passaram a permitir a geração de resumos mais coesos e naturais, especialmente por meio da abordagem abstrativa. Trabalhos como PEGASUS, proposto por [Zhang et al. 2020a], demonstraram que pré-treinamentos especializados podem melhorar drasticamente a qualidade dos resumos, especialmente em textos longos ou complexos. Esses avanços tornaram os Large Language Models (LLMs) parte central da pesquisa atual em sumarização.

Os LLMs, por sua vez, representam uma evolução direta dos Transformers, ampliando sua capacidade por meio de treinamento em grandes corpora textuais. Como discutido por [Bommasani et al. 2022], esses modelos chamados de foundation models apresentam forte generalização e podem ser adaptados a diferentes tarefas por técnicas como o fine-tuning. Esse avanço possibilitou a produção de resumos mais precisos, coerentes e sensíveis ao contexto semântico. Além disso, a difusão dos LLMs transformou significativamente a forma como sistemas computacionais compreendem e geram textos. No entanto, o desempenho desses modelos depende não apenas de sua arquitetura, mas também de estratégias de adaptação, como o fine-tuning. Avaliar o impacto desse processo em textos escritos em português brasileiro é fundamental para entender como os modelos podem ser otimizados para tarefas específicas, ampliando sua precisão e utilidade prática.

No contexto da língua portuguesa, estudos recentes têm buscado adaptar modelos originalmente treinados para o inglês às particularidades do português brasileiro. Entre esses trabalhos, destaca-se o PTT5, desenvolvido por [Carmo et al. 2020], que adapta o modelo T5 ao português e apresenta resultados promissores em tarefas de sumarização, incluindo datasets como CSTNews e SciSumPT. Tais iniciativas respondem à carência histórica de modelos robustos para o português e evidenciam a relevância de investigar ajustes específicos, como fine-tuning, cuja eficácia foi destacada por [Howard and Ruder 2018] ao demonstrar que ajustes direcionados ao domínio podem melhorar substancialmente o desempenho de modelos de linguagem em tarefas específicas.

Diante desse cenário, este trabalho tem como objetivo analisar a performance de

LLMs aplicados à sumarização de textos em português brasileiro, avaliando especificamente o impacto do fine-tuning na qualidade dos resumos produzidos. Busca-se comparar abordagens tradicionais e modernas, identificar limitações, discutir avanços recentes e contribuir para o desenvolvimento de sistemas automáticos mais eficientes e precisos para o processamento de textos em português.

O restante do trabalho está organizado da seguinte forma: a Seção 2 apresenta os conceitos fundamentais de sumarização automática; a Seção 3.1 descreve os principais modelos de linguagem e arquiteturas relevantes; a Seção 4 discute os trabalhos relacionados; a Seção 5 detalha a metodologia adotada; a Seção 6 descreve o corpus e os experimentos realizados; a Seção 7 apresenta e discute os resultados; e, por fim, a Seção 8 reúne as conclusões e perspectivas de trabalhos futuros.

2. Sumarização de Textos

A sumarização de textos é uma tarefa central no campo do Processamento de Linguagem Natural (PLN), cujo objetivo consiste em reduzir um ou mais documentos a uma versão concisa que preserve suas informações essenciais, o contexto global e a coerência semântica. Em um cenário marcado pelo crescimento exponencial de dados textuais (artigos científicos, notícias, publicações acadêmicas e conteúdo digital), a sumarização automática torna-se uma ferramenta fundamental para otimizar o consumo e a análise de grandes volumes de informação.

Historicamente, os primeiros métodos de sumarização eram estritamente estatísticos, fundamentados em características superficiais do texto, como frequência de palavras, posição das sentenças ou presença de termos chave. Embora eficientes para documentos simples, esses sistemas tinham dificuldades em capturar nuances semânticas mais profundas. A evolução das técnicas de PLN, especialmente após o surgimento das redes neurais profundas, ampliou drasticamente o escopo e a qualidade dos sistemas de sumarização.

A sumarização automática pode ser dividida em duas grandes abordagens: sumarização extrativa; e sumarização abstrativa. A abordagem extrativa consiste em selecionar, literalmente, trechos do texto original (geralmente sentenças completas) considerados mais relevantes. Esses trechos são então combinados para compor o resumo final. Essa seleção pode ser feita por meio de métricas estatísticas, análise da posição das sentenças, modelos baseados em grafos ou representações vetoriais. Por não gerar novos enunciados, mas sim recortar fragmentos da entrada, a abordagem extrativa tende a preservar a fidelidade factual; porém, pode apresentar falta de fluidez, repetições, desconexões, ausência de parafraseamento e incapacidade de reescrever conceitos complexos.

Em contraste, a abordagem abstrativa busca gerar novas sentenças, assim como um ser humano faria ao compreender o conteúdo original. O sistema interpreta o texto, identifica suas ideias centrais e produz uma reformulação condensada, frequentemente utilizando vocabulário e estruturas sintáticas diferentes das presentes no documento. Esse processo exige capacidades mais avançadas, como modelagem semântica profunda, resolução de correferências, abstração, reescrita e geração de linguagem natural coerente e fluida.

2.1. Uso de LLMs na Sumarização Automática

A arquitetura Transformer [Vaswani et al. 2017] e o mecanismo de atenção ampliaram a capacidade de capturar dependências semânticas em sequências longas, permitindo modelar relações complexas entre palavras e, conseqüentemente, gerar resumos mais naturais e informativos.

Em língua portuguesa, destaca-se o PTT5 [Carmo et al. 2020], ajustado a dados do português brasileiro, que apresentou ganhos na qualidade de geração, sobretudo em domínios como notícias e artigos científicos. O modelo formula a sumarização como uma conversão *text-to-text*, em que o documento é codificado e o resumo é produzido a partir de representações internas altamente contextualizadas.

Treinados em vastos corpora multilíngues, os LLMs elevaram a qualidade da sumarização ao produzir saídas mais coerentes e fluidas, com parafraseamento mais avançado, adaptação a diferentes estilos e comprimentos, maior fidelidade semântica e menor incidência de repetições e inconsistências.

Nesse cenário, trabalhos associados à métrica ROUGE [Lin 2004] e a arquiteturas como o BART [Lewis et al. 2021] evidenciam como a combinação entre modelos modernos e critérios de avaliação mais robustos contribuiu para aprimorar a qualidade dos resumos. Além disso, o *fine-tuning* em corpora específicos (por exemplo, textos científicos) permite alinhar vocabulário e estilo às necessidades do domínio.

3. Modelos de Linguagem

Modelos de linguagem são sistemas de aprendizado computacional que estimam a probabilidade de palavras e sentenças em contexto, viabilizando tarefas de PLN, como tradução automática, sumarização e detecção de sentimentos. As primeiras abordagens, baseadas em N-gramas, capturavam dependências locais; posteriormente, RNNs e LSTMs passaram a modelar relações de longo alcance, embora ainda com limitações para contextos muito extensos e para a paralelização do treinamento.

Para superar essas restrições, mecanismos de atenção passaram a ser incorporados, permitindo que o modelo recupere, de forma dinâmica, informações relevantes de trechos anteriores. O estudo de [Salton 2017], por exemplo, propõe um módulo de atenção para RNN-LM com o objetivo de melhorar o processamento de expressões idiomáticas e dependências de longa distância.

Com a consolidação desses conceitos, surgiram LLMs como BERT [Devlin et al. 2019], GPT-3 [Brown et al. 2020] e T5 [Raffel et al. 2019], treinados em grandes corpora e com bilhões de parâmetros. Essas arquiteturas utilizam representações distribuídas de alta dimensionalidade para capturar nuances semânticas e contextuais, mostrando desempenho superior na sumarização ao compreender relações profundas e reformular o conteúdo de modo coerente e fiel. Modelos *encoder-decoder*, como BART [Lewis et al. 2021] e T5, exemplificam essa abordagem ao codificar o texto-fonte e gerar um resumo informativo e bem estruturado.

3.1. Modelos de Linguagem de Grande Porte (LLMs)

Os LLMs (*Large Language Models*) são modelos de aprendizado profundo capazes de compreender, gerar e manipular linguagem natural com alto grau de coerência

e contextualização [Bommasani et al. 2022]. Em geral, baseiam-se em Transformers [Vaswani et al. 2017], substituindo a recorrência por atenção para ponderar a relevância de cada token no contexto global, o que favorece o tratamento de dependências de longo alcance e a modelagem de estruturas linguísticas complexas.

Como base de diversas tarefas de PLN, esses modelos podem ser adaptados por *fine-tuning* com quantidades relativamente pequenas de dados rotulados. Na sumarização, formulam o problema como geração condicional: a partir de um documento, produzem uma saída mais curta que preserva informações essenciais, com o apoio de arquiteturas *encoder-decoder* (como BART [Lewis et al. 2021] e T5) para garantir fluência e informatividade.

A literatura tem investigado LLMs em contextos como textos científicos e jornalísticos: [Cohan et al. 2018] propõe um modelo sensível à estrutura discursiva, enquanto [Narayan et al. 2018] analisam ganhos oriundos de pré-treinamento em grandes corpora seguido de *fine-tuning* em sumarização. Em português, destaca-se novamente o PTT5 [Carmo et al. 2020], versão do T5 ajustada para o idioma, com resultados promissores em bases como CSTNews e SciSumPT. Em síntese, os LLMs constituem um marco para o avanço da sumarização automática em diferentes idiomas e domínios, incluindo o português brasileiro.

3.2. Modelos Utilizados no Estudo

3.2.1. PTT5

O modelo PTT5 foi apresentado no artigo “PTT5: Pretraining and validating the T5 model on Brazilian Portuguese data” [Carmo et al. 2020], cujo objetivo central foi adaptar o modelo T5 (Text-to-Text Transfer Transformer) especificamente para o português brasileiro. A motivação surge da constatação de que modelos originalmente pré-treinados em inglês apresentam limitações ao serem aplicados a tarefas em português, especialmente devido às diferenças linguísticas e à escassez de modelos *encoder-decoder* robustos pré-treinados na língua.

O PTT5 foi desenvolvido a partir do T5 original, mas totalmente pré-treinado em um grande corpus monolíngue em português brasileiro, o BrWac, composto por aproximadamente 2,7 bilhões de tokens coletados de páginas web brasileiras. Esse pré-treinamento seguiu o objetivo de denoising do T5, onde trechos do texto são corrompidos e o modelo precisa reconstruí-los, permitindo aprender representações linguísticas amplas e robustas.

De maneira geral, o artigo conclui que pré-treinar o T5 em português brasileiro resulta em melhorias significativas em tarefas variadas da língua, demonstrando a importância de adaptar modelos a idiomas com características próprias. Os resultados mostram que o PTT5 se torna uma alternativa competitiva a modelos monolíngues consolidados, sobretudo em tarefas de geração e transformação de texto, nas quais arquiteturas *encoder-decoder*, como o T5, são naturalmente superiores.

3.2.2. FLAN-T5

O FLAN-T5 é uma versão do T5 aprimorada por meio de instruction-tuning, técnica em que o modelo é ajustado usando um grande conjunto de tarefas formuladas como instruções em linguagem natural (como “resuma”, “explique”, “traduza para X”, etc.). Esse processo foi introduzido no trabalho Scaling Instruction-Finetuned Language Models [Chung et al. 2022b], desenvolvido por pesquisadores do Google.

Esse ajuste adicional torna o FLAN-T5 mais versátil, especialmente em cenários zero-shot e few-shot, nos quais o modelo precisa executar tarefas novas sem muitos exemplos. Embora não seja especializado em português, o FLAN-T5 apresenta desempenho robusto na geração de texto e segue instruções de forma mais confiável do que o T5 original.

No contexto deste estudo, o FLAN-T5 serve como um modelo “moderno” que permite comparar o impacto do instruction-tuning frente a modelos monolíngues e multilíngues tradicionais.

3.2.3. BERT Multilíngue (mBERT)

O BERT Multilingual (mBERT) [Devlin et al. 2019] foi lançado pelo Google como uma extensão do BERT original para 104 idiomas, incluindo o português. Ele utiliza apenas a arquitetura encoder e foi treinado com textos da Wikipedia multilíngue usando Masked Language Modeling (MLM) e Next Sentence Prediction (NSP).

Apesar de não ser treinado especificamente para português, o mBERT tornou-se um modelo de referência para tarefas como classificação, detecção de entidades e outras aplicações que envolvem compreensão textual. Em tarefas de sumarização extrativa, por exemplo, suas representações podem ser usadas para medir a similaridade semântica e selecionar sentenças relevantes.

No presente trabalho, o mBERT representa uma linha de modelos clássicos e amplamente utilizados no PLN, permitindo avaliar como um modelo multilíngue tradicional se comporta frente a alternativas mais recentes ou especializadas.

3.2.4. RoBERTa

O RoBERTa (*Robustly Optimized BERT Pretraining Approach*) é uma variante do BERT que mantém uma arquitetura *encoder-only* e foi otimizada por alterações no pré-treinamento, como um maior volume de dados, ajustes de hiperparâmetros e remoção do objetivo *Next Sentence Prediction*. Por não possuir um decodificador nativo, não é um modelo gerativo; assim, no contexto de sumarização, tende a ser empregado como componente de pipelines extrativos, por exemplo, via representações vetoriais e medidas de similaridade para seleção de sentenças.

4. Trabalhos Relacionados

Um dos estudos mais relevantes no cenário de sumarização de textos na língua portuguesa é o de [Sarmiento and de Oliveira 2024]. A pesquisa compara abordagens ex-

trativas clássicas, técnicas baseadas em programação linear inteira e modelos neurais, incluindo o ajuste fino de modelos pré-treinados (PTT5 e FLAN-T5) e o uso de grandes modelos de linguagem (LLMs), de código aberto e proprietários.

Conduzido em três bases de dados de notícias em português, o estudo demonstrou que modelos ajustados para sumarização abstrativa, como o PTT5, alcançaram desempenho competitivo em comparação com modelos muito maiores, como o GPT-4o. A análise qualitativa indicou que, embora abordagens extrativas pudessem alcançar resultados quantitativos similares, os resumos abstrativos eram mais sucintos e apresentavam maior coesão, coerência e qualidade textual. Assim, embora os LLMs sejam promissores, modelos extrativos ou especializados seguem como alternativas viáveis em cenários de recursos limitados, reforçando a importância da relação custo-benefício.

Apesar desses avanços na sumarização de notícias, a aplicação de LLMs para a sumarização de artigos científicos em português ainda é uma área com lacuna significativa. Pesquisas como [Reis and Araújo 2022] e [Sousa and Lima 2025] têm focado majoritariamente em métodos extrativos, eficientes, mas que pouco exploram o potencial dos modelos abstrativos para criar resumos com alta qualidade textual. Portanto, este trabalho busca contribuir para essa área, explorando a capacidade de LLMs de gerar resumos com qualidade e precisão no domínio acadêmico em português.

O estudo de [Sousa and Lima 2025] investigou a sumarização automática de textos científicos a partir de técnicas de Inteligência Artificial no contexto da Organização do Conhecimento, sob o paradigma da Inteligência Aumentada (IAu), que entende a IA como assistente e enfatiza a supervisão profissional para garantir qualidade, ética e relevância. Os autores desenvolveram um software com spaCy e o modelo `pt_core_news_sm` para processar artigos da BDPA/EMBRAPA por uma pipeline de PLN (extração, identificação de termos, classificação de frases e seleção das mais informativas); na avaliação com 100 artigos, os resumos foram analisados qualitativamente (“adequado”, “difícil de ler” e “inadequado”). Como pontos fortes, destaca-se a viabilidade de modelos leves em computadores pessoais, com processamento local (segurança e privacidade), e a proposta de Inteligência Aumentada como alternativa equilibrada entre automação e supervisão humana; como limitações, apontam-se a exclusão deliberada dos LLMs, a dependência da qualidade do texto-fonte, o desafio de escalabilidade e a ausência de métricas padronizadas e de análises mais aprofundadas sobre vieses éticos e culturais.

Outro estudo na área de sumarização é o de [Nascimento Rodrigues and Conceição 2025], que investigou sumarizadores extrativos na organização do conhecimento científico, avaliando ferramentas para identificar informações-chave em artigos no contexto da língua portuguesa. A partir de levantamento bibliográfico e documental, os autores compararam qualitativamente os resumos e observaram que, embora eficientes para identificar conteúdos importantes, abordagens extrativas apresentam limitações quanto à fluência textual e à coesão; além disso, o estudo não explorou abordagens abstrativas nem modelos de linguagem pré-treinados, mantendo lacunas sobre o potencial de LLMs para a sumarização científica em português. Dessa forma, reforça-se a necessidade de explorar técnicas mais avançadas, como o fine-tuning de LLMs e o prompt engineering adaptado à língua portuguesa, visando gerar resumos abstrativos de maior qualidade, coesão e confiabilidade.

Nessa mesma linha, [Reis and Araújo 2022] desenvolveu uma plataforma web para sumarização automática extrativa de artigos científicos em português do Brasil, implementando dois algoritmos clássicos (Luhn e distância do cosseno) para auxiliar pesquisadores. A metodologia envolveu o processamento dos artigos, a aplicação dos algoritmos e avaliações quantitativas (distância do cosseno e containment), além do monitoramento do tempo de execução. Os resultados indicaram similaridades superiores a 70% usando 10% das frases, com desempenho superior do algoritmo de Luhn. Como pontos fortes, destaca-se a ferramenta prática e acessível e a comparação entre algoritmos; por outro lado, permanecem limitações como a restrição ao português do Brasil, o caráter extrativo e a ausência de avaliação humana, que poderia qualificar melhor a qualidade dos resumos.

5. Metodologia

A Figura 1 sintetiza a metodologia adotada neste trabalho. O processo foi organizado em etapas que contemplam desde a seleção dos modelos de linguagem até a análise dos resultados. Inicialmente, foram definidos os modelos PTT5, FLAN-T5, BERT Multilíngue e RoBERTa e, a partir de um conjunto de textos científicos em português, foram conduzidos experimentos com e sem *fine-tuning*. Em seguida, os modelos foram utilizados para gerar resumos automáticos, que foram avaliados por meio das métricas ROUGE e BERTScore. Por fim, os resultados foram comparados e discutidos a fim de verificar o impacto do *fine-tuning* na qualidade da sumarização.

O planejamento do estudo constituiu a primeira etapa e teve como objetivo orientar as fases subsequentes da pesquisa. Nessa fase, foram definidos o tema, a pergunta de pesquisa e os objetivos do trabalho, assegurando um direcionamento claro e consistente. Também foi elaborado um cronograma de atividades, com prazos para o levantamento de referências, organização do *dataset*, implementação dos experimentos e análise dos resultados.

Em paralelo, foi realizado o levantamento de referências teóricas e artigos científicos relacionados, com foco em sumarização automática, modelos de linguagem de grande porte e abordagens voltadas a textos em português. Esse levantamento permitiu identificar lacunas e oportunidades na literatura e fundamentou a escolha das ferramentas, modelos e métricas empregadas nas etapas seguintes. Ainda no planejamento, foram definidos e aplicados critérios de inclusão e exclusão para compor o *dataset*, considerando: período de publicação (a partir de 2020), disponibilidade do texto completo em português e relevância científica. Com isso, buscou-se garantir um conjunto de dados consistente, adequado e passível de análise automatizada.

A revisão de literatura foi conduzida para contextualizar o estudo, apresentar o estado da arte e identificar lacunas relacionadas à sumarização automática de textos científicos em português. Para isso, os trabalhos analisados foram organizados em categorias: técnicas extrativas, técnicas abstrativas, modelos baseados em LLMs e pesquisas com foco no idioma português. Essa categorização viabilizou uma análise sistemática das abordagens existentes, destacando tendências, limitações e oportunidades de melhoria. Durante a revisão, foram observados os métodos mais utilizados, os tipos de dados empregados, as métricas de avaliação mais recorrentes e desafios específicos ao português.

Com base nessa análise, foram evidenciadas lacunas como a escassez de estudos

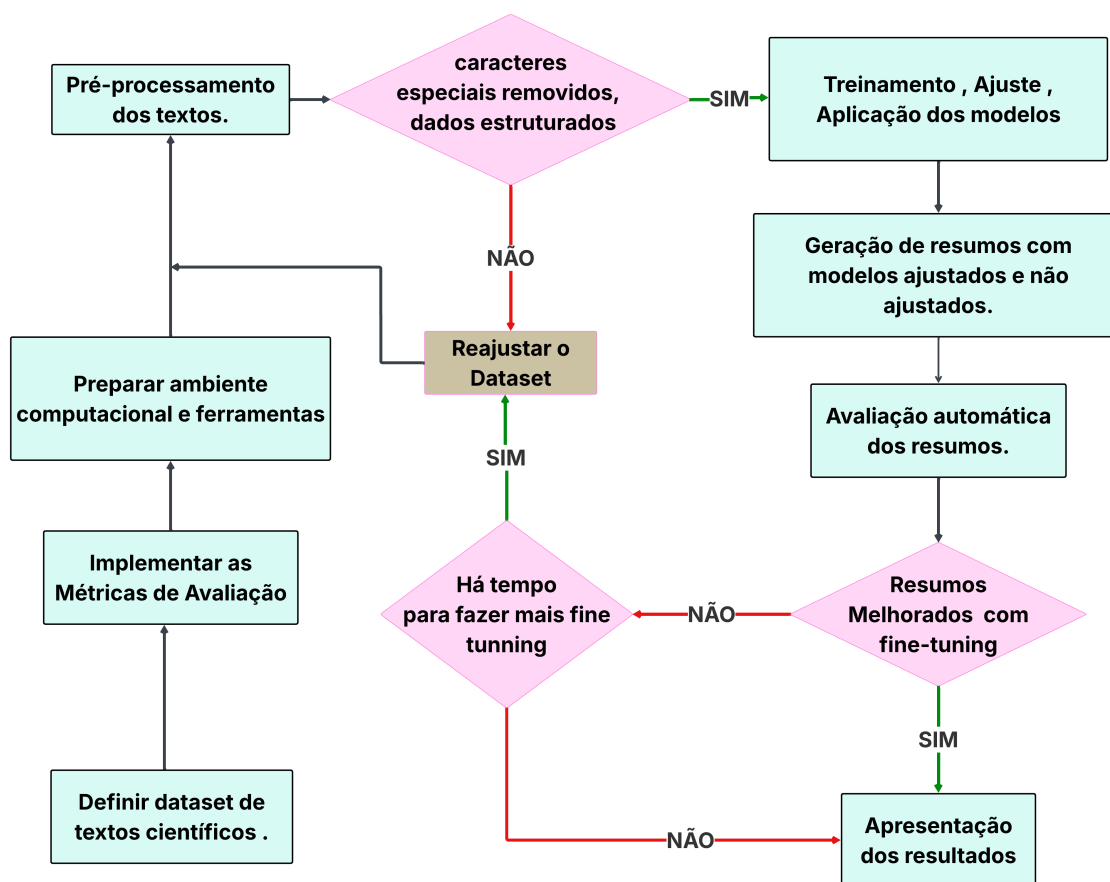


Figure 1. Metodologia de trabalho.

que aplicam LLMs em *datasets* de artigos científicos em português e a necessidade de avaliações mais detalhadas de coerência, coesão e fidelidade nos resumos gerados. Essas lacunas justificam a relevância do presente estudo, que investiga o uso dos modelos citados para gerar resumos de textos científicos em português. Por fim, a revisão de literatura serviu como base para a definição do desenho experimental, influenciando a escolha dos modelos, métricas e estratégias de avaliação.

A metodologia experimental definiu os procedimentos adotados para a geração de resumos de textos científicos em português. Foi definido um *dataset* composto por aproximadamente 100 artigos completos em português, publicados a partir de 2020, contemplando diferentes áreas do conhecimento (Saúde, Tecnologia, Ciências Sociais, Ciências Exatas e Agronomia), a fim de garantir diversidade temática e variações de estilo de escrita. O conjunto de dados foi organizado de forma sistemática para facilitar o pré-processamento e as análises subsequentes. Em seguida, foram selecionados os modelos de linguagem utilizados nos experimentos, considerando suas capacidades de processar textos em português e produzir resumos coerentes. Essa etapa consolidou o material e os procedimentos necessários para a implementação prática, incluindo o *fine-tuning* e a avaliação dos modelos.

A etapa de implementação e experimentos consistiu na execução dos procedimentos planejados para gerar resumos a partir dos artigos do *dataset* e avaliar a performance

dos modelos. Inicialmente, os textos passaram por pré-processamento, incluindo limpeza de caracteres especiais, normalização de formatos e padronização das estruturas textuais, visando garantir dados consistentes e adequados ao treinamento. Na sequência, os modelos foram submetidos a *fine-tuning* com o *dataset* preparado, ajustando seus parâmetros para melhorar a capacidade de produzir resumos coerentes e fiéis ao conteúdo original. Paralelamente, os modelos também foram avaliados em sua forma não ajustada, permitindo comparar diretamente o impacto do *fine-tuning*. Para viabilizar os experimentos, o ambiente computacional foi configurado com Python, PyTorch e a plataforma Hugging Face, assegurando que as etapas de treinamento, *fine-tuning* e avaliação fossem executadas de forma eficiente e reproduzível.

A análise dos resultados correspondeu à etapa final e teve como objetivo avaliar a eficácia dos modelos na geração de resumos de textos científicos em português. Os resumos gerados pelos modelos, nas versões ajustada e não ajustada, foram comparados aos textos de referência por meio das métricas definidas previamente.

As métricas automáticas ROUGE [Lin 2004] e BERTScore [Zhang et al. 2020b] foram empregadas para quantificar o nível de correspondência entre os resumos e os artigos originais, oferecendo uma medida objetiva de fidelidade e completude. A comparação entre as versões ajustadas e não ajustadas permitiu identificar ganhos atribuíveis ao *fine-tuning* e discutir seu impacto sobre a capacidade de síntese dos modelos.

6. Experimentos

6.1. Coleta e Tratamento de Dados

O corpus deste trabalho é composto por 100 artigos científicos distribuídos equitativamente entre cinco áreas distintas do conhecimento: Saúde, Tecnologia, Ciências Sociais, Ciências Exatas e Agronomia. A coleta dos documentos foi realizada por meio de plataformas de busca acadêmica, especificamente o SciELO e o ResearchGate. Como critério de seleção, foram considerados apenas artigos publicados a partir de 2020, escritos em língua portuguesa e publicados no Brasil.

Após o download, os dados foram estruturados em um conjunto de dados composto pelas seguintes colunas: *id*, *título*, *resumo*, *conteúdo* e *area*. Durante a montagem desse arquivo, o resumo original (escrito pelo autor) foi rigorosamente separado do conteúdo principal do texto. Essa separação é uma etapa metodológica indispensável para evitar o risco de vazamento de dados (*data leakage*) durante as fases posteriores de treinamento (*fine-tuning*) e inferência dos modelos.

Durante essa etapa de separação textual, foi aplicado um processo de filtragem e limpeza para remover metadados e ruídos que não agregavam valor semântico ao experimento. Nessa etapa, eliminaram-se elementos textuais secundários, tais como informações de linhas editoriais, links externos, referências bibliográficas, nomes de colaboradores secundários, índices, cabeçalhos, rodapés e palavras-chave. Como resultado final desta etapa de engenharia de dados, consolidou-se um dataset limpo com 100 entradas válidas, composto estritamente pelas informações essenciais para a execução do trabalho.

6.2. Geração dos resumos com os modelos bases

Nesta etapa, os experimentos foram conduzidos com dois modelos da família T5 (PTT5 e FLAN-T5) e duas arquiteturas *Encoder-Only* (mBERT e RoBERTa). No caso dos modelos T5, a janela de contexto padrão é limitada a 512 *tokens* de entrada. Para contornar essa restrição, implementou-se uma técnica de fatiamento do texto (*chunking*) em blocos menores com tamanho máximo de 512 *tokens*, aplicando-se uma sobreposição (*overlap*) de 100 *tokens* entre os blocos adjacentes.

A aplicação dessa estratégia de *chunking* alterou temporariamente a estrutura do conjunto de dados: as 100 entradas originais foram expandidas para 1.857 registros. Em seguida, adotou-se uma abordagem análoga ao mapeamento e redução (*map-reduce*): gerou-se um resumo parcial para cada fragmento e, posteriormente, esses mini-resumos foram concatenados de forma ordenada para constituir o resumo definitivo de cada documento, agregando-se novamente os dados para reconstituir o *dataset* original de 100 linhas.

Diferenciando-se das arquiteturas PTT5 e FLAN-T5, que pertencem à família *Encoder-Decoder* (arquitetura do T5), os modelos mBERT (*Multilingual BERT*) e RoBERTa estruturam-se sob uma abordagem *Encoder-Only* (arquitetura do BERT), o que implica a ausência de um módulo decodificador nativo capaz de gerar sequências de texto palavra por palavra (resumo). Neste trabalho, adotou-se a estratégia de sumarização extrativa baseada na similaridade de cosseno: extraíram-se os vetores de representação semântica de cada frase do conteúdo principal e o *embedding* correspondente ao resumo original do autor e, na sequência, calculou-se a proximidade angular entre eles.

Esse procedimento permitiu pontuar e ordenar as sentenças do documento. Para cada documento, selecionaram-se as 7 frases com as maiores pontuações de similaridade, reorganizando-as em ordem cronológica de aparição para preservar a coesão textual e constituir o resumo extrativo final. Por fim, aplicou-se a avaliação automatizada por meio das métricas ROUGE e BERTScore, cujos resultados quantitativos serão discutidos na seção dedicada aos resultados deste trabalho.

6.2.1. PTT5

O PTT5 consiste em uma versão adaptada para a língua portuguesa do modelo de linguagem T5 (*Text-to-Text Transfer Transformer*). Para a geração dos resumos nesta etapa de modelos bases, utilizou-se o *checkpoint* unicamp-dl/ptt5-base-portuguese-vocab, desenvolvido e implementado por [Paiola et al. 2022]. A inferência foi executada seguindo o pipeline descrito anteriormente, respeitando a janela de contexto de 512 *tokens* por meio do *chunking* com sobreposição.

6.2.2. FLAN-T5

Diferente do PTT5, cujo foco primário reside na especialização e alinhamento ao idioma português, o FLAN-T5 (*Finetuned Language Net T5*) introduz uma abordagem baseada no aprendizado por instruções (*instruction-following tuning*). Embora ambos os

modelos compartilhem a arquitetura *Encoder-Decoder* do T5 original desenvolvido pela Google, seus processos de pré-treinamento e objetivos de convergência são distintos. No caso desse trabalho foi usado o modelo `google/flan-t5-base` implementado por [Chung et al. 2022a].

Para a execução do FLAN-T5, manteve-se rigorosamente a mesma estratégia de fatiamento em blocos (*chunking*) com sobreposição utilizada no PTT5. Contudo, por tratar-se de um modelo ajustado por instruções, o pipeline de inferência exigiu a estruturação de um *prompt*. Dessa forma, foi adicionado o seguinte prompt no início de cada *chunk*: "*Resume this text in portuguese:*". Essa inserção do prompt no início foi indispensável para guiar o comportamento do modelo, induzindo-o a compreender a natureza da tarefa de sumarização e a obrigatoriedade de gerar o texto de saída no idioma alvo.

6.3. Geração de Resumos Após *Fine Tuning* dos modelos

Para o *fine-tuning* dos modelos, o ambiente experimental foi estruturado na plataforma Google Colab, utilizando uma GPU NVIDIA T4. Previamente ao treinamento, foram separados 50% do dataset para o *fine-tuning*, enquanto a outra metade foi reservada para a etapa posterior de inferência e geração de resumos com os modelos ajustados. O conjunto de dados de *fine-tuning* foi particionado de forma a destinar 80% das entradas para o conjunto de treino e 20% para o de teste.

Como os modelos-base possuem uma estrutura robusta composta por milhões de parâmetros, realizar o ajuste fino completo (*Full Fine-Tuning*) exigiria uma infraestrutura de *hardware* massiva e de alto custo computacional, além de expor os modelos ao risco de esquecimento catastrófico (*catastrophic forgetting*). Para contornar essa limitação e viabilizar os experimentos, integrou-se a técnica de Adaptação de Baixo Posto, ou LoRA (*Low-Rank Adaptation*), que congela os pesos do modelo pré-treinado e atualiza apenas adaptadores de baixo posto inseridos nas camadas de atenção. Para cada modelo, foram avaliadas diferentes porcentagens de parâmetros treináveis no processo de (*Fine-Tuning*) com LoRA. Ao final dos experimentos, foi selecionada a configuração que apresentou o melhor desempenho para cada modelo, o que explica a diferença na quantidade de parâmetros treináveis utilizada entre os modelos. Essa abordagem reduz drasticamente o volume de parâmetros treináveis, mantendo a viabilidade computacional do processo sem perda significativa de generalização.

Durante a execução do ciclo de treinamento, as métricas de convergência foram monitoradas e registradas em tempo real por meio do MLflow, permitindo acompanhar a evolução das funções de perda (*loss*), avaliar o comportamento do modelo a cada época e realizar ajustes controlados de hiperparâmetros. Ao término do processo, os pesos otimizados do adaptador LoRA foram consolidados, gerando modelos prontos para a inferência de resumos de artigos científicos na língua portuguesa.

6.3.1. PTT5 - Fine-Tuning

Para a etapa de ajuste fino do PTT5, utilizou-se o modelo base `unicamp-dl/ptt5-base-portuguese-vocab` juntamente com o seu respectivo

tokenizador, ambos carregados a partir do ecossistema Hugging Face, mantendo-se a mesma estratégia de fatiamento em blocos (*chunking*) adotada na etapa de inferência.

Por se basear na arquitetura *Text-to-Text*, onde todas as tarefas utilizam o mesmo formato de entrada e saída, inseriu-se o prefixo "*summarize:* " no início de cada *chunk*, atuando como um gatilho de instrução para a tarefa de sumarização.

O modelo PTT5 base possui 224.673.024 parâmetros em sua configuração original. Com a integração da técnica LoRA, esse volume de parâmetros treináveis foi reduzido para 1.769.472, correspondendo a menos de 1% do total do modelo.

6.3.2. Flan-T5

No *fine-tuning* do FLAN-T5, o pipeline exigiu uma engenharia de *prompts* específica devido à natureza de aprendizado por instruções do modelo. Para cada fragmento de texto, inseriu-se o comando "*Resume this text in português:*" no início e a string "*Resumo:*" ao final, delimitando a tarefa e orientando a geração no idioma correto.

Com a aplicação da técnica LoRA sobre essa arquitetura, o volume de parâmetros treináveis foi otimizado de 254.360.832 para 6.782.976, cerca de 2,66% do total.

6.3.3. mBERT e RoBERTa

Para o *fine-tuning* das arquiteturas *Encoder-only* (mBERT e RoBERTa), o pipeline de execução foi o mesmo, alterando-se apenas os modelos e seus respectivos tokenizadores importados do ecossistema Hugging Face.

Previamente ao treinamento, o conjunto de dados foi reestruturado para adequar os artigos à tarefa de classificação. Utilizou-se a similaridade de cosseno para comparar cada sentença do corpo do artigo com o resumo do autor, classificando se a sentença deveria ser incluída ou descartada do resumo final. As sentenças foram mapeadas de forma binária: atribuiu-se o rótulo 1 para as 7 frases de maior proximidade semântica e o rótulo 0 para as demais. Para garantir rastreabilidade, cada sentença manteve o vínculo com seu documento de origem por meio do identificador do artigo, resultando em um novo *dataset* com as colunas: *article_id*, *title*, *original_abstract*, *sentence* e *label*.

Aplicando LoRA ao mBERT, que possui originalmente 182.575.108 parâmetros, o volume foi reduzido para 4,720,130 parâmetros treináveis (aproximadamente 2,53%). Para o RoBERTa, cuja arquitetura inicial compreende 279.227.140 parâmetros, o volume foi reduzido para 1.181.954 parâmetros (0,42%).

7. Resultados e Discussão

7.1. PTT5

No modelo base, os resumos gerados para as diferentes áreas apresentaram boa qualidade de escrita: uso adequado do português, coerência textual e preservação de citações bibliográficas relevantes. Observou-se também a capacidade de reconhecer marcadores estruturais das seções originais (por exemplo, introduções e revisões históricas), comportamento consistente com o pré-treinamento do modelo em língua portuguesa.

Ainda assim, a análise evidenciou limites da estratégia de fatiamento (*chunking*) e das restrições de *tokens*. Ao resumir trechos isolados, sem visão global do artigo, o modelo tende a supervalorizar informações secundárias (contextos históricos e detalhes metodológicos) em detrimento das ideias centrais, argumentos e conclusões. Além disso, ocorreram alucinações e truncamentos: no primeiro caso, surgiram referências a figuras/tabelas inexistentes no resumo (p.ex., “conforme descrito na Tabela 1”); no segundo, o texto era interrompido antes da conclusão, em função da janela máxima de geração do decodificador antes da emissão do *token* `<eos>`.

No *fine tuning*, o monitoramento foi conduzido via MLflow, com registro de métricas por época. A Tabela 1 sintetiza os resultados: ROUGE relativamente baixo (ROUGE-1=14,93; ROUGE-2=3,74), em contraste com BERTScore F1=65,92.

Table 1. Métricas do modelo PTT5 coletadas durante o fine tuning

ROUGE (%)		BERTScore (%)	
Métrica	Valor	Métrica	Valor
ROUGE-1	14,9313	BERTScore F1	65,9219
ROUGE-2	3,7408	Precision	-
ROUGE-L	10,9367	Recall	-

Após o *fine tuning*, o modelo ajustado foi utilizado para gerar novos resumos, cujas métricas consolidadas estão na Tabela 2.

Table 2. Métricas Consolidadas de Avaliação: PTT5 Base vs. PTT5 Ajustado

PTT5 - Modelo Base (%)				PTT5 Fine-Tuning (%)			
ROUGE		BERTScore		ROUGE		BERTScore	
ROUGE-1	37,14	BERTScore F1	68,41	ROUGE-1	11,97	BERTScore F1	66,29
ROUGE-2	11,14	Precision	66,64	ROUGE-2	5,42	Precision	66,55
ROUGE-L	17,84	Recall	70,30	ROUGE-L	7,42	Recall	66,10

A comparação indica que o PTT5 Base obteve maior sobreposição lexical (ROUGE-1=37,14; ROUGE-2=11,14), enquanto o *fine-tuning* reduziu os n-gramas do ROUGE (ROUGE-1=11,97; ROUGE-2=5,42) e também levou a uma queda leve no BERTScore (de 68,41 para 66,29), com *Precision* de 66,55% e *Recall* de 66,10%.

As curvas de perda (*loss*) em treino e validação apresentaram convergência descendente (Figura 2, painéis (a) e (b)), sugerindo generalização e ausência de sobreajuste (*overfitting*). Ainda que o modelo ajustado utilize estrutura sintática e vocabulário distintos dos resumos humanos, ele preserva o núcleo conceitual e o contexto dos artigos com fidelidade aceitável.

7.2. Flan-T5

Nos resumos gerados pelo *Flan-T5 Base*, observou-se dificuldade em lidar com abstrações teóricas densas em português. Em alguns casos, o decodificador entrou em *loops* redundantes, reformulando a mesma sentença sucessivamente e produzindo termos incomuns, como “burocrânio” e “agência crática”. Em outros exemplos, o modelo respeitou o gênero textual e produziu um parágrafo conciso, porém também houve contaminação por outras

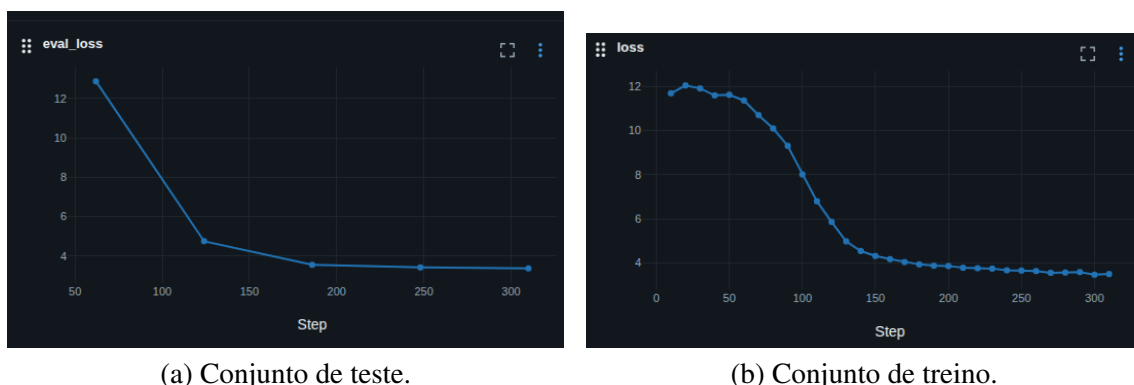


Figure 2. Curvas de convergência da perda (*loss*) do PTT5 nos conjuntos de teste e de treino.

línguas românicas do pré-treinamento, com alternância para espanhol (“información”) e desinências francesas (“Técnicó-Cientfique”).

Após o *fine-tuning*, verificou-se degradação acentuada das saídas e métricas substancialmente baixas, compatíveis com a hipótese de colapso linguístico. A Tabela 3 apresenta as métricas coletadas durante o ajuste.

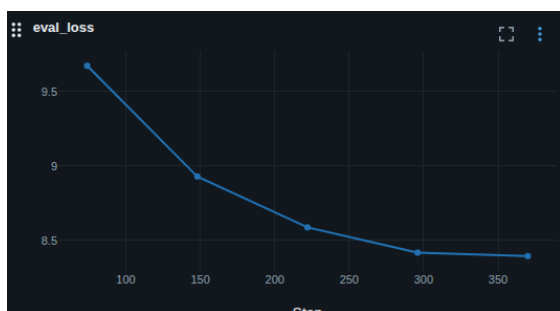
Table 3. Métricas do modelo Flan-T5 coletado durante o fine tuning

ROUGE (%)		BERTScore(%)	
ROUGE-1	7,2532	BERTScore F1	50,2024
ROUGE-2	0,1023	Precision	-
ROUGE-L	6,3774	Recall	-

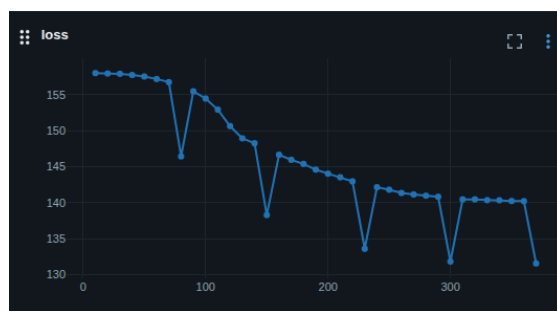
A perda de validação (*eval loss*) atingiu 8,3945, evidenciando dificuldade de convergência e de adaptação de pesos originalmente ancorados no inglês à distribuição do português. Isso se refletiu no ROUGE-1=7,25%, ROUGE-2=0,10% e ROUGE-L=6,37%, com BERTScore F1=50,20%, indicando baixa sobreposição e comprometimento semântico.

Embora as curvas de *loss* em validação sugiram aprendizado (Figura 3, painéis (a) e (b)), a avaliação qualitativa indicou colapso linguístico, com cadeias pouco inteligíveis, repetições silábicas desconexas e termos incompletos. Uma das principais hipóteses para esse comportamento está relacionada à incompatibilidade entre o vocabulário utilizado pelo Flan-T5 e a língua portuguesa. Como o modelo não foi originalmente pré-treinado com foco em português, muitos termos científicos são segmentados em subpalavras pouco representativas, dificultando a construção de representações semânticas estáveis durante o ajuste fino. Esse problema torna-se ainda mais evidente em domínios especializados, nos quais há elevada frequência de terminologia técnica e baixa redundância lexical.

Atribui-se esse desempenho principalmente à incompatibilidade de vocabulário/tokenização do Flan-T5, não pré-treinado em português. Sob a pressão do *fine tuning* via LoRA para codificar conceitos científicos densos, os adaptadores mostraram-se insuficientes para sustentar a sintaxe, e o *chunking* hierárquico exacerbou a fragmentação dos mecanismos de atenção. As métricas consolidadas estão na Tabela 4.



(a) Conjunto de teste.



(b) Conjunto de treino.

Figure 3. Curvas de convergência da perda (*loss*) do *Flan-T5* nos conjuntos de teste e de treino.

Table 4. Métricas Consolidadas de Avaliação: *Flan-T5* Base vs. *Flan-T5* Ajustado

Flan-T5 - Modelo Base (%)				Flan-T5 - Fine-Tuning (%)			
ROUGE		BERTScore		ROUGE		BERTScore	
ROUGE-1	23,87	BERTScore F1	63,48	ROUGE-1	26,26	BERTScore F1	53,63
ROUGE-2	3,12	Precision	63,08	ROUGE-2	1,48	Precision	53,45
ROUGE-L	12,73	Recall	63,91	ROUGE-L	15,07	Recall	53,82

7.3. mBERT

Na sumarização extrativa com modelos *Encoder-Only*, o *mBERT Base* apresentou limitações como alucinações, redundâncias e alternância linguística. Alguns resumos exibiram citações diretas de autores externos ao texto-fonte, além de falhas de completude (término abrupto e truncado). Também houve variações ortográficas do português europeu (p.ex., “perspetiva”) e repetições cíclicas, sugerindo saturação dos mecanismos de atenção ao tentar preencher o limite de *tokens*. Notou-se, ainda, comportamento fortemente extrativo, com cópia integral de trechos do resumo original e alternância inesperada para o inglês.

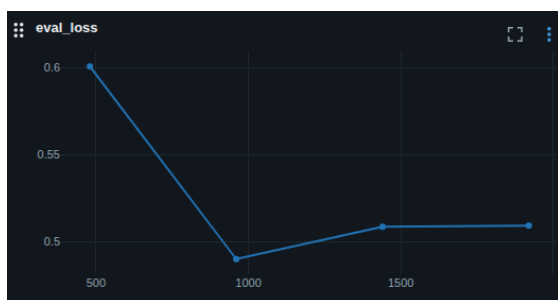
Com *fine-tuning* via LoRA, o monitoramento no MLflow mostrou dinâmica distinta de arquiteturas *Encoder-Decoder*. A Tabela 5 resume as métricas de validação.

Table 5. Métricas Avaliadas no Conjunto de Validação durante o Fine-Tuning do mBERT

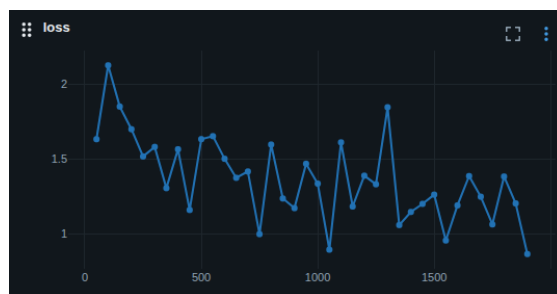
Métricas de Erro e Desempenho (%)		Métricas de Classificação (%)	
Acurácia	96,96	F1-Score	51,9297
loss	0,5094	Precision	73,5198
-	-	Recall	51,3837

As curvas de *loss* oscilaram de forma ruidosa ao longo das épocas, embora com tendência descendente (Figura 4, painéis (a) e (b)). Na fase final, observou-se início de sobreajuste (*overfitting*), sugerindo necessidade de controle estrito dos *steps*.

O efeito do ajuste aparece nas métricas consolidadas (Tabela 6): o mBERT Ajustado mitigou ciclos repetitivos sem degradar drasticamente a estrutura lexical, com ROUGE-1 de 40,30% para 38,74%, ROUGE-2 de 12,69% para 14,86% e BERTScore F1 de 68,97% para 70,04% (*Recall*=71,48%). Esses resultados indicam refinamento dos



(a) Conjunto de teste.



(b) Conjunto de treino.

Figure 4. Curvas de convergência da perda (*loss*) do mBERT nos conjuntos de teste e de treino.

mecanismos de atenção, com extrações mais relevantes e fiéis ao texto-fonte, reduzindo alucinações.

Table 6. Métricas Consolidadas de Avaliação: mBERT Base vs. mBERT Ajustado

mBERT - Modelo Base (%)				mBERT - Fine-Tuning (LoRA) (%)			
ROUGE		BERTScore		ROUGE		BERTScore	
ROUGE-1	40,30	BERTScore F1	68,97	ROUGE-1	38,74	BERTScore F1	70,04
ROUGE-2	12,69	Precision	68,19	ROUGE-2	14,86	Precision	68,68
ROUGE-L	19,70	Recall	69,82	ROUGE-L	19,24	Recall	71,48

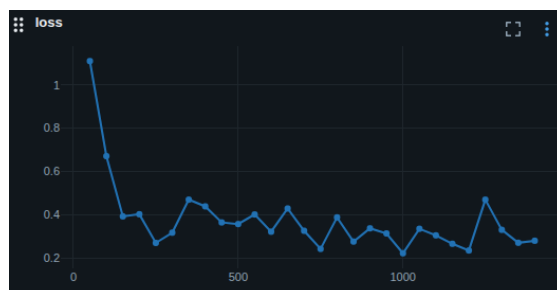
7.4. RoBERTa

Os sumários do *RoBERTa Base* revelaram limitações típicas na sumarização: inícios incompletos, uso de expressões informais (“*Pois é...*”) e alucinações, sobretudo em temas mais complexos. Esses indícios sugerem que o modelo recorria a conhecimentos do pré-treinamento, reduzindo a fidelidade ao documento.

Com o *fine-tuning*, as curvas de perda (MLflow) indicaram aprendizado, porém com convergência irregular (Figura 5, painéis (a) e (b)), possivelmente devido ao tamanho limitado do *dataset* e à complexidade do domínio científico.



(a) Conjunto de teste.



(b) Conjunto de treino.

Figure 5. Curvas de convergência da perda (*loss*) do RoBERTa nos conjuntos de teste e de treino.

Nas métricas (Tabela 7), o *fine-tuning* trouxe ajustes pontuais: ROUGE-1 manteve-se quase estável (36,00% para 35,72%), e o ROUGE-2 aumentou (9,78% para

12,25%), sugerindo melhora na captura de *bigrams*. O BERTScore F1 permaneceu estável (68,86% para 68,60%), com *Recall* de 70,09%. Em conjunto, os resultados indicam redução parcial de alucinações e melhor preservação do conteúdo essencial, ainda que sem ganhos expressivos nas métricas semânticas.

Table 7. Métricas do modelo RoBERTa coletadas durante o fine tuning

RoBERTa - Modelo Base (%)				RoBERTa - Fine-Tuning (%)			
ROUGE		BERTScore		ROUGE		BERTScore	
ROUGE-1	36,00	BERTScore F1	68,86	ROUGE-1	35,72	BERTScore F1	68,60
ROUGE-2	9,78	Precision	69,42	ROUGE-2	12,25	Precision	67,20
ROUGE-L	17,87	Recall	68,36	ROUGE-L	17,22	Recall	70,09

7.5. Discussão geral dos resultados

De forma geral, os experimentos indicam que o *fine-tuning* tende a trazer ganhos mais consistentes em aspectos de sobreposição lexical (especialmente em ROUGE-2), enquanto as melhorias em similaridade semântica (BERTScore) foram mais discretas. Esse comportamento é coerente com o cenário de sumarização científica: pequenas alterações na escolha de termos e na estrutura frasal podem aumentar a coincidência de *n-grams* com o sumário de referência, sem necessariamente elevar, na mesma proporção, a equivalência semântica global do texto gerado.

Ao comparar os modelos, o mBERT apresentou melhor equilíbrio após o ajuste, com aumento do BERTScore F1 e do *Recall*, sugerindo maior preservação do conteúdo relevante do documento, ainda que com leve queda em ROUGE-1 e ROUGE-L. Já no RoBERTa, a melhora em ROUGE-2 aponta avanço na coesão local, mas a estabilidade (ou pequena queda) do BERTScore F1 sugere que o modelo continua sensível a variações de conteúdo e a possíveis alucinações em tópicos mais complexos.

Esses resultados também evidenciam limitações do contexto experimental: o tamanho do *dataset*, a heterogeneidade dos temas e a natureza altamente especializada dos textos científicos podem dificultar uma convergência estável e limitar o ganho obtido por ajustes leves (como LoRA). Assim, além de ampliar e balancear o conjunto de treino, estratégias como controle de comprimento, filtragem/normalização do texto de entrada e validações manuais de factualidade podem ser decisivas para melhorar a qualidade percebida dos sumários, mesmo quando as métricas automáticas variam pouco.

8. Conclusão

Este trabalho investigou o impacto do *fine-tuning* na tarefa de sumarização de textos científicos em português brasileiro, comparando modelos representativos de arquiteturas *encoder-decoder* (PTT5 e Flan-T5) e *encoder-only* (mBERT e RoBERTa). A avaliação combinou análise qualitativa das saídas e métricas automáticas (ROUGE e BERTScore), considerando também limitações práticas impostas pela janela de contexto e pela estratégia de *chunking*.

Os resultados indicaram que o desempenho depende fortemente do alinhamento do pré-treinamento com a língua-alvo e com o domínio do texto. O PTT5 Base apresentou saídas coerentes e bem formadas em português, com maior

sobreposição lexical (p.ex., ROUGE-1=37,14%); entretanto, o ajuste realizado degradou as métricas (ROUGE-1=11,97%; BERTScore F1=66,29%), sugerindo que o processo de adaptação e/ou o conjunto de ajuste não foram suficientes para melhorar a correspondência com resumos de referência. Para o Flan-T5, observou-se instabilidade linguística já no modelo base e, após o *fine-tuning*, evidências de colapso linguístico, atribuídas principalmente à ausência de pré-treinamento em português e à incompatibilidade de vocabulário/tokenização. A estratégia de chunking adotada devido à limitação da janela de contexto dos modelos impactou negativamente a compreensão global dos artigos. Como cada artigo foi fragmentado em múltiplos trechos para possibilitar seu processamento, os minirresumos produzidos para cada fragmento, na maioria dos casos, não foram capazes de representar de forma integrada as informações essenciais do documento, comprometendo a coerência e a visão global do conteúdo.

Nos modelos *encoder-only*, o *fine-tuning* mostrou-se mais promissor: o mBERT Ajustado melhorou a qualidade extrativa e aumentou o BERTScore F1 (de 68,97% para 70,04%), com ganho em ROUGE-2 (de 12,69% para 14,86%), indicando extrações mais relevantes e redução de redundâncias. Já o RoBERTa apresentou ganhos pontuais (principalmente em ROUGE-2), mas sem avanços expressivos em similaridade semântica, reforçando que a qualidade final também é sensível ao tamanho e à diversidade do *dataset*, bem como ao controle de *overfitting*.

Como limitações, destacam-se: (i) a fragmentação do conteúdo causada por *chunking*, que reduz a visão global do documento e pode privilegiar detalhes secundários; (ii) restrições de *tokens* que levam a truncamentos; e (iii) a dependência de métricas automáticas, que nem sempre capturam fidelidade factual e utilidade do resumo para o leitor.

Como trabalhos futuros, recomenda-se ampliar e diversificar o *dataset* com resumos de alta qualidade em português, realizar validação humana (fluência, cobertura, factualidade e concisão), incorporar as palavras-chave de cada artigo como informação adicional de entrada para os modelos e avaliar o impacto dessa estratégia nos resultados finais, explorar estratégias de sumarização hierárquica com agregação global, ajustar hiperparâmetros e regularização (incluindo variantes de LoRA) e avaliar modelos com maior janela de contexto e/ou técnicas de recuperação (RAG) para reduzir alucinações e melhorar a cobertura das ideias centrais.

References

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C.,

- Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2022). On the opportunities and risks of foundation models.
- Bornmann, L., Haunschild, R., and Mutz, R. (2021). Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):224.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.
- Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2022a). Scaling instruction-finetuned language models.
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S. S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E. H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q. V., and Wei, J. (2022b). Scaling instruction-finetuned language models.
- Cohan, A., Dernoncourt, F., Kim, D. S., Bui, T., Kim, S., Chang, W., and Goharian, N. (2018). A discourse-aware attention model for abstractive summarization of long documents.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding.
- Fire, M. and Guestrin, C. (2019). Over-optimization of academic publishing metrics: Observing goodhart’s law in action. *GigaScience*, 8.
- Gambhir, M. and Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, 47:1–66.
- Howard, J. and Ruder, S. (2018). Universal language model fine-tuning for text classification.
- Jurafsky, D. and Martin, J. (2023). *Speech and Language Processing (3rd ed. draft)*. Accessed: 2025-01-10.

- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Narayan, S., Cohen, S. B., and Lapata, M. (2018). Don’t give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Nascimento Rodrigues, J. M. and Conceição, V. P. d. (2025). Uso da sumarização automática de textos na representação da informação e do conhecimento. *ISKO Brasil*, (8).
- Nenkova, A. and McKeown, K. (2012). *A Survey of Text Summarization Techniques*, pages 43–76. Springer US, Boston, MA.
- Paiola, P. H., de Rosa, G. H., and Papa, J. P. (2022). Deep learning-based abstractive summarization for brazilian portuguese texts.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2019). Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.
- Reis, G. S. R. and Araújo, J. (2022). Plataforma de sumarização automática de artigos científicos.
- Salton, G. (2017). *Representations of Idioms for Natural Language Processing: Idiom Type and Token Identification, Language Modelling and Neural Machine Translation*. Doctoral thesis, Dublin Institute of Technology (DIT).
- Sarmiento, M. and de Oliveira, H. (2024). Sumarização automática de artigos de notícias em português: Da extração à abstração com abordagens clássicas e modelos de neurais. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 139–148, Porto Alegre, RS, Brasil. SBC.
- Sousa, H. T. d. and Lima, G. d. (2025). Sumarização automática de textos científicos com inteligência aumentada: Integrando ia na organização do conhecimento. *ISKO Brasil*, (8).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *CoRR*, abs/1706.03762.
- Zhang, J., Zhao, Y., Saleh, M., and Liu, P. J. (2020a). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020b). Bertscore: Evaluating text generation with bert.