Comparação de Desempenho dos Modelos T5 e BART na Tarefa de Sumarização de Textos

1st Otávio Augusto Correia Novais
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
otavio.novais@unifesp.br

2nd Victor Augusto Reis Marques
Instituto de Ciência e Tecnologia
Universidade Federal de São Paulo
São José dos Campos, Brasil
marques.victor@unifesp.br

I. Resumo

Este trabalho apresenta uma análise comparativa semântica entre os modelos de Processamento de Linguagem Natural (PLN) T5 e BART, focando na tarefa de sumarização de avaliações de produtos, mais especificamente livros. Com o crescente volume de avaliações disponíveis, há uma demanda por ferramentas que possam sintetizar essas informações de maneira eficiente. O estudo avalia o desempenho dos dois modelos em três configurações diferentes: sem fine-tuning, com fine-tuning aplicado ao T5, e com fine-tuning aplicado a ambos os modelos. A métrica BERTScore foi utilizada para medir a similaridade semântica entre os resumos gerados e os resumos de referência. Os resultados indicam que o BART, especialmente após o fine-tuning, apresente um desempenho superior na tarefa de sumarização, sendo mais eficaz na preservação da qualidade semântica dos textos.

II. Introdução e Motivação

O crescente volume de avaliações de produtos online oferece uma riqueza de informações úteis para consumidores e empresas. No entanto, a quantidade de dados pode ser avassaladora e difícil de processar manualmente. Nesse ponto que entra o sumarizador de avaliações de produtos, uma ferramenta de inteligência artificial projetada para sintetizar comentários e opiniões de usuários de maneira concisa e significativa.

Um sumarizador de avaliações de produtos utiliza técnicas avançadas de processamento de linguagem natural (PLN) para analisar textos escritos por consumidores. Através de algoritmos de aprendizado de máquina, ele identifica os principais temas, sentimentos e insights expressos nas avaliações. Esse processo resulta em resumos que destacam os pontos positivos e negativos mais frequentes, proporcionando uma visão geral clara e rápida sobre o desempenho de um produto.

Este trabalho é motivado pela demanda crescente por ferramentas eficazes para lidar com o vasto volume de avaliações de produtos online. Modelos de processamento de Linguagem Natural (PLN), como o T5 e o BART, têm sido amplamente utilizados para a tarefa de sumarização automática. No entanto, a eficácia desses modelos em capturar detalhes críticos, como a preservação do conteúdo semântico das frases após a sumarização, ainda requer uma análise mais aprofundada. Com o intuito de aprofundar nessa análise, este estudo foi desenvolvido com o objetivo de avaliar e comparar o desempenho desses dois modelos, buscando identificar qual deles oferece uma maior precisão.

III. Conceitos Importantes e Trabalhos Relacionados

A. Conceitos Importantes

1) Aprendizado de Máquina: O Aprendizado de Máquina é um campo da inteligência artificial focada no desenvolvimento de algoritmos que, a partir de dados, consigam aprender e otimizar determinada tarefa sem a necessidade de uma programação especifica para tal tarefa (MITCHELL, 1997). Esse método é utilizado em PLN para identificação de algumas características da fala, como sarcasmos, ironia, metáforas, variações na estrutura da frase, exceções gramaticais, entre outras características.

2) Aprendizado Profundo: O Aprendizado Profundo é uma subárea do Aprendizado de Máquina que usa várias camadas de redes neurais para processamentos complexos, gerando uma rede semelhante ao cérebro. Através dessa área, os computadores conseguem reconhecer, classificar e correlacionar padrões nos dados de entrada (GOODFELLOW; BENGIO; COURVILLE, 2016).

3) BART - Bidirectional and Auto-Regressive Transformers: BART é um modelo de transformação baseado em uma arquitetura híbrida que combina características dos autoencoders e autoregressivos. Ele é treinado primeiro para corromper texto de entrada e, em seguida, para reconstruí-lo, permitindo que o modelo compreenda a estrutura global do texto e corrija erros. Isso resulta em um modelo altamente eficaz para tarefas de geração de texto, como sumarização e tradução, oferecendo um entendimento profundo e uma capacidade robusta de capturar contextos complexos. (LEWIS et al., 2020).

4) BERT - Bidirectional Encoder Representations from Transformers: É um modelo de linguagem desenvolvido pela Google apresentado em por DEVLIN et al. em 2018. Esse modelo é baseado em transformadores e causou grande repercussão na época do lançamento por considerar

o contexto bidirecional das palavras, possibilitando um entendimento mais preciso das palavras em diferentes contexto.

5) BERTScore: É uma métrica para tarefas de geração de texto que mede a similaridade semântica entre dois textos com base nos embeddings contextuais (representações vetoriais dos dados) derivados do modelo BERT. Essa métrica permite a captura de nuances semânticas que outras métricas, como ROUGE ou Distância dos Cossenos, não conseguem identificar. (ZHANG et la., 2019).

Essa métrica é composta por três métricas:

- Recall BERTScore (RBERT): Essa métrica compara a similaridade de cada token da sentença de referência com o token mais similar da sentença candidata. O objetivo é avaliar o quão bem o candidato capturou o conteúdo referência.
- Precision BERTScore (PBERT): Realiza a avaliação ao sentido contrário da RBERT, candidato para referência. O foco da PBERT é medir o quão relevante é o conteúdo do candidato em relação à referência.
- F1 BERTScore (FBERT): O FBERT, ou F1-score, oferece uma visão geral da qualidade da sentença candidata unindo o RBERT e a PBERT, realizando uma avaliação da capacidade de captura do conteúdo da referência e da relevância do conteúdo gerado.

Neste trabalho, foi utilizado o F1-score como métrica para a comparação dos textos gerados e os textos presentes na base de dados.

$$R_{\text{BERT}} = \frac{1}{|x|}\sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \ , \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|}\sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \mathbf{x}_i^\top \hat{\mathbf{x}}_j \ , \quad F_{\text{BERT}} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \ .$$

Figura 1: Fórmulas para cálculo das três métricas que compõem o BERTScore.
Fonte: ZHANG et la., 2019.

6) Cross-validation - Validação Cruzada: Este método divide o conjunto de dados em cinco subconjuntos (folds) aproximadamente iguais. Em cada iteração, um dos subconjuntos é utilizado como conjunto de validação, enquanto os quatro restantes são utilizados para treinamento do modelo. O processo é repetido cinco vezes, alternando o subconjunto de validação a cada iteração, garantindo que todos os dados sejam utilizados tanto para treinamento quanto para validação (Kohavi, 1995).

7) Linguística Computacional: Segundo Manning e Schütze (1999), a Linguística Computacional é a ciência que busca desenvolver modelos e algoritmos capazes de processar e analisar a linguagem humana. Para atingir esse objetivo, os pesquisadores utilizam métodos como análise sintática, análise semântica e identificação de sintagmas, entre outros, para compreender a estrutura e o funcionamento da linguagem natural (MANNING; SCHÜTZE, 1999).

8) Pragmática: A pragmática é o ramo que estuda a maneira como o contexto altera o significado da linguagem (LEVINSON, 1983).

9) Processamento de Linguagem Natural: O Processamento de Linguagem Natural (PLN) é um ramo da Inteligência Artificial (IA) focado na interação entre computadores e a linguagem humana, com o objetivo de capacitar os computadores a processar e analisar grandes volumes de dados de linguagem natural (JURAFSKY; MARTIN, 2021). Uma outra definição para PLN é o desenvolvimento de modelos computacionais que utilizam entradas de informações em formato de linguagem natural para executar alguma tarefa (PEREIRA; 2011).

Através do PLN, é possível realizar diversas tarefas, como a análise de sentimentos, conversão de fala em texto escrito, tradução automática, extração de informações, geração de novos textos com base em conjuntos de dados, criação de chatbots e assistentes virtuais, resumos automáticos e análise de textos. O PLN combina três conceitos fundamentais para processar a linguagem natural: Linguística Computacional, Aprendizado de Máquina (Machine Learning) e Aprendizado Profundo (Deep Learning).

10) Semântica: Ramo da linguística focado no estudo do significado, interpretação e entendimento das palavras, frases e textos. Contemplando o estudo de sinônimos, antônimos, homônimos, parônimos, polissemia, conotação e denotação (PALMER, 1981).

11) Sintaxe: É o estudo de como uma frase é estruturada, considerando as regras e hierarquia que moldam essa formação (CARNIE, 2013).

12) Transformers: São modelos de aprendizado profundo focado em processar dados sequenciais, como textos. Esses modelos são compostos 3 tipos de camadas: Mecanismo de Atenção, que calcula a importância de diferentes partes da entrada, Camada de Codificador, responsável por refinar a entrada em representações hierarquicamente mais ricas, e Camada de Decodificador, similar às camadas de codificador com o adicional do mecanismo de atenção para focar nas saídas do codificador (VASWANI et al. 2017).

13) Text-To-Text Transfer Transformer - T5: O T5 é um modelo que converte todas as tarefas de processamento de linguagem natural em um formato de entrada e saída de texto, permitindo que o mesmo modelo seja aplicado a uma variedade de tarefas. A arquitetura é baseada em transformadores e é treinada em um grande corpus de dados através de uma tarefa de preenchimento de lacunas. Essa abordagem unificada simplifica a adaptação do modelo a diferentes aplicações, mantendo a consistência e eficiência em tarefas como tradução, sumarização e classificação de textos. (RAFFEL et al., 2020).

14) Token: O token é a unidade básica de processamento em PLN, podendo ser uma palavra, uma sílaba ou uma letra, a depender da tarefa a ser realizada (JURAFSKY; MARTIN, 2021).

B. Trabalhos Relacionados

1) AI-Assisted Summarization of Radiological Reports: Nesse artigo é analisado o desempenho de diferentes

modelos de processamento de linguagem natural (NLP) na sumarização de relatórios radiológicos. Para avaliar a qualidade dos resumos gerados, o estudo utilizou métricas como ROUGE e BERTScore, sendo esta última utilizada para medir a similaridade semântica entre os resumos gerados pelos modelos e os resumos de referência criados por especialistas na área de radiologia (CHIEN et al, 2024).

2) Fine-tuning T5 and RoBERTa Models for Enhanced Text Summarization and Sentiment Analysis: Em MENGI, GHORPADE, KAKADE (2023) usaram o transformador T5 e RoBERTa com foco na melhoria na sumarização de textos e na análise de sentimentos. O T5 (Text-to-Text Transfer Transformer) trata-se de uma técnica de PLN criado pelo Google Research que consideram a entrada e a saída como sequências de texto, permitindo uma maior flexibilidade e consistência na realização das tarefas. A RoBERTa (Robustly Optimized BERT Pretraining Approvach) é uma versão otimizada do BERT, apresentando melhor desempenho devido a modificações no pré-treinamento e melhoria de hiper-parâmetros. Após ajustes finos, o modelo T5 apresentou grande eficácia na geração de resumo coerentes e concisos e a RoBERTa conseguiu analisar com precisão os sentimentos nos dados utilizados.

3) Summarizing Business News: Evaluating BART, T5, and PEGASUS for Effective Information Extraction: Nesse estudo é explorado a eficácia de três modelos de ponta para sumarização automática de notícias de negócios: BART, T5 e PEGASUS. O estudo compara os resumos gerados pelos modelos com os resumos de referência para identificar o modelo que melhor preserva a qualidade semântica e a precisão da informação. Para a análise da eficácia é usados as métricas ROUGE-1, ROUGE-P e METEOR. A métrica ROUGE-1 é ideal em medir a presença de termos específicos e a cobertura geral do conteúdo essencial, o ROUGE-P é utilizado para medir a exatidão das informações e a inclusão de informações irrelevantes ou incorretas, por fim, o METEOR é aplicado na avaliação quando a qualidade linguística e a similaridade semântica são importantes (DHARRAO et al, 2024).

IV. Objetivos

O objetivo deste projeto é conduzir uma análise comparativa detalhada entre os modelos de sumarização T5 e BART, com o intuito de avaliar e determinar qual deles oferece um desempenho superior na tarefa específica de sumarizar reviews de usuários. Este estudo abrangerá, principalmente, a qualidade semântica dos resumos gerados, focando a análise em identificar qual modelo conseguiu produzir resumos que mantivesse a qualidade dos reviews com base em um resumo pré-gerado com base na métrica BERTScore. Ao final, espera-se identificar o modelo que proporciona o melhor desempenho para essa aplicação específica.

V. Metodologia Experimental

A. Coleta de dados:

1) Fonte de dados e coleta: Os dados para o trabalho foram obtidos do *dataset "Amazon Books Reviews"*[1] do autor *Mohamed Bekheet*, hospedado no site do *Kaggle*. A coleta do banco de dados foi realizada no dia 24 de junho de 2024.

2) Critérios de seleção: Este *dataset* foi escolhido devido à sua vasta quantidade de informações, contendo aproximadamente 1,5 milhões de avaliações escritas de livros. Além disso, o *dataset* inclui resumos das avaliações dos livros, o que permite o treinamento supervisionado dos modelos. Para otimizar o processamento e o treinamento, foi restringido o banco de dados a reviews com um comprimento entre 30 e 60 palavras.

3) Sobre o dataset: O *dataset* possui dois conjuntos de dados, *"Books_rating"* e *"Books_data"*, que estão relacionados pelo campo *Title* (Figura 2). Os dados do conjunto *Books_rating* foram coletados no intervalo de maio de 1996 até setembro de 2023 [2]. Já os dados do conjunto *Books_data* foram obtidos através do Google Books API [3], não havendo a data de coleta dos dados.

O conjunto de dados *"Books_rating"* conta com o reviews do livro e o resumo desses reviews feitos através do modelo ChatGPT.

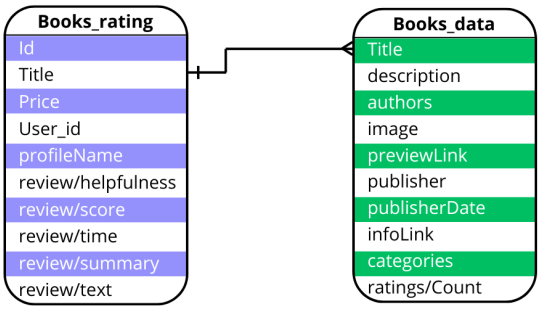

Figura 2: Relacionamento entre banco de dados *Books_rating* e *Books_data*.
Fonte: Autores

4) Limpeza dos dados: Para a limpeza dos dados, os seguintes processos foram aplicados nos dois conjuntos de dados:

- Remoção de linhas duplicadas: A função *"drop_duplicates()"* foi utilizada para remover linhas duplicadas, considerando como duplicadas apenas aquelas em que todas as colunas possuíam valores idênticos em outra linha.
- Remoção de colunas irrelevantes: Foram descartadas colunas com informações que não agregariam

[1]https://www.kaggle.com/datasets/mohamedbakhet/amazon-books-reviews
[2]https://amazon-reviews-2023.github.io/
[3]https://developers.google.com/books/docs/overview

ao aprendizado. No conjunto *"Book_rating"*, colunas como Id, use_id, review/time foram removidas. No conjunto *Book_data*, colunas como description, image, previewLink foram excluídas.
- Remoção de linhas com valores ausentes: Após a remoção das colunas desnecessárias, todas as linhas que possuíam pelo menos um valor ausente foram eliminadas.

5) Formatação dos dados: Os dois *DataFrames* foram unidos com base na coluna comum *Title*, gerando o novo banco de dados *df_base*. A partir desse momento, todos os procedimentos citados serão realizados sobre esse novo *DataFrame*.

Para formatar os dados presentes na coluna *"categories"*, utilizamos uma expressão regular para extrair apenas o conteúdo relevante, removendo caracteres indesejados.

6) Criação dos conjuntos de validação (treinamento e teste): A criação dos conjuntos de validação será realizada utilizando a técnica de validação cruzada com k-fold, onde k = 5. Essa abordagem é eficaz para minimizar o viés na avaliação do modelo, proporcionando uma estimativa mais robusta de seu desempenho geral, especialmente em conjuntos de dados com variabilidade significativa. Ao final, os resultados das cinco iterações são agregados, fornecendo uma média das métricas de avaliação que reflete melhor a capacidade do modelo de generalizar para dados não vistos.

B. Implementação do Modelos T5 e BART:

Os modelos foram implementados utilizando a biblioteca Transformers, desenvolvida e mantida pela Hugging Face, que fornece uma ampla gama de ferramentas para trabalhar com modelos de linguagem baseados em transformadores, incluindo aqueles criados pelo Google, como BERT, T5, e outros.

C. Protocolo de Experimentação:

O trabalho foi estruturado em três experimentos, cada um projetado para analisar o desempenho dos modelos em configurações diferentes.

Para garantir uma avaliação robusta em cada uma das configurações e minimizar variações decorrentes da amostragem dos dados, utilizou-se a técnica de validação cruzada com k-fold, onde k = 5, em todos os experimentos. Esse método permite ajustes mais precisos nos experimentos em que foi realizado o fine-tuning.

1) T5 e BART sem Fine-Tuning: Nesse experimento, o objetivo é avaliar o desempenho dos modelos em suas versões pré-treinadas, sem qualquer ajuste fino adicional (fine-tuning). Essa abordagem permite uma compreensão da capacidade dos modelos em generalizar conhecimentos adquiridos durante o pré-treinamento para novas tarefas.

2) T5 com Fine-Tuning e BART sem Fine-Tuning: Nessa configuração, o foco é analisar o impacto do fine-tuning no desempenho do modelo T5 em tarefas especificas, comparando-o diretamente com o modelo BART em sua versão pré-treinada, sem qualquer ajuste fino adicional.

3) T5 e BART com Fine-Tuning: Nesta última configuração, o objetivo é analisar o impacto do ajuste fino tanto no T5 quanto no BART para tarefas especificas. Essa abordagem permite uma análise profunda sobre qual modelo possui melhor desempenho em diferentes configurações.

D. Avaliação:

A avaliação será realizada após cada iteração do processo de validação cruzada. Em cada etapa, os resumos gerados pelos modelos serão comparados com os resumos de referência presentes no banco de dados utilizando a métrica BERTScore. Essa métrica permite uma avaliação da similaridade semântica entre os textos, avaliando se os resumos gerados mantém o significado original.

VI. Resultados

A. Experimento 1:

Apesar da ausência de fine-tuning, os resumos gerados pelos modelos demonstraram desempenho ótimo, com similaridade média acima de 80%, quando comparados aos resumos da base de dados, conforme avaliado pela métrica BERTScore.

As Figuras 3 e 4, apresentam as médias das cinco etapas de validação cruzada para cada um dos modelos. A partir dessas figuras, conclui-se que o BART obteve um melhor desempenho no Experimento 1. Esse desempenho indica que a versão pré-treinada do BART é mais eficaz para a tarefa de sumarização possuindo uma qualidade semântica maior e demonstrando maior capacidade de generalização para novas tarefas.

As Figuras 5 até 14 ilustram o desempenho dos modelos em cada um dos folds.

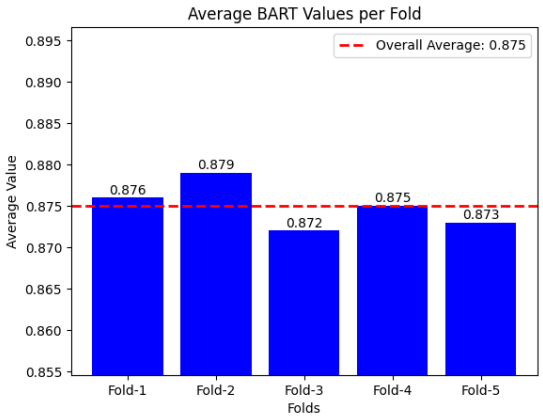

Figura 3: Valores médios de BART por fold - Experimento 1

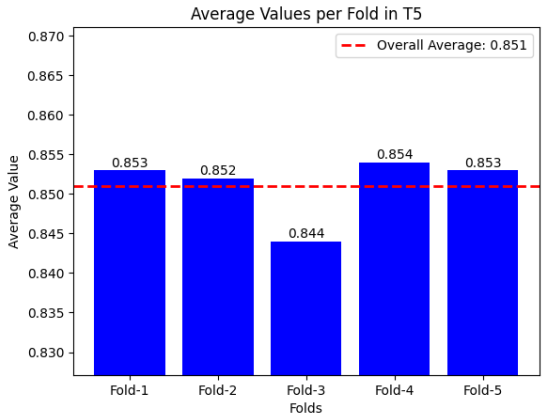

Figura 4: Valores médios de T5 por fold - Experimento 1

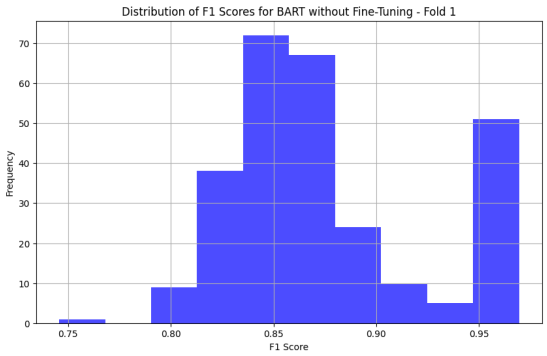

Figura 5: Experimento 1 - 1-fold BART

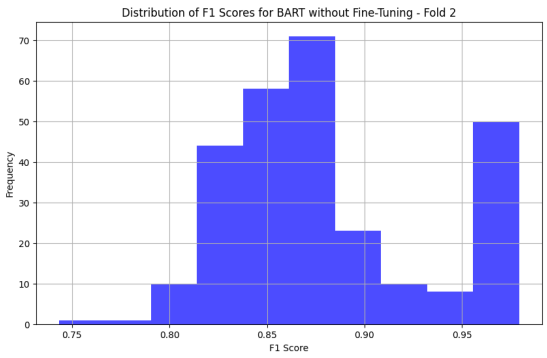

Figura 6: Experimento 1 - 2-fold BART

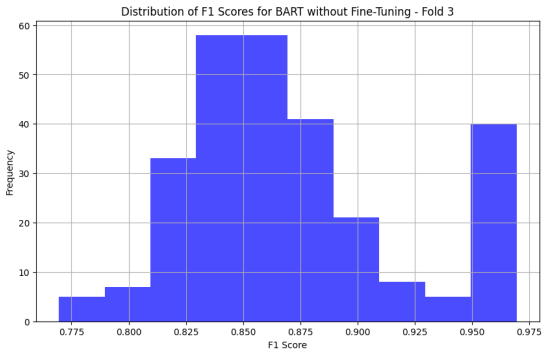

Figura 7: Experimento 1 - 3-fold BART

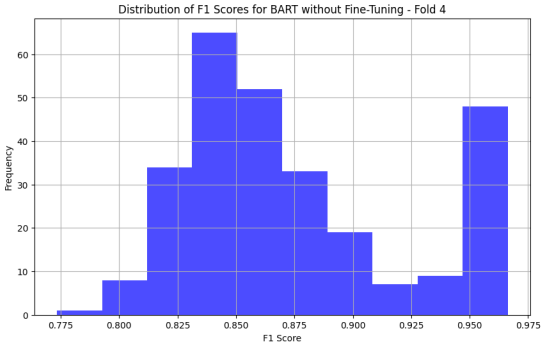

Figura 8: Experimento 1 - 4-fold BART

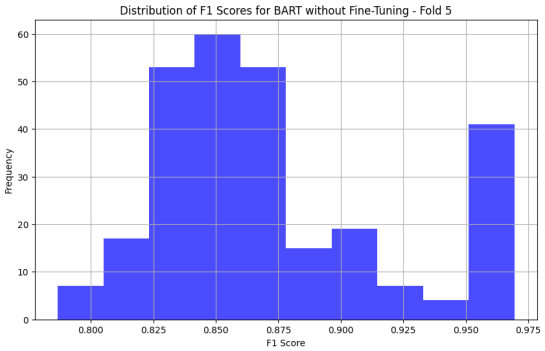

Figura 9: Experimento 1 - 5-fold BART

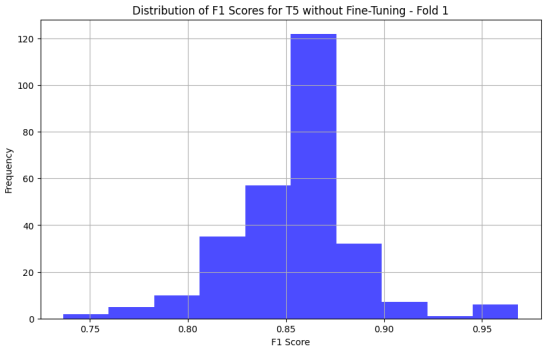

Figura 10: Experimento 1 - 1-fold T5

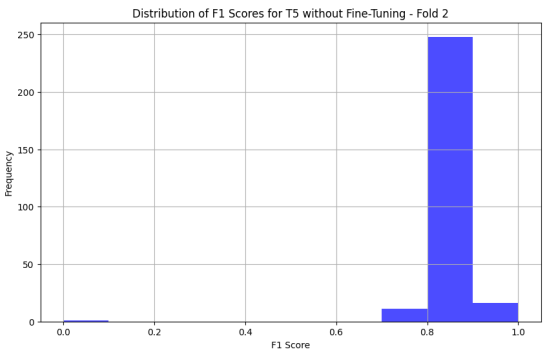

Figura 11: Experimento 1 - 2-fold T5

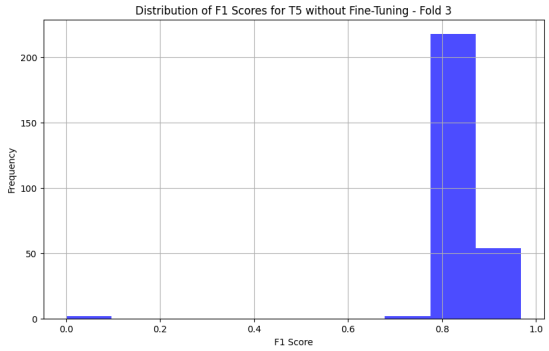

Figura 12: Experimento 1 - 3-fold T5

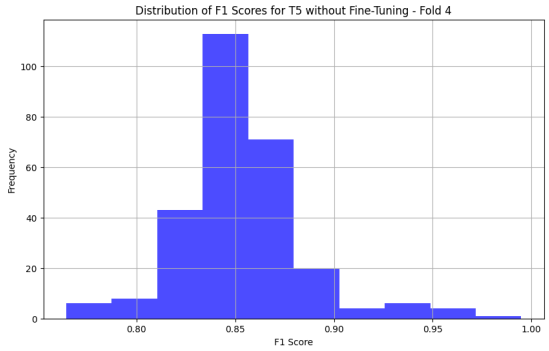

Figura 13: Experimento 1 - 4-fold T5

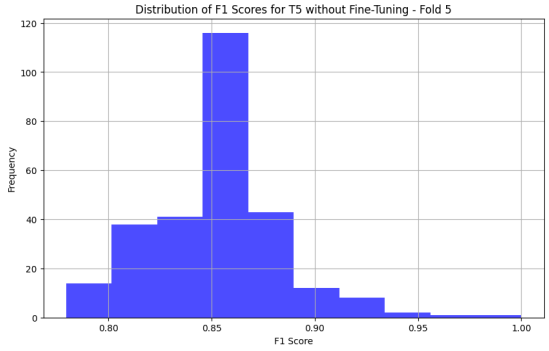

Figura 14: Experimento 1 - 5-fold T5

B. Experimento 2:

Como a configuração do BART nos Experimentos 1 e 2 é idêntica, utilizando apenas os modelos pré-treinados, não houve variação nos valores observados entre os dois experimentos (Figuras 3 e 15).

Quando comparado o valor médio de similaridade entre o Experimento 1 (Figura 4) e o Experimento 2 (Figura 16), observa-se um aumento de 4,1% do valor médio de similaridade. Essa melhoria é justificável devido ao fine-tuning, que ajusta o modelo de forma mais precisa à tarefa proposta.

Com o fine-tuning, o modelo T5 superou o BART sem fine-tuning, apresentando resultados melhores (Figuras 16 e 15, respectivamente). O fine-tuning capacitou o T5 a capturar padrões mais específicos dos dados de sumarização, algo que o BART em sua versão pré-treinada não foi capaz de alcançar. Esse aprimoramento ressalta a importância do fine-tuning na otimização de modelos para tarefas específicas, resultando em um desempenho mais eficaz.

As Figuras 17 até 26 apresentam o desempenho dos modelos em cada um dos folds.

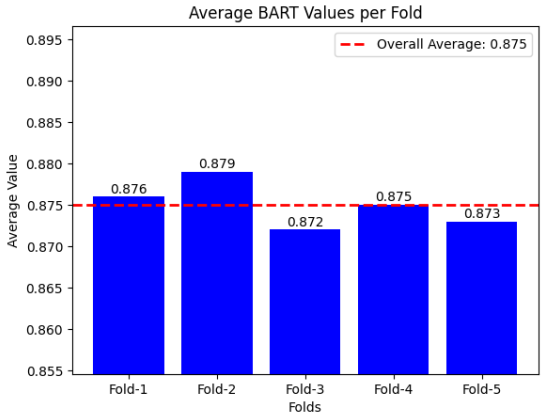

Figura 15: Valores médios de BART por fold - Experimento 2

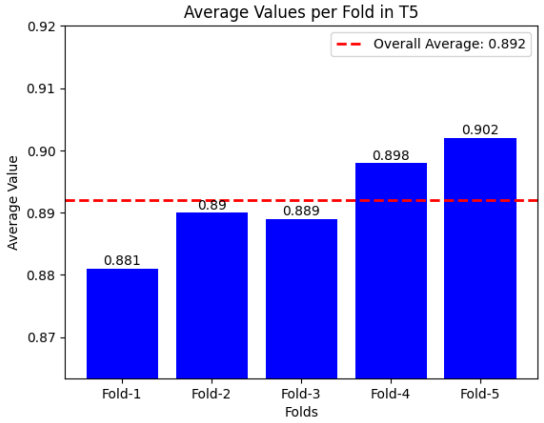

Figura 16: Valores médios de T5 por fold - Experimento 2

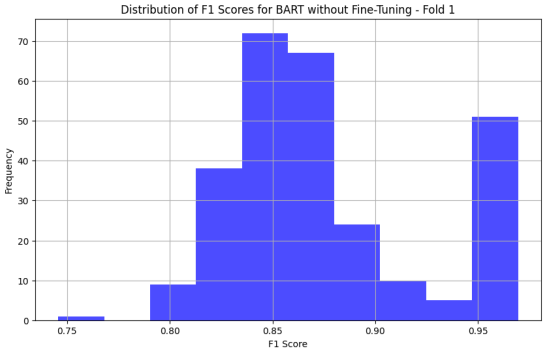

Figura 17: Experimento 2 - 1-fold BART

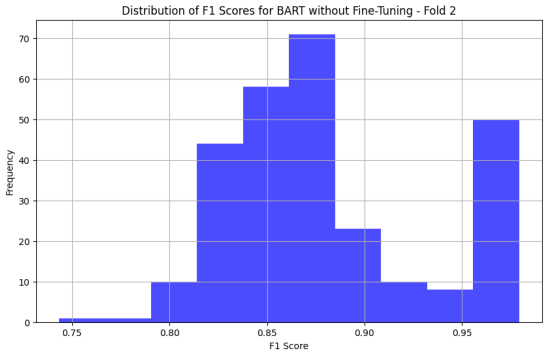

Figura 18: Experimento 2 - 2-fold BART

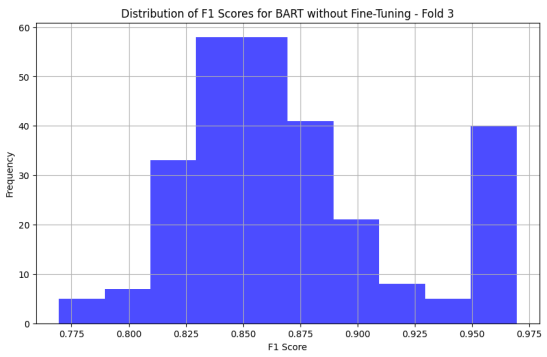

Figura 19: Experimento 2 - 3-fold BART

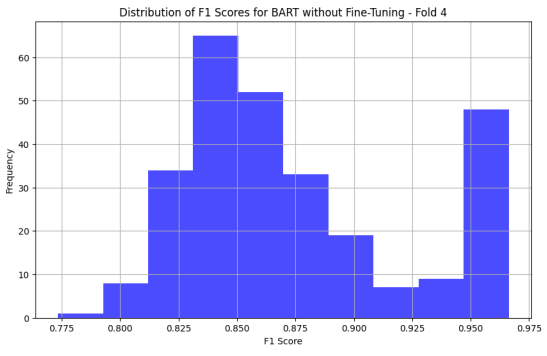

Figura 20: Experimento 2 - 4-fold BART

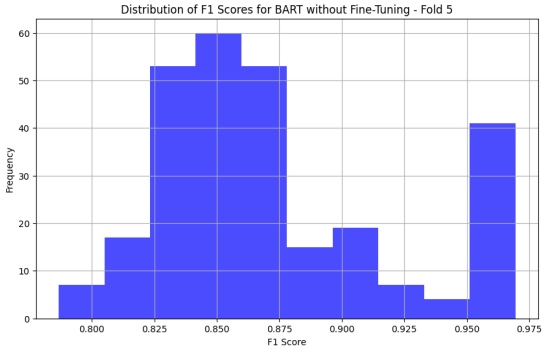

Figura 21: Experimento 2 - 5-fold BART

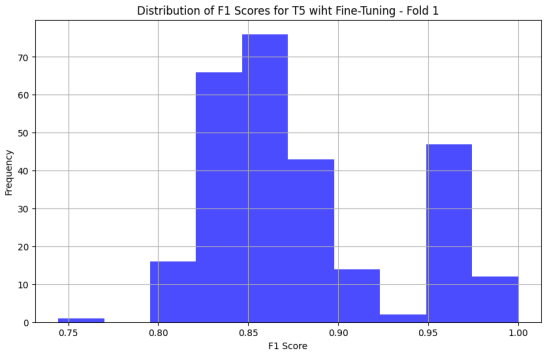

Figura 22: Experimento 2 - 1-fold T5

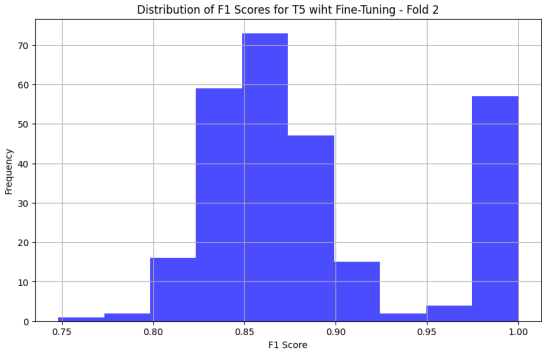

Figura 23: Experimento 2 - 2-fold T5

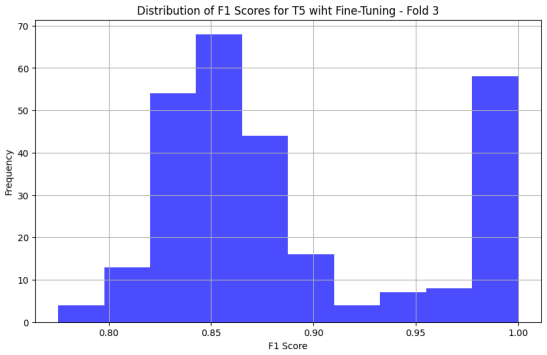

Figura 24: Experimento 2 - 3-fold T5

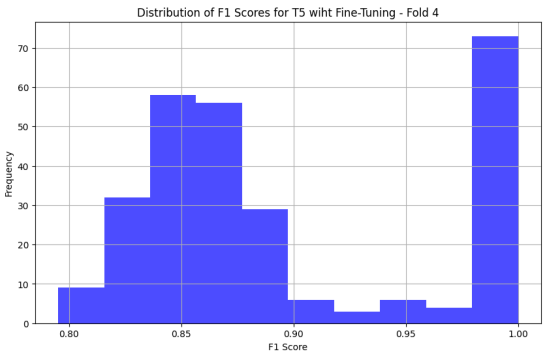

Figura 25: Experimento 2 - 4-fold T5

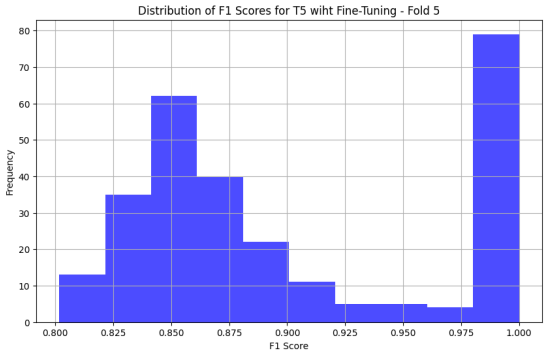

Figura 26: Experimento 2 - 5-fold T5

C. *Experimento 3:*

Como a configuração do T5 nos experimentos 2 e 3 é idêntica, utilizando fine-tuning em ambos, não houve variação entre os valores observados entre os experimentos para esse modelo(Figuras 34 a 38) .

Com a aplicação do fine-tuning no modelo BART, detectou-se um aumento de 10,3% na similaridade média (Figuras 15 e 27). Esse aumento evidencia um ajuste eficaz do modelo à tarefa proposta, indicando a importância da técnica para adequação ao contexto específico. As Figuras 29 a 33 demonstram o desempenho do modelo com o fine-tuning em cada k-fold e na Figura 27 é indicado o desempenho médio.

Nesse último experimento, o modelo BART com fine-tuning apresentou resultados melhores que o T5 também com fine-tuning, constatando-se que o BART é mais adequado para tarefas de sumarização de texto em comparação ao T5 (Figuras 27 e 28, respectivamente). Esse resumo é corroborado por estudos recentes que avaliaram o desempenho dos modelos BART, T5 e PEGASUS, onde o BART se destacou na tarefa de sumarização, sendo capaz de recriar fielmente o conteúdo de referência (DHARRAO et al, 2024).

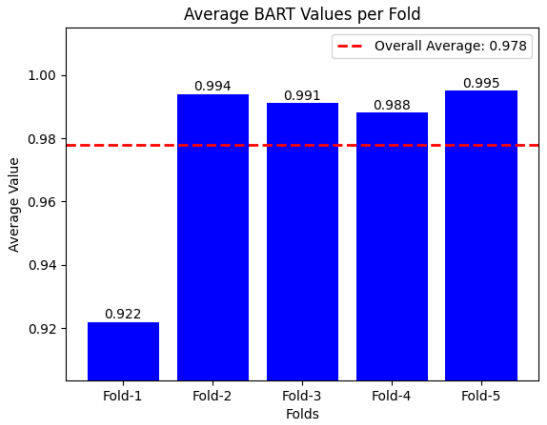

Figura 27: Valores médios de BART por fold - Experimento 3

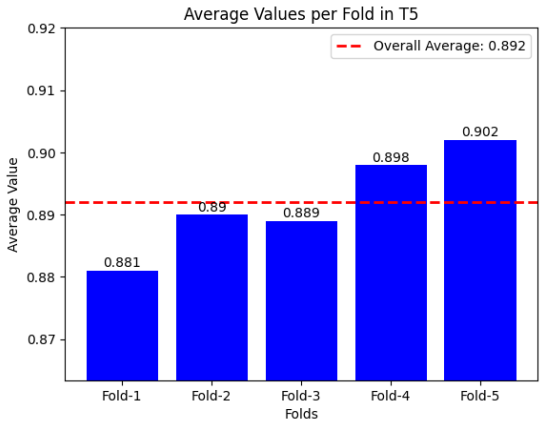

Figura 28: Valores médios de T5 por fold - Experimento 3

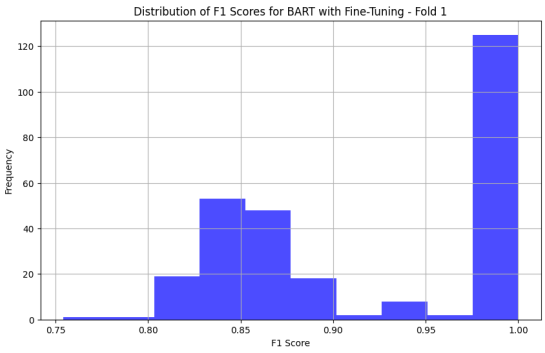

Figura 29: Experimento 3 - 1-fold BART

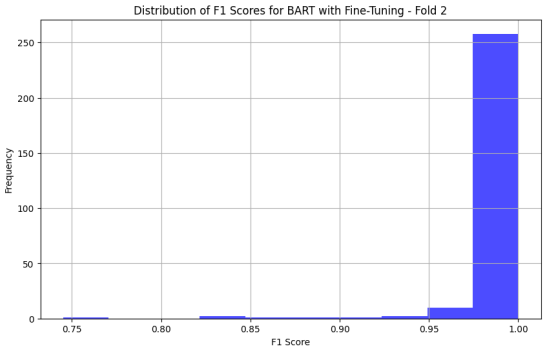

Figura 30: Experimento 3 - 2-fold BART

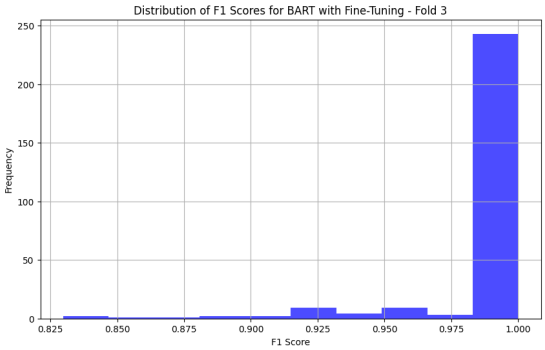

Figura 31: Experimento 3 - 3-fold BART

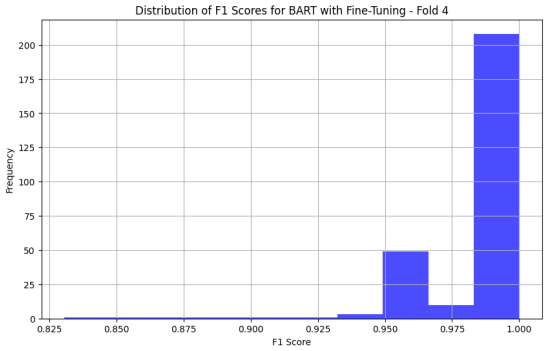

Figura 32: Experimento 3 - 4-fold BART

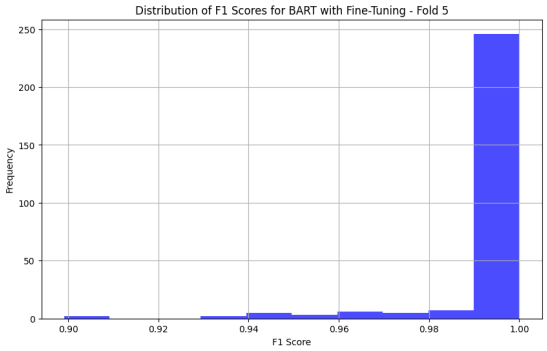

Figura 33: Experimento 3 - 5-fold BART

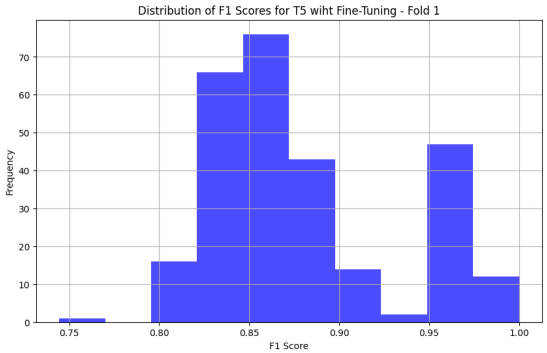

Figura 34: Experimento 3 - 1-fold T5

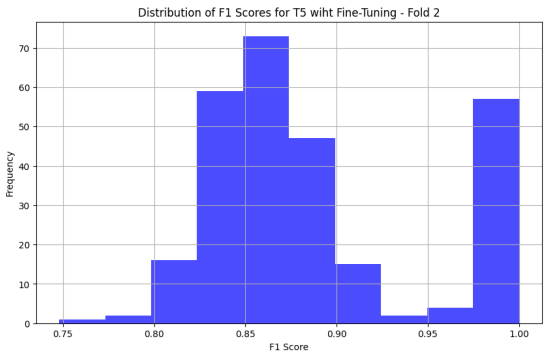

Figura 35: Experimento 3 - 2-fold T5

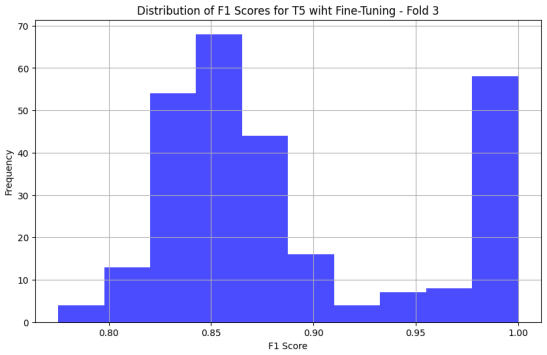

Figura 36: Experimento 3 - 3-fold T5

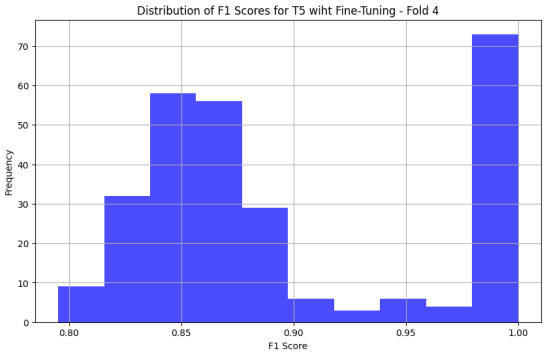

Figura 37: Experimento 3 - 4-fold T5

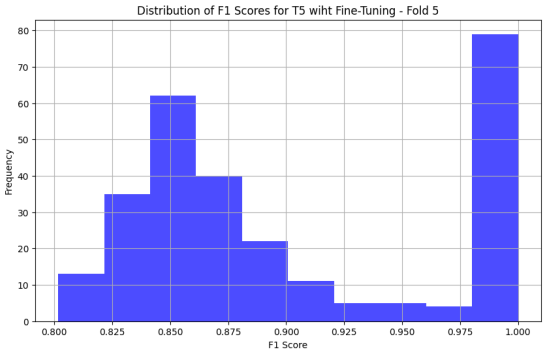

Figura 38: Experimento 3 - 5-fold T5

D. Visão Geral:

Na Figura 39, é possível comparar a performance de cada modelo em todos os experimentos efetuados. Após analisar os resultados,é evidente que o modelo BART é melhor para a tarefa de sumarização do que o modelo T5, independentemente de haver fine-tuning ou não.

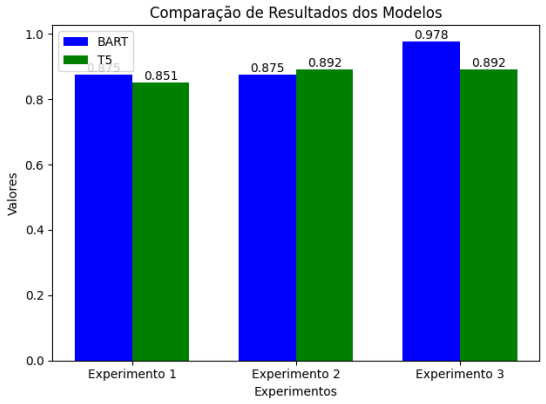

Figura 39: Gráfico comparativo

VII. Conclusões e Trabalhos Futuros

Este estudo teve como objetivo identificar entre os modelos T5 e BART, aquele que apresenta melhor desempenho na tarefa de sumarização de reviews de produtos. Nossos resultados indicam que o BART possui melhor capacidade de conservar a similaridade semântica, independentemente de haver fine-tuning ou não. Embora ambos os modelos tenham demonstrado competência na tarefa de sumarização, o BART se destacou no aspecto de qualidade dos resumos gerados, especificamente após o fine-tuning. Esse achado é siginificado, pois destaca o BART como ferramente mais robusta para tarefas de sumarização quando a precisão semântica é crucial.

Para trabalhos futuros, seria interessante expandir a análise para outros modelos amplamente usados, como o PEGASUS, fornecendo insights sobre os prós e contras de cada modelo. Outro caminho é adicional outras métricas para a a avaliação dos modelos, como ROUGE, BLEU, METEOR e MoverScore.

VIII. AGRADECIMENTOS

Agradecimentos especiais para André Almeida, CEO da Dom Rock, que nos ajudou na idealização do trabalho e apoiou no desenvolvimento do trabalho.

Referências

[1] ARONOFF, M.; FUDEMAN, K. What is Morphology? 2. ed. Wiley-Blackwell, 2011.

[2] CARNIE, A. Syntax: A Generative Introduction. 3. ed. Wiley-Blackwell, 2013.

[3] CHEN, X.; HE, Z.; HOU, Y.; LI, J.; MCAULEY, J.; YAN, A. BRIDGIND LANGUAGE AND ITEMS FOR RETRIEVAL AND RECOMMENDATION, 2024.

[4] CHIEN, A., et al. (2024). AI-Assisted Summarization of Radiological Reports: Evaluating GPT3davinci, BARTcnn, LongT5booksum, LEDbooksum, LEDlegal, and LEDclinical. AJNR Am J Neuroradiol.

[5] DHARRAO, D.; MISHRA, M.; KAZI, A.; PANGAVHANE, M.; PISE, P.; BONGALE, A. SUMMARIZING BUSINESS NEWS: EVALUATING BART, T5 AND PEGASUS FOR EFFECTIVE INFORMATION EXTRACTION, 2024

[6] DEVLIN, J.; CHANG, M.-W.; LEE, K.; TOUTANOVA, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). doi:10.18653/v1/N19-1423. 2019

[7] EDGE, Darren et al. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. 2024.

[8] GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep Learning. MIT Press, 2016.

[9] HOSKING, T.; TANG, H.; LAPATA, M. Hierarchial Indexing for Retrieval-Augmented Opinion Summarization. arXiv, 2024.

[10] JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. 3. ed. Prentice Hall, 2021.

[11] JURAFSKY, D.; MARTIN, J. H. Speech and Language Processing (3rd ed.). Prentice Hall, 2019.

[12] KOHAVI, R. (1995). A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137–1143.

[13] LEVINSON, S. C. Pragmatics. Cambridge: Cambridge University Press, 1983.

[14] LEWIS, P., PEREZ, E., PIKTUS, A., PETRONI, F., KARPUKHIN, V., GOYAL, N., ... & RIEDEL, S. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401.

[15] LEWIS, M.; LIU, Y.; GOYAL, N.; GHANDEHARI, M.; MOHAMED, A.; LEVINE, E.; STOYANOV, V.; ZETTLEMAIER, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv preprint arXiv:1910.13461, 2020.

[16] MANNING, C. D.; SCHÜTZE, H. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press, 1999.

[17] MENGI, R; GHORPADE, H; KAKADE, A. Fine-tuning T5 and RoBERTa Models for Enhanced Text Summarization and Sentiment Analysis. The Great Lakes Botanist, 2023.

[18] MITCHELL, T. M. Machine Learning. New York: McGraw-Hill, 1997.

[19] OVED, N.; LEVY, R.. PASS: Perturb-and-Select Summarizer for Product Reviews. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, August 1–6, 2021, pages 351–365. ©2021 Association for Computational Linguistics

[20] PALMER, F. R. Semantics. Cambridge: Cambridge University Press. 1981.

[21] PEREIRA, S.L. Processamento de Linguagem Natural. USP. 2011.

[22] RAFFEL, C.; SHAZIER, N.; ROBERTS, A.; LEE, K.; NARANG, S.; MATENA, M.; ZHOU, Y.; LI, W.; LIU, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 2020.

[23] TABOSA, H. R., SOUZA, O. de, CÂNDIDO, J. C. dos S., MELO, A. C. A. U., & REIS, K. G. B. Avaliação do desempenho de um software de sumarização automática de textos. Informação & Informação, 189–210, 2020.

[24] VASWANI, A.; SHAZEER, N.; PARMAR, N.; USZKOREIT, J.; JONES, L., GOMEZ, A. N.; ... POLOSUKHIN, I. (2017). Attention is All You Need. arXiv preprint arXiv:1706.03762.

[25] ZHANG, T., KISHORE, V., WU, F., WEINBERGER, K.Q., & ARTZI, Y. (2019). BERTScore: Evaluating Text Generation with BERT. arXiv preprint arXiv:1904.09675.